IDENTIFYING OPTIMAL LIGHT WAVELENGTH ON THE PLANT GROWTH FOR INDOOR VERTICAL FARM

Christopher Vukmir[1] | Tamjid Azad[2] | Abdul Rabbani Syed[3]
Abdul Mannan Mohammed[4] | Mohd Aleem Uddin[5]

Abstract

This research aims to optimize light wavelength for plant growth in vertical farms, a crucial factor due to the substantial energy and associated costs required for artificial LED lighting. Traditional vertical farming practices face challenges with high electricity expenses, limiting their widespread adoption. Our study addresses this by employing advanced machine learning (ML) algorithms and image analysis techniques to predict the most effective light wavelength in the Electromagnetic spectrum for enhancing plant growth rates. Previously, the optimal growth rate was gauged using an R-CNN network model based on leaf area measurements over time, yielding average accuracy. We have extended this approach by comparing the efficacy of Multiple Linear Regression and deep learning neural network models, particularly the Artificial Neural Network (ANN), in a controlled vertical farm environment. Our comparative analysis focuses on the accuracy of these models in predicting plant growth area with minimal loss. In our experimental setup, four trays were equipped with cameras and lighting systems, and real-time datasets were analyzed. The ANN model demonstrated superior performance over traditional ML models. Additionally, we explored the relative contributions of environmental factors like soil quality, moisture, humidity, nutrients, and the consumption of carbon dioxide and oxygen. These factors are identified as critical features in our predictive models, further enhancing their accuracy and practical applicability in optimizing lighting for indoor vertical farms.

1. Introduction

As the global population is projected to reach 8.5 billion by 2030 and 9.7 billion by 2050, and with the challenges of unpredictable climate patterns becoming more prominent, there is a critical need to rethink our agricultural practices. This study addresses this imperative by focusing on the revolutionary approach of indoor vertical farming, a method poised to redefine agricultural efficiency and sustainability.

Vertical farming is particularly characterized by the strategic arrangement of plants in stacked layers within controlled environments. It offers an unfamiliar solution to the challenges of urban food production. It supports year-round production, reduces reliance on long-distance transportation, and minimizes water usage and the need for chemicals or pesticides. However, one of the major hurdles to widespread adoption is the energy cost, especially for artificial lighting.

Our group employed a combination of analytical methods and predictive models, including image analysis, Multiple Linear Regression, Decision Tree, Random Forest, and Artificial Neural Networks, to develop a predictive framework for indoor vertical farming. We have also utilized PlantCV, an open-source image processing software, for in-depth plant phenotyping. This integration of computer vision and machine learning techniques is designed to determine the most effective combinations of light wavelengths for indoor farming.

Our study hopes to contribute to the growing body of knowledge in vertical farming and demonstrates the potential of utilizing machine learning practices in agriculture. This paper details our methodology, analysis, and the implications of our research. The objective of this research is to offer insights into the intricate relationship between light wavelengths and plant growth in controlled environments, thereby paving the way for more sustainable and efficient agricultural practices.

2. Literature Review

2.1 Vertical Farming and Plant Phenotyping

Vertical farming is a concept popularized in the early 2000s. It represents a paradigm shift in agriculture, particularly in urban areas. Despommier introduced vertical farming as a sustainable solution to the world's increasing food demands and diminishing arable land. By integrating agriculture with urban ecology, vertical farming conserves space and resources, a critical factor in densely populated urban areas [1]. Benke and Tomkins focused on reducing water usage, pesticide use, and carbon emissions from transportation due to proximity to urban consumers [2]. Despite these advantages, vertical farming faces significant challenges especially in energy consumption for artificial lighting, impacting its overall sustainability and cost-effectiveness [3].

Plant phenotyping plays a vital role in vertical farming by providing quantitative data on plants' physical and functional characteristics. This field is crucial for calibrating models using information extracted from plants [9]. Traditionally, gathering specific plant details and statistics has been labor-intensive, often requiring specialized instruments. Recent advancements in computer vision have significantly reduced manual efforts, enabling the acquisition of data measurements from plant images [10].

2.2 The Role of Light in Plant Growth

Light is a fundamental component of plant growth, with photosynthesis being the primary process impacted by light quality and intensity. According to Kozai, Niu, and Takagaki, different wavelengths of light influence various aspects of plant growth and development. Red and blue light spectra are known to enhance photosynthesis, a critical process for plant growth [4]. Bantis, Smirnakou and Ouzounis further explore how light quality affects plant morphology, signaling that light manipulation can optimize plant growth in controlled environments like vertical farms [5].

2.3 Energy Efficiency in Lighting

The choice of lighting technology is primary in vertical farming because it significantly impacts energy consumption. The use of energy-efficient lighting such as LEDs is often advocated as a solution. Barbosa demonstrated that LEDs with their lower energy emission and longer lifespan are more cost-effective and sustainable for vertical farming [6]. However, identifying the most effective light wavelengths for plant growth while maintaining energy efficiency remains a challenge.

2.4 Image Analysis and Machine Learning in Agriculture

Recent advancements in technology have paved the way for the integration of image analysis and machine learning in agriculture. PlantCV, an open-source image processing software, has been instrumental in plant phenotyping, providing tools for analyzing plant traits from images [7]. Machine learning models, as explored by Liakos, have shown promise in predicting plant growth and health, offering a data-driven approach to precision agriculture [8].

Gehan made significant updates to an open-source software called PlantCV for plant phenotype analysis. They integrated new tools and methods for image processing, data handling, and machine learning which enhanced the software's capabilities in analyzing multiple plants, leaf segmentation, and morphometrics [11].

Research by Montes and Martin was focused on feature selection in regression models to predict energy consumption in LED lights with specific light recipes in Closed Plant Production Systems (CPPS). Their research emphasized the importance of integrating AI methods for predicting crucial variables like light color and specific wavelengths for cost-effective and sustainable CPPS [12].

2.5 Application in Indoor Vertical Farming

Bhama Krishna Pillutla conducted research on developing a non-intrusive mechanism using image analysis and AI techniques to monitor plant growth in indoor vertical farms. Their study, which collected data across three growth cycles of basil plants, demonstrated precise tracking of leaf growth using techniques like centroid tracking and intersection comparison [13].

2.6 Optimizing Light Conditions for Plant Growth

Our study builds upon the existing body of research by focusing on optimizing light conditions for plant growth in vertical farming using image analysis and machine learning techniques. By examining the effects of different light wavelengths on plant growth, we aim to identify the most effective combinations that promote growth while being energy efficient. This endeavor aligns with the global shift towards sustainable agricultural practices, addressing both the challenges of urbanization and the pressing need for environmental conservation.

3. Data

3.1 Data Collection Process

Our study on optimizing light wavelengths for plant growth in vertical farming was conducted in a regulated indoor environment that was set up by Professor Dr. Kevin B. Martin from the College of Engineering and Technology at Northern Illinois

University. The study was structured to capture comprehensive pictures of plants that are crucial for understanding the impact of light on plant growth.

1. **Vertical Farm Setup and Environmental Control**: The vertical farm consists of four separate trays; each contains a plant bed, and it is subjected to different light conditions. These sections were equipped with advanced LED light systems that are capable of precisely adjusting light intensity and spectrum, focusing largely on the red, green, and blue wavelengths of light. The light treatment for each tray was maintained consistently for an entire growth cycle that typically lasts for 5-7 days.
2. **Data Types and Collection Techniques**: The initial data that was collected included visual images to notice the growth area of the leaves. Cameras were installed at the top of each plant tray that captured these images regularly that provided a continuous visual record of the plant's development under different light conditions. In addition to the images, we also recorded the data on the combinations of different light wavelengths and their intensities for each individual tray. This data was important in correlating specific lights that observe plant growth patterns and health.
3. **Consistency in Environmental Conditions**: To ensure that light was the only variable to be considered to see the effects in plant growth, other environmental factors such as temperature, humidity, and water quality were kept constant across all trays. This control was vital in isolating the effects of light wavelengths from other potential growth factors.

3.2 Data Analysis Process

The data analysis phase was a fundamental process in deriving meaningful insights from the collected data. This process involved a combination of image processing, statistical analysis, and pattern recognition.

1. **Initial Data Processing**: Preprocessing was done on the raw image data, which included measures of plant development, readings from the environment, and images recorded at different times. This included cleaning the data, normalizing the measurements of data, and structuring it into an analyzable format. A Python-based script processed this raw data into a structured DataFrame making it the foundation for the following analyses.
2. **Image Analysis Using PlantCV**: The focus of our data analysis was the use of PlantCV for image processing. PlantCV allowed the extraction of quantitative data from visual images, such as plant location and growth patterns.

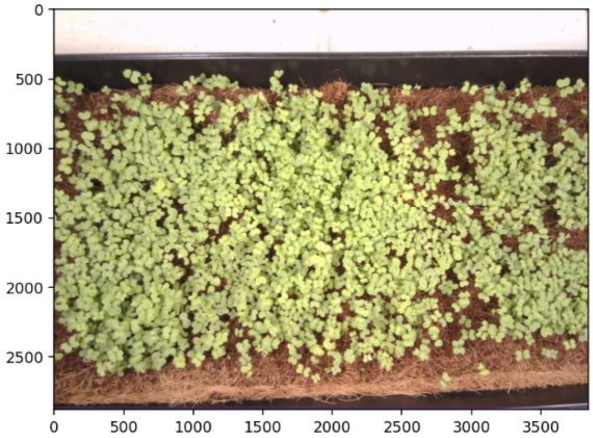


Figure 1: Initial image before image segmentation

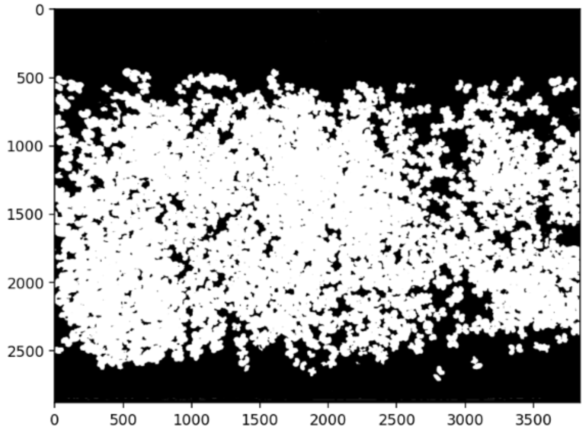


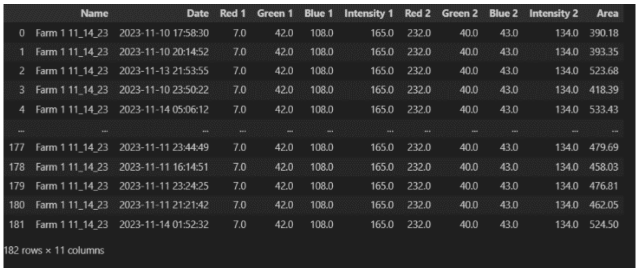
Figure 2: Segmented image after applying thresholding

3. **Statistical Analysis and Correlation**: After the data was preprocessed, we used statistical methods to explore relations between light conditions and plant growth metrics. The focus was on identifying patterns that can indicate how different light wavelengths influenced the plant growth, particularly in terms of leaf area.
4. **Comparative Analysis – Different Trays**: The study included data from four different trays, where each tray was exposed to unique light treatments. A comparative analysis was essential to extract information on the effects of light conditions on plant growth.
   - The analysis revealed that there might be a correlation between specific light wavelengths

and growth of the plant. For example, trays that were exposed to higher red and blue wavelengths showcased larger growth areas.

- The consistent growth in each tray in the cycle of harvest showed us that there is a positive impact of light conditions in indoor farming. However, the growths and the rate varied among different trays indicating some wavelengths were more conductive than others.
- Each tray's data provided unique insights. For instance, Farm 1 displayed the most considerable increase in growth area, suggesting that its specific light treatment was highly effective. In contrast, Farm 3, despite showing growth, had a relatively lower increase, indicating a less optimal light combination for the plant species grown in that tray.

5. **Insights from Data Analysis**: The culmination of our data analysis yielded significant insights. Certain light wavelengths were found to be more conducive to plant growth, enhancing both the rate and quality of growth.
   - **Growth Area Observations**: The visual image data, as illustrated in the outputs, showcased the growth areas of leaves under different lighting conditions. For instance, in Farm 1, the growth area measured 390.18 cm^2 initially and showed a significant increase to 524.50 cm^2, indicating robust growth under the specific light treatment administered in this tray.

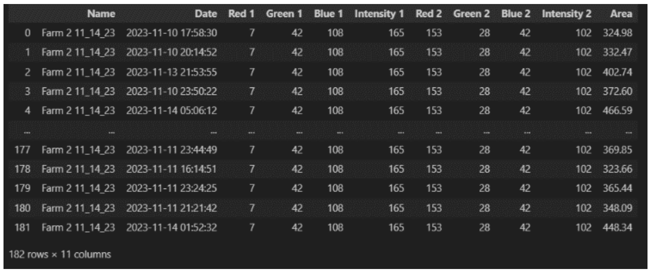


	Name	Date	Red 1	Green 1	Blue 1	Intensity 1	Red 2	Green 2	Blue 2	Intensity 2	Area
0	Farm 1 11_14_23	2023-11-10 17:58:30	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	390.18
1	Farm 1 11_14_23	2023-11-10 20:14:52	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	393.35
2	Farm 1 11_14_23	2023-11-13 21:53:55	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	523.68
3	Farm 1 11_14_23	2023-11-10 23:50:22	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	418.39
4	Farm 1 11_14_23	2023-11-14 05:06:12	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	533.43
...	...	...	...	...	...	...	...	...	...	...	...
177	Farm 1 11_14_23	2023-11-11 23:44:49	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	479.69
178	Farm 1 11_14_23	2023-11-11 16:14:51	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	458.03
179	Farm 1 11_14_23	2023-11-11 23:24:25	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	476.81
180	Farm 1 11_14_23	2023-11-11 21:21:42	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	462.05
181	Farm 1 11_14_23	2023-11-14 01:52:32	7.0	42.0	108.0	165.0	232.0	40.0	43.0	134.0	524.50

182 rows × 11 columns

Figure 3: Data frame containing information for Farm 1

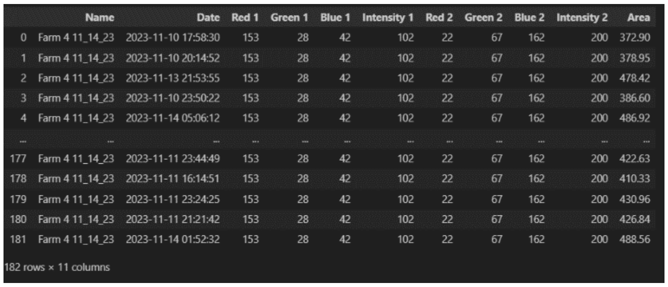
- **Comparative Growth Analysis:** Now we are comparing the growth areas across different farms that depict to us how varying light wavelengths influenced plant development. For example, Farm 2 started with a growth area of 324.98 cm^2, developing to 448.34 cm^2, while Farm 4 showed growth from 372.90 cm^2 to 488.56 cm^2. By running these variations, the study underlined the variations of different wavelength combinations that impacted the growth effectively.



	Name	Date	Red 1	Green 1	Blue 1	Intensity 1	Red 2	Green 2	Blue 2	Intensity 2	Area
0	Farm 2 11_14_23	2023-11-10 17:58:30	7	42	108	165	153	28	42	102	324.98
1	Farm 2 11_14_23	2023-11-10 20:14:52	7	42	108	165	153	28	42	102	332.47
2	Farm 2 11_14_23	2023-11-13 21:53:55	7	42	108	165	153	28	42	102	402.74
3	Farm 2 11_14_23	2023-11-10 23:50:22	7	42	108	165	153	28	42	102	372.60
4	Farm 2 11_14_23	2023-11-14 05:06:12	7	42	108	165	153	28	42	102	466.59
...	...	...	...	...	...	...	...	...	...	...	...
177	Farm 2 11_14_23	2023-11-11 23:44:49	7	42	108	165	153	28	42	102	369.85
178	Farm 2 11_14_23	2023-11-11 16:14:51	7	42	108	165	153	28	42	102	323.66
179	Farm 2 11_14_23	2023-11-11 23:24:25	7	42	108	165	153	28	42	102	365.44
180	Farm 2 11_14_23	2023-11-11 21:21:42	7	42	108	165	153	28	42	102	348.09
181	Farm 2 11_14_23	2023-11-14 01:52:32	7	42	108	165	153	28	42	102	448.34

182 rows × 11 columns

Figure 4: Data frame containing information for Farm 2



	Name	Date	Red 1	Green 1	Blue 1	Intensity 1	Red 2	Green 2	Blue 2	Intensity 2	Area
0	Farm 4 11_14_23	2023-11-10 17:58:30	153	28	42	102	22	67	162	200	372.90
1	Farm 4 11_14_23	2023-11-10 20:14:52	153	28	42	102	22	67	162	200	378.95
2	Farm 4 11_14_23	2023-11-13 21:53:55	153	28	42	102	22	67	162	200	478.42
3	Farm 4 11_14_23	2023-11-10 23:50:22	153	28	42	102	22	67	162	200	386.60
4	Farm 4 11_14_23	2023-11-14 05:06:12	153	28	42	102	22	67	162	200	486.92
...	...	...	...	...	...	...	...	...	...	...	...
177	Farm 4 11_14_23	2023-11-11 23:44:49	153	28	42	102	22	67	162	200	422.63
178	Farm 4 11_14_23	2023-11-11 16:14:51	153	28	42	102	22	67	162	200	410.33
179	Farm 4 11_14_23	2023-11-11 23:24:25	153	28	42	102	22	67	162	200	430.96
180	Farm 4 11_14_23	2023-11-11 21:21:42	153	28	42	102	22	67	162	200	426.84
181	Farm 4 11_14_23	2023-11-14 01:52:32	153	28	42	102	22	67	162	200	488.56

182 rows × 11 columns

Figure 5: Data frame containing information for Farm 4

The outcome of our rigorous data collection and comprehensive analysis gave us valuable insights into the optimal light conditions for plant growth and health in a vertical farming setup. The integration of superior image processing with statistical evaluation enabled a detailed knowledge of plant responses to the light, paving the way for extra green and sustainable agricultural practices. This research holds sizeable implications for the future of vertical farming, where precision and performance are key to maximizing productivity in confined areas.

4. Methodology

4.1 Regression

Regression analysis involves finding the connections between variables. When using regression, we examine a particular phenomenon using multiple observations. Each observation contains various characteristics. The idea is to establish a relationship among these characteristics based on the assumption that at least one of them is influenced by the others [19]. The objective is to discover a function that effectively relates certain variables to others. The variables that are influenced are referred as dependent variables, outputs, or responses. On the other hand, the variables that are considered to influence the dependent variables are termed independent variables, inputs, regressors, or predictors [19].

Typically, regression problems involve predicting a continuous and unbounded outcome variable using different types of input data. These inputs can take the form of continuous values, discrete numbers, or even categorical data such as gender, nationality, or brand preferences. In simpler terms, regression analysis is about understanding how various factors might affect or relate to each other and using that understanding to predict or explain certain outcomes. It's a method for uncovering connections and patterns in data to help explain or forecast future behaviours and trends. We will use this approach to answer whether and how light and plant growth are related and useful when you want to forecast a response using a new set of predictors.

4.1.1 Linear Regression

Linear regression aims to establish a relationship between a dependent variable (light) and one or more independent variables (Plant growth) by fitting a linear equation to the observed data. Linear regression serves as a foundational tool in statistics, machine learning, and various fields for understanding and modelling relationships between variables. The primary goal is to find the coefficients that minimize the difference between the observed and predicted values. The general equation for a simple linear regression with one predictor variable is:

$$Y=\beta 0 + \beta 1X + \varepsilon$$

Here, Y is the dependent variable, X is the independent variable, $\beta 0$ is the intercept, $\beta 1$ is the coefficient for the independent variable, and ε represents the error term accounting for unexplained variability.

4.1.2 Multiple Linear Regression:

Multiple linear regression emerged from the need to understand more complex relationships between dependent and multiple independent variables that is likely an upgrade from linear regression. While simple linear regression allowed for modelling relationships between two variables, it became evident from our study that real-world scenarios involve multiple factors influencing an outcome.

The history of multiple linear regression started from the pioneering work of Francis Galton, Karl Pearson, and Ronald Fisher in the late 19th and early 20th centuries. Galton explored regression and the concept of fitting a line to a scatter plot of data points. Pearson introduced the method of least squares, which formed the basis for fitting linear regression models. However, these early developments primarily focused on relationships between two variables only.

By allowing for the addition of multiple predictors, multiple linear regression provided a more detailed understanding of the relationships between various factors and the outcome of interest. This modelling technique gained popularity across various fields due to its ability to capture the multidimensional nature of real-world scenarios, enabling more accurate predictions and deeper insights into complex systems. Over time, with advancements in computing power and statistical methodologies, multiple linear regression has become a pioneer in data analysis, playing a crucial role in research, prediction, and decision-making across diverse disciplines.

In multiple linear regression, the model extends to accommodate multiple predictors:

$$Y= \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon$$

In this equation, Y is still the dependent variable as it was in simple linear regression, and $X1$, $X2$, …, Xp represent multiple independent variables. The process involves using the method ordinary of least squares, which minimizes the sum of squared differences between observed and predicted values. The model estimates the coefficients by adjusting them iteratively until it finds the values that minimize the overall error. These coefficients represent the slope of the line for each independent variable and the intercept term in the case of multiple linear regression. The fitted model allows for predicting the dependent variable based on new values of the independent variables, enabling insights into relationships, and making predictions within the given data context.

4.2 Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The objective is to have model that predicts the value of a target variable by learning simple decision rules gained from the data features. A tree can be seen as a piecewise constant approximation. Decision trees learn from data with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model. [20]

The process involves iteratively partitioning the data based on input features, creating a hierarchical structure resembling a tree. At each step, the data is divided into subsets using specific features. This establishes nodes representing attribute tests, where each branch signifies the outcome of a test. The endpoints of these branches, termed as leaf nodes, present either a class label or a predicted value for the given input.

4.2.1 Math behind the model:

To consider a classification task, let us consider a dataset with *N* samples and *K* features, a decision tree algorithm constructs a series of binary decision rules to partition the data. The decision tree builds a tree structure where each node corresponds to a feature and a threshold value.

At each node *m* in the tree:

- A feature j and a threshold value t are selected to split the data into two subsets based on a condition: Xjm ≤ t (left branch) and Xjm > t (right branch), where Xjm represents the value of feature j at node m.
- The splitting criterion is typically chosen to optimize a measure of node impurity, such as **Gini impurity** or **information gain** (entropy), to maximize the homogeneity of the classes in the resulting subsets.

This process continues recursively until a stopping condition is met. The tree structure is essentially a series of if-else statements based on feature values, leading from the root node to the leaf nodes.

4.2.2 Explanation:

Root Node:

- The first node of the tree is the root node, representing the entire dataset.
- It selects the feature and threshold value that best splits the data into two subsets based on a chosen information gain.

Internal Nodes:

- Following the root node, each internal node represents a test condition on a feature.
- It further partitions the data based on the feature value and threshold selected to minimize impurity.

Leaf Nodes:

- Leaf nodes are reached when the splitting process stops.
- Leaf nodes contain the final decision, representing the predicted class for classification tasks or the predicted value for regression tasks.

4.2.3 Some advantages of decision trees are:

- Simple to understand and easy to interpret. Trees can be visualized.
- Able to handle both numerical and categorical data.
- Able to handle multi-output problems.

4.2.4 The disadvantages of decision trees include:

- Sometimes, decision trees, they can get too detailed and only work well for the specific data used to create the model. This is called overfitting. To overcome this, Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

To have better accuracy on the model of our research, we went a step further to use random forest algorithm.

4.3 Random Forest

In 2001, Leo Breiman and Adele Cutler introduced the Random Forest algorithm as a solution to address the limitations of decision trees. It builds on the idea of bagging by creating an ensemble of decision trees. However, it introduces randomness during both the training process and the feature selection.

A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. At each node of each tree, a random subset of features is considered for splitting. This randomness ensures that each tree in the forest is different from the others.

4.3.1 Key Features of Random Forest:

Random Forests train multiple decision trees on random subsets of the training data (bootstrap samples) and at each node of each tree, they consider only a random subset of features for splitting. This randomness helps in decorrelating the trees, reducing overfitting.

- **Reduced Overfitting:** By averaging the predictions of multiple decision trees and introducing randomness, Random Forests tend to generalize better to unseen data compared to individual decision trees.
- **Improved Accuracy:** They often produce more accurate results where the combined predictions of multiple trees tend to be more reliable than any single tree.
- **Feature Importance:** Random Forests provide measures of feature importance, allowing insights into which features are more influential in making predictions.

Overall, the Random Forest algorithm emerged as a powerful ensemble learning method that overcomes the limitations of individual decision trees, providing improved accuracy, robustness, and insights into feature importance. Its development was a response to the need for more reliable and accurate predictive models in machine learning.

4.3.2 Random Forest Regression

In our research study we used random forest classifier which involves statistical techniques such as bootstrapping, random feature selection, and aggregation of predictions to create a robust and accurate ensemble model. The mathematics largely revolves around decision tree construction, handling randomness in feature selection, and combining predictions from multiple trees.

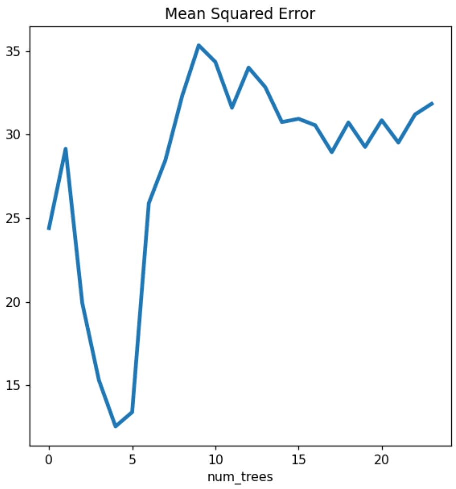


Figure 6: Determine the optimal n_estimators using the MSE metric

4.4 Artificial Neural Networks (ANN)

Artificial neural networks are information processing structures providing the connection between input and output data by artificially simulating the physiological structure and functioning of human brain structures.[13] The natural neural network, instead, consists of a very large number of nerve cells (about ten billion in humans), said neurons, and linked together in a complex network. The intelligent behaviour is the result of extensive interaction between interconnected units. The input of a neuron is composed of the output signals of the neurons connected to it. When the contribution of these inputs exceeds a certain threshold, the neuron through a suitable transfer function generates a bioelectric signal, which propagates through the synaptic weights to other neurons.

The ANN simulate a model like the following:

- The parallel processing, like the neurons that process the information simultaneously.
- The bi-fold (two functionalities in one) function of the neuron, that acts simultaneously as memory and signal processor.
- The distributed nature of the data representation, i.e. knowledge is distributed throughout the network.
- The whole network's ability to learn from experience.
- An artificial neural network captures this attitude in an appropriate "learning" stage.

4.4.1 Structure of Neural Networks

Artificial neural networks are composed of elementary computational units called neurons combined according to different architectures. Meaning, they can be arranged in layers (multi-layer network), or they may have a connection topology [18].

Layered networks consist of:

- Input layer, made of P neurons (one for each network input).
- Hidden layer, composed of one or more intermediate layers consisting of Q neurons.
- Output layer, consisting of R neurons (one for each network output).

The feedback architecture, with connections between neurons of the same or previous layer; The feedforward architecture without feed- back connections (signals go only to the next layer's neurons) [15].

Each neuron receives *N* input signals *Xi* (with connection weights *Wi*) which sum to an "activation" value *y*. A suitable activation function *F* transforms it into the output *F(y)*. The operational capability of a network is contained in the connection weights, which assume their values from the training phase.

4.4.2 Architecture and Models of Neural Networks

In 1943 American neurophysiologist and cybernetician of the University of Illinois at Chicago Warren McCulloch and self-taught logician and cognitive psychologist Walter Pitts published the first mathematical model of a neural network. Neural network applications can be grouped into three major areas [16]:

1. Classification.
2. Time series forecasting.
3. Function approximation.

4.4.3 Training & Building an ANN Model

The neural network isn't directly programmed; rather, it undergoes training via a learning algorithm to predict the outcome. This training process represents the essence of "learning through experience." The learning algorithm shapes the neural network's specific configuration, thus governing and defining the network's capability to offer accurate solutions to specific problems. The network's ability to provide correct answers is reliant upon the conditions set by the learning process and the resultant network configuration.[17]

4.4.4 Network Parameters

The construction of a neural network necessarily involves some steps requiring us to set the appropriate parameters. To get an optimal training and network implementation, it is necessary to divide the dataset into subsets, which determine the learning environment: training and validation set. In essence, the network learns by trying to recognize the dynamics of the training set, checks how it fits on the test set and then applies itself to a set of data (validation dataset) never observed before. There are no universally valid rules for the subdivision of the dataset: the solutions adopted are (60% − 20% − 20%) and (60% − 30% − 10%) respectively for the training, test and validation set.

4.4.5 Number of Hidden layers and Neurons

We found Several studies utilize a single hidden layer in neural networks, finding it adept at accurately approximating complex, nonlinear functions. However, this approach often demands many neurons, potentially limiting the learning process. Alternatively, networks featuring two hidden layers prove more effective, especially for forecasting high-frequency data. This choice, backed by specific theories and practical experience, demonstrates that employing more than two hidden layers seldom leads to substantial enhancements in network performance.

It's worth noting that an excessive number of neurons can result in overfitting, where neurons fail to generate reliable predictions due to an overtraining on specific inputs. This situation mirrors the network memorizing correct answers without the ability to generalize. Conversely, an insufficient number of neurons hampers the network's capacity to learn effectively. Striking the right balance is crucial to ensure optimal learning without falling into the traps of overfitting or underfitting.

The formula proposed in literature is [15]:

$$h =(n+m)2 +t12$$

where:

h = number of hidden neurons;
n = number of input neurons;
m = number of output neurons;
t = number of observations in the training set.

The empirical results show that none of these rules appears generalizable to every problem, although a not meaningless number of positive results seem to prefer.

4.4.6 Learning Rules & Connecting Weights

Once the initial properties of the neural network are set, determining the stopping criteria becomes crucial. For networks intended for evaluating learning on the test set is recommended. Conversely, for accurately describing the studied condition, assessing learning on the training set is preferable.

Learning parameters are interlinked with the network's error indicators, such as average error, maximum error, and the frequency of instances where there's no improvement in error. These indicators serve as guides in understanding the network's performance and adjusting the learning parameters for optimal results.

The next problem is the choice of the learning rate: in particular, it is necessary to decide the rate at which the network changes weights compared to error. The learning rate η is normally initially set to the value:

$$\eta = [\max(x) - \min(x)]N$$

where x represents the training set with N input records. Through the compensation for the input values x and relative number N this initial value of η should be fine for many classification problems, regardless of the number of training samples and their values' range. However, although highest values of η may accelerate the learning process, this could induce oscillations that can slow convergence.

In another approach, the inclusion of learning rate in training the neural network entails adjusting a fraction of the prior weight change into the new weight adjustment. This nuanced addition allows the network to learn at an accelerated pace while averting abrupt fluctuations. By integrating a substantial portion of the previous weight alterations, this mechanism fosters a steadier learning path.

The iterative process of fine-tuning the network's parameters involves a series of attempts that may yield substantially varied outcomes.

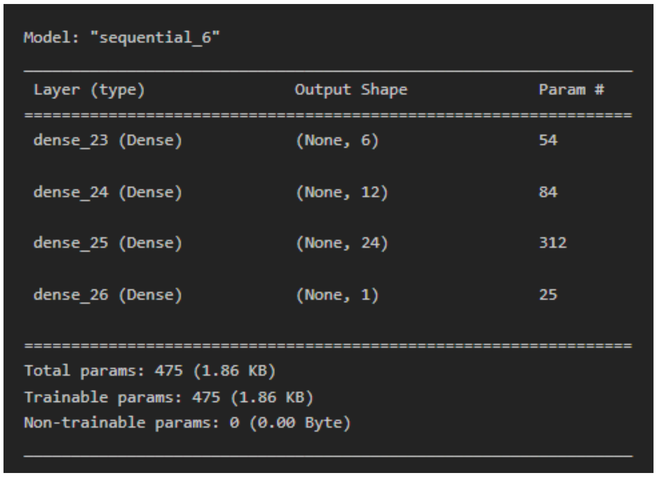
4.4.7 Code and Our Algorithm

The necessary libraries to implement our neural network will involve importing the TensorFlow and Keras library. From that library, we will use the dense and sequential classes to create the architecture of our model. The next step will be splitting the data into target variable and predictors. The predictors are the three color channels (RGB) and their intensity while the target is the plant area.

This is followed by standardizing the data that is referred to as a data pre-processing step where we used a 'standardscaler' method from our 'sklearn' module. Then for training and building the ANN model using the test dataset of 20% data and training dataset of 80% by splitting the data.

Then we have defined input layer and our first hidden layer followed by second layer of the model with dimensionality of '8' and activation function of Rectified Linear Unit for the first and second hidden layer and then for the third layer the activation function 'tanh'.

Following steps were to train our ANN model which has a structure of three hidden layers with respective activation functions and optimizer as 'adam' trained for each iteration of batch size 20 and the number of iterations is 100. The final design of our network has a total of 475 parameters and made a prediction on the test data with the following results predicting an area value against ground truth.



```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_23 (Dense)            (None, 6)                 54

 dense_24 (Dense)            (None, 12)                84

 dense_25 (Dense)            (None, 24)                312

 dense_26 (Dense)            (None, 1)                 25

=================================================================
Total params: 475 (1.86 KB)
Trainable params: 475 (1.86 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 7: Framework of our Neural Network

5. Results

Our result attempts to identify the optimal light wavelength for plant growth within vertical farming. We fed 28 different combinations of light channels at an intensity of 150 for each light into our Random Forest model as shown in *Figure 9*. From the output, we learned that the ideal light wavelength was hyper red and true blue, since it produced the largest plant area (414.31 cm²) out of all the other light combinations. We also notice that the second and third highest areas all involve hyper red as one of the input light channels.

Another unit we tracked was the change in growth between each image for each farm as presented in *Figure 9* and *Figure 10*. For the first cycle, the farm that had the largest change in area growth was Farm 3 at 125.49% between Nov 7th from 12:01 am to Nov 7th 1:58 pm. The wavelengths utilized in that tray were true blue and true red. In the second iteration, the farm that had the largest change in growth was Farm 3 at 99.57% from Dec 5th from 3: 56 am to Dec 5th at 7:53 pm. The light combinations in this part were amber and 5k white light.

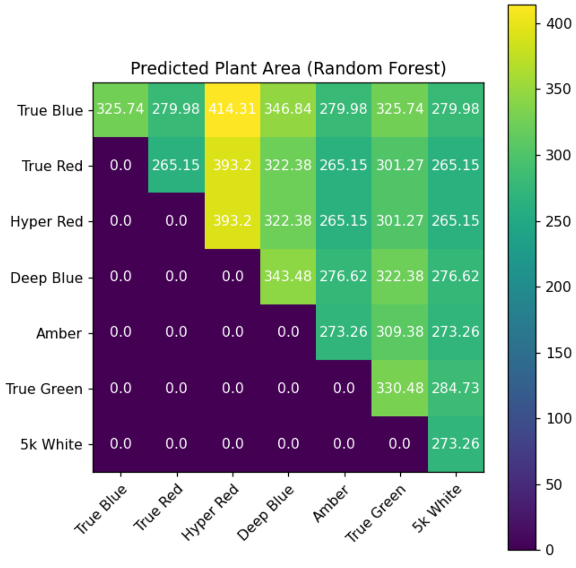


Figure 8: Predicted area using 10 light channel combinations

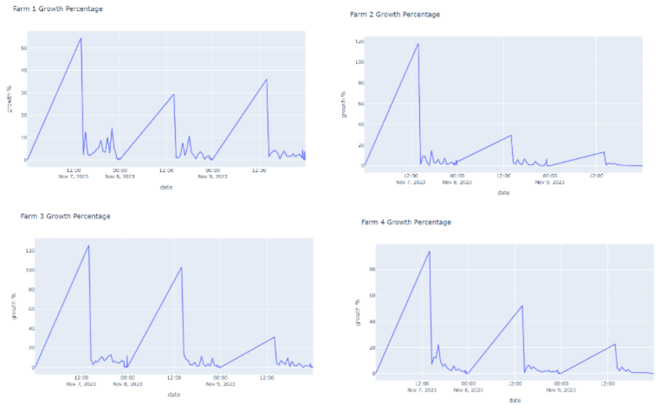


Figure 9: The growth % between each image for the first iteration

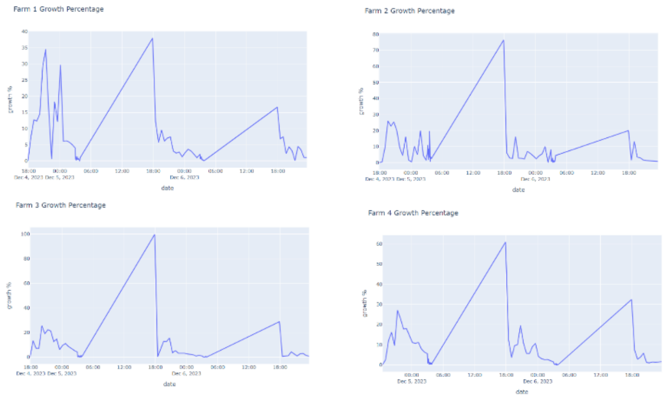
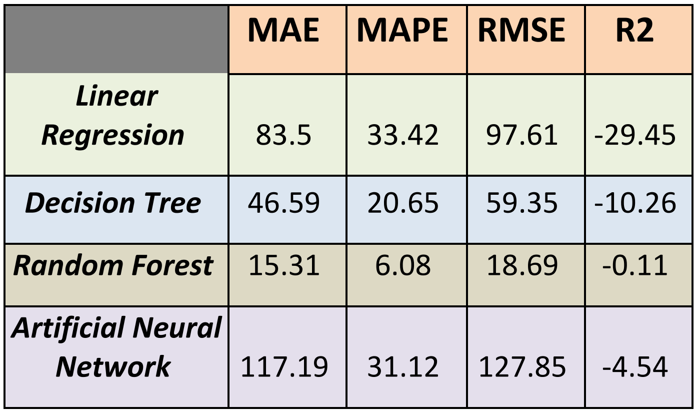


Figure 10: The growth % between each image for the second iteration



	MAE	MAPE	RMSE	R2
Linear Regression	83.5	33.42	97.61	-29.45
Decision Tree	46.59	20.65	59.35	-10.26
Random Forest	15.31	6.08	18.69	-0.11
Artificial Neural Network	117.19	31.12	127.85	-4.54

Figure 11: Model Evaluation Metrics

6. Evaluation of Predictive Models Using Key Performance Metrics

In our analysis, we employed several metrics as shown in *Figure 11* to evaluate the performance of different predictive models - Linear Regression, Decision Tree, Random Forest, and Artificial Neural Network (ANN) - in optimizing light wavelengths for plant growth in vertical farming. Below is an explanation of each metric used and the comparative performance of the models.

6.1 Mean Absolute Error (MAE)

MAE is a measure of the average magnitude of errors in predictions, disregarding their direction. It represents the average distance between the predicted and actual values, offering a clear insight into the model's prediction accuracy. Lower MAE values indicate a model's higher accuracy in predictions. The Random Forest model demonstrated superior performance with the lowest MAE of 15.31, indicating its high accuracy in predicting plant growth under various light conditions.

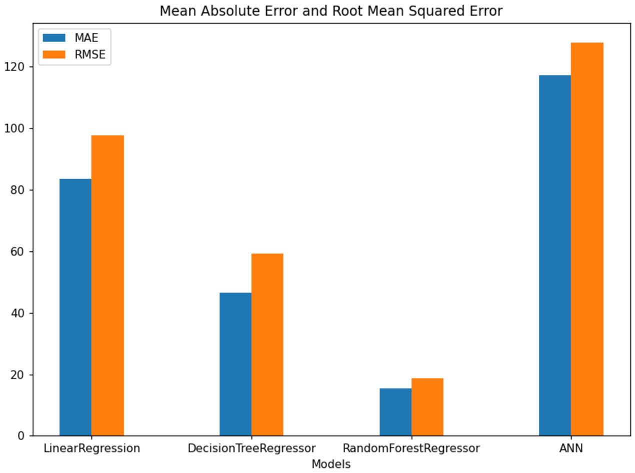


Figure 12

6.2 Root Mean Squared Error (RMSE)

RMSE calculates the square root of the average of the squared differences between the predicted and actual values. This metric is sensitive to large errors, making it a stringent measure of prediction accuracy. A model with a lower RMSE is considered to have a better fit. The Random Forest model outperformed others with an RMSE of 18.69, indicating that on average, its predictions deviated least from the observed values.

6.3 Mean Absolute Percent Error (MAPE)

MAPE expresses prediction accuracy as a percentage, calculating the absolute error relative to the true values. This metric is particularly useful for comparing the accuracy of models across datasets with different scales or magnitudes. A lower MAPE value denotes a more accurate model. Here again, the Random Forest model excelled with the lowest MAPE of 6.08%, suggesting its consistent reliability in proportion to the actual data.

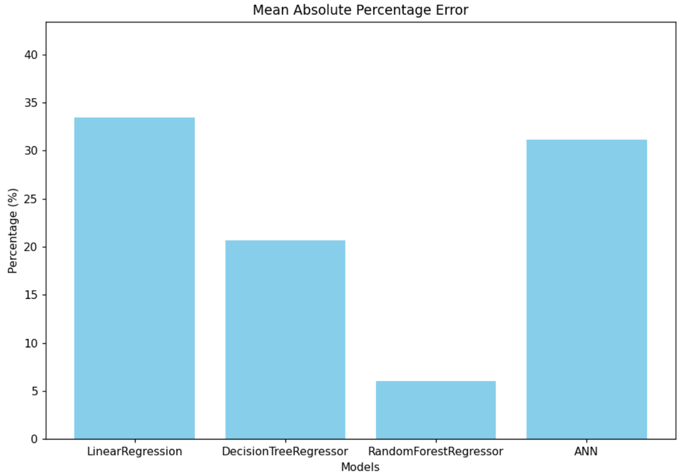


Figure 13

6.4 R^2 (Coefficient of Determination)

R^2 quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It assesses the goodness of fit of a model, where a higher R^2 value (closer to 1) indicates a stronger correlation between observed and predicted values. Although all models exhibited negative R^2 values, the Random Forest model had the least negative value at -0.11. This suggests that despite the challenging nature of the data, the Random Forest model had a comparatively better fit than the others.

6.5 Conclusion on Model Selection

Upon considering all the evaluated metrics - MAE, MAPE, RMSE, and R^2- the Random Forest model consistently demonstrated the most reliable and accurate performance. Its capacity to effectively handle complex, non-linear relationships between multiple variables in the dataset makes it the optimal choice for predicting plant growth responses to varying light wavelengths in vertical farming scenarios.

7. Discussion

This research aimed to optimize light wavelengths for enhancing plant growth in vertical farming setups, employing various machine learning models. The study's findings, drawn from the analysis of these models, provide several insightful conclusions and raise important considerations for future research and practical applications in the field of vertical farming.

7.1 Insights from Model Analysis

Random Forest's Superiority: The Random Forest model consistently outperformed others across multiple metrics. Its ability to manage the intricacies of the dataset, particularly non-linear relationships, and interactions among variables, was notably effective. This suggests that ensemble methods, which combine multiple models to improve prediction accuracy, are highly suitable for complex agricultural data.

Challenges in Model Fitting: The negative R^2 values across all models highlight the challenges inherent in modeling natural growth processes, which are influenced by a multitude of factors, some of which may not have been fully captured in the dataset. This underscores the need for more comprehensive data collection and possibly the development of more sophisticated models tailored to the complexities of plant biology and environmental interactions.

7.2 Implications for Vertical Farming

Energy Efficiency: The study’s focus on optimizing light wavelengths is directly linked to energy efficiency in vertical farms. The ability to determine the most effective light conditions for plant growth can lead to significant energy savings, a crucial factor in the sustainability and economic viability of vertical farms.

Precision Agriculture: The application of machine learning in determining optimal growing conditions is a step towards precision agriculture in vertical farming. This approach allows for fine-tuning environmental factors, leading to more efficient resource use and potentially higher yields.

7.3 Future Directions

Broader Data Collection: Future studies should aim to collect a more diverse range of data, covering various plant species, growth stages, and environmental conditions. This would help in building more robust models that can generalize better across different vertical farming scenarios.

Technological Advancements: The integration of more advanced image analysis and sensor technologies could provide richer datasets. This might include more detailed phenotypic data, real-time monitoring of environmental conditions, and even automated adjustments to lighting and other factors based on model predictions.

In conclusion, this study contributes to the growing body of knowledge in vertical farming and demonstrates the potential of machine learning in revolutionizing agricultural practices. By continuously refining our models and approaches, we move closer to achieving sustainable, efficient, and productive agricultural systems that can meet the challenges of a rapidly changing world.

8. Conclusion and Future Work

The concept of vertical farming will continue to be prevalent in the sector of agriculture and will eventually be a necessary portion of food production due to external factors like global warming. For this reason alone, the necessity to determine the optimal parameters for plant growth is critical to continuing innovation and investment into this area. As discussed in this paper, an important feature is determining the optimal light settings for plant growth. This means finding the correct combination of color and intensity is found to maximize the plants growth area, health, growth rate, and food quality while also maintaining the lowest energy consumption possible. All these factors together will significantly increase productivity and reduce the amount of energy and resources needed to create a vertical farm.

With this reasoning in mind, there is still significant work that needs to be done to provide accurate and generalizable results for the light settings. Most notable is the lack of data that was available at the time of this report. The first step that needs to be taken to generalize the results provided in this study is to give more data instances using different light settings to give a wider range that the models can train off. Once given, the models can be more accurately fine-tuned for this specific task which will in turn provide better results. In addition to more datasets broadly, additional data fields over the course of the growth cycle. One such data measurement is NDVI images which can be used to analyze plant health. Another would be using cameras along the sides of the trays which can help with monitoring plant growth in terms of height and not just area. Also, the analysis of the overall quality of the food product at the end of the trial would also be useful. These three measurements alone would be invaluable in adding additional feedback to the model for additional accuracy.

As a continuation of this research, the last piece that should be mentioned for future endeavors would be the implementation of a better image recognition software. The PlantCV system used in this study is perfect for the limited scope that is offered here but it lacks some important traits. The first being the lack of options regarding image classification. Currently, the software allows for edge detection which is sufficient at getting an accurate estimate of the total growth area of the plants. What it lacks is the accounting for overlapping leaves which means there is an extra area that is not traceable. A more advanced image classification software or library would be valuable to create a better picture of what is occurring not just on the surface, but for each individual plant that it could see even from a partial glimpse. As a first attempt at creating the tools necessary to further research in the area of vertical farming, this paper identifies some of the nuances that come with this topic and hopes to extend the lessons learned here to future work.

References

[1] Despommier, D. (2009). "The Vertical Farm: Feeding the World in the 21st Century."

[2] Benke Kurt, & Tomkins, Bruce. (2017). "Future food-production systems: Vertical farming and controlled-environment agriculture." https://doi.org/10.1080/15487733.2017.1394054

[3] Specht, K., et al. (2014). "Urban agriculture of the future: An overview of sustainability aspects of food production in and on buildings." DOI:10.1007/s10460-013-9448-4

[4] "Plant Factory: An Indoor Vertical Farming System for Efficient Quality Food Production." By Toyoki Kozai , Genhua Niu , Michiko Takagaki

[5] Bantis, F., Smirnakou, S., & Ouzounis, T. (2018). "Artificial light sources in plant growth and development." doi: 10.3390/plants12051075

[6] Barbosa, G. L., (2015). "Comparison of land, water, and energy requirements of lettuce grown using hydroponic vs. conventional agricultural methods." doi: 10.3390/ijerph120606879

[7] Fahlgren, N. (2015). "A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in Setaria." https://doi.org/10.1016/j.molp.2015.06.005

[8] Liakos, K. G., (2018). "Machine learning in agriculture: A review." https://doi.org/10.3390/s18082674

[9] “Image-Based Plant Phenotyping.” Plant-Phenotyping - The University of Nottingham, www.nottingham.ac.uk/research/groups/cvl/projects/plant-phenotyping/plant-phenotyping.aspx#:~:text=Plant%20phenotyping%20is%20a%20rapidly,and%20functional%20properties%20of%20plants. Accessed 17 Nov. 2023.

[10] Pietrzykowski, S. K., & Wymysłowski, A. (n.d.). Application of the OpenCV library in indoor hydroponic plantations for automatic height assessment of plants. https://sciendo.com/pdf/10.14313/jamris-2-2022-16

[11] Gehan, Malia A., et al. “PlantCV V2: Image Analysis Software for High-Throughput Plant Phenotyping.” PeerJ, PeerJ Inc., 1 Dec. 2017, peerj.com/articles/4088/.

[12] Montes Rivera, Martín, et al. “Feature Selection to Predict LED Light Energy Consumption with Specific Light Recipes in Closed Plant Production Systems.” MDPI, Multidisciplinary Digital Publishing Institute, 9 June 2022, www.mdpi.com/2076-3417/12/12/5901.

[13] Pillutla, Bhama Krishna. “BhamaPillutla.pdf.” Dropbox, May 2022, www.dropbox.com/s/9d9aayl29almak5/BhamaPillutla.pdf?dl=0. Accessed 20 Nov. 2023.

[14] “A Logical Calculus of the ideas Imminent in Nervous Activity,” describing the “McCulloch - Pitts neuron

[15] Multilayer feedforward networks are universal approximators by Kurt Hornik, Maxwell Stinchcombe, Halbert White

[16] Neural Information Processing. Models and Applications 17th International Conference, ICONIP 2010, Sydney, Australia, November 21-25, 2010, Proceedings, Part II

[17] Haykin, S. (1999) Neural Networks—A Comprehensive Foundations. Prentice-Hall International, New Jersey.

[18] Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Su-Hyun Han, Ko Woon Kim, SangYun Kim, Young Chul Youn doi: 10.12779/dnd.2018.17.3.83

[19] Alan O. Sykes, "An Introduction to Regression Analysis" (Coase-Sandor Institute for Law & Economics Working Paper No. 20, 1993).

[20] “Decision trees: from efficient prediction to responsible AI”. Hendrik Blockeel, Laurens Devos, Benoît Frénay, Géraldin Nanfack, Siegfried. - https://doi.org/10.3389/frai.2023.1124553