Classification of Tomatoes growth degree from Spectrum hue information measured Applying RBF-Kernel method

1st Yoshitsugu Nakagawa
Electronics Technology Group, Tama-Techno Plaza
Tokyo Metropolitan Industrial Technology Research Institute
Tokyo, Japan
nakagawa.yoshitsugu@iri-tokyo.jp

2nd Hiroyasu Sano
Electronics Technology Group, Tama-Techno Plaza
Tokyo Metropolitan Industrial Technology Research Institute
Tokyo, Japan
sano.hiroyasu@iri-tokyo.jp

3rd Toshiro Takata
Nozomi Corporation
Tokyo, Japan
takata@nozomicorp.jp

***Abstract*— In recent years, efforts have been made in the agricultural field to improve crop quality and streamline agricultural management by collecting and visualizing various data on crop growth. In this study, we focus on the "spectrum of light" as information on sunlight that is essential for crop growth and aim to capture its functions. Specifically, a spectral sensor is used to measure the reflectance of light for each wavelength. Measurement data generally contains variations, making classification difficult using machine learning based on linear separation. In this study, we attempt classification using a nonlinear separation method capable of high-dimensional analysis and report its effectiveness. Crop growth varies depending on location, lushness, etc. Traditionally, this work relied mainly on experience and subjectivity, but by quantitatively evaluating the distribution and changes in the visible light band, farmers can visually grasp the growth status of crops and determine the optimal harvest time.**

***Keywords*— Nonlinear SVM, RBF-Kernel, Visible light spectrum, Photomorphogenesis information, Tomatoes growth degree**

I. INTRODUCTION

In recent years, the adoption of IoT across various industries has been progressing. As a first step, efforts to visualize production activities and enhance management efficiency through academic research and pilot experiments have been reported.

In the agricultural sector, various technologies such as remote monitoring using IT, UAV-based pesticide spraying, unmanned transport vehicles to reduce labor, and modeled plant factories simulating natural environments are being introduced. These innovations continue to contribute to the productivity improvement of relatively large-scale agricultural businesses. Research and practical experiments related to agricultural IT for crop growth are diverse. Particularly, by utilizing AI and machine learning, it is expected that environmental control information collected and stored through ubiquitous environment control systems (UECS) integrated with numerous sensors can be used to optimize crop growth, shorten cultivation periods, achieve higher yields, and ensure stable growth.

Traditionally, farmers have relied on their long-standing experience, subjective judgment, and physical labor—such as irrigation, fertilization, and pesticide application—based on climate fluctuations and past knowledge, investing significant time to maintain crop quality through visual inspection and manual oversight. The utilization of IT and AI technologies will soon enable us to detect subtle, unseen changes in crop growth and take proactive actions. This will lead to multifunctional, automated environmental control systems.

Additionally, techniques using hyperspectral cameras mounted on UAVs have been reported, allowing analysis of the color distribution of crops in the field to understand growth trends and pest damage [1][2]. These methods enable the visualization of large-scale fields, including greening zones and water sources, and help assess environmental deterioration due to climate change and vegetation health over wide areas. Moreover, such macro-level data contributes to more accurate harvest forecasts.

On the other hand, the costs associated with implementing UECS and increased energy consumption for environmental control pose trade-offs with the goal of improving agricultural management efficiency. Excluding farm operators capable of absorbing these costs, many small-scale farmers are hesitant to adopt agricultural IT. In urban areas with dense residential neighborhoods and limited cultivated land nearby, it is crucial to focus on high-quality, high-value cultivation techniques that not only increase yield but also enhance sweetness and coloration to stabilize agricultural management. Furthermore, the decreasing number of new farmers and the aging farming population are regarded as urgent crises threatening food self-sufficiency and sustainability. Attracting new farmers and transferring technical skills are integral parts of deepening and researching agricultural IT—an important responsibility [3][4].

Historically, farmers have judged crop growth and environmental conditions based on their experience, know-how, and subjective perception, passing this knowledge to successors. To address this challenge, this research aims to develop technologies that use spectral light to quantify how well crops are growing via photosynthesis and nutrition, thus visually confirming and validating the conditions suggested by subjective impressions. Additionally, by accumulating and analyzing data, we hope to estimate the crop growth process and aid in future predictions. Focusing on the light spectrum necessary for crops to grow under sunlight, the system extracts wavelengths in the visible range (400–700 nm) from scattered light that possesses a broad spectrum. Multispectral sensors feature silicon filter lenses arranged in an array, each with different wavelengths, allowing measurement of light reflectance through the lenses. This enables capturing the distribution of crop growth based on spectral information.

In this paper, we sample the fruit parts of tomatoes as target crops and analyze the changes in spectral reflectance distribution that are invisible to the naked eye. Since tomatoes change color during growth through photosynthesis and eventually accumulate acidity and sweetness before harvest, quantifying the spectral distribution at this stage allows us to interpret the optimal harvest timing. In agricultural information involving nature, data uncertainty (variability) is often present. Variability can stem from individual differences in growth, measurement inconsistencies, and differences in sunlight reflectance caused by crop shapes. Conventional linear analysis methods struggle to categorize growth levels reliably under such variability [5][6]. This paper introduces a methodology using nonlinear classification methods as an alternative to linear approaches. While farmers have traditionally relied on visual cues such as color to judge fruit ripeness, using this quantified indicator could significantly assist new farmers. Furthermore, accumulating quantified data sets can be utilized for machine learning to classify the most appropriate harvest time based on future data samples.

II. PROTOTYPE FOR MEASUREMENT

A. Preparation of Equipment

The primary sources of crop growth are water and sunlight. Additionally, the weight, nutritional content, and texture of crops are determined by photosynthesis. Photosynthesis converts specific wavelengths of light energy into chemical energy necessary for plant growth, producing carbohydrates and oxygen from carbon dioxide. Among scattered light, the wavelengths effective for photosynthesis lie within the visible spectrum (400–700 nm). Recently, combining the visible spectrum with near-infrared (NIR) wavelengths has been recognized as an effective method to broadly visualize vegetation and water source distributions on Earth, providing an overview of crop growth conditions. The multispectral sensor used in this research adopts AMS technology and can measure the NIR band at 900 nm. This sensor can simultaneously quantify multiple visible light regions through digital processing, and a portable prototype has been developed.

Figure 1. Show prototype of equipment (Using Multi-spectral Sensor AS7341).

The concept of light morphological information refers to capturing not only visible colors but also the roles of information obtained from light. By extracting specific wavelength components from the broad spectrum of scattered light emitted from the sensor and observing the reflectance of crops, this method aims to characterize crop features and monitor changes during the growth process. The growth rate of crops varies depending on conditions such as the location of fruit on the plant and leaf density. Even when the same color appears to the human eye, analyzing the information contrast allows us to distinguish differences in the progress of growth. The prototype device captures eight wavelengths in the visible spectrum: 415, 445, 480, 515, 555, 590, 630, and 680 nm.

B. Need for Normalization

The amount of sunlight received from the sun varies with seasons and weather conditions. When measuring light intensity, it is necessary to adapt the measurement approach to these changes. The semiconductor used in the prototype adjusts exposure based on measurement time and aperture size, converting the reflected light into a measurable signal. If the aperture remains large and the sensor is exposed to strong light for an extended period, light reflectance can become saturated.

Conversely, in indoor environments where direct sunlight does not reach, increasing the aperture size enhances sensitivity. In controlled cultivation facilities, the amount of scattered light also fluctuates over time, so maintaining a fixed exposure and aperture may lead to saturation or low resolution at some wavelengths. Under such conditions, changing the exposure settings can cause variations in the measured light reflectance levels at the same wavelength, depending on the exposure conditions. To address this, normalization processing was applied to compare relative values across different wavelengths. The normalization formula is represented as Equation (1).

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

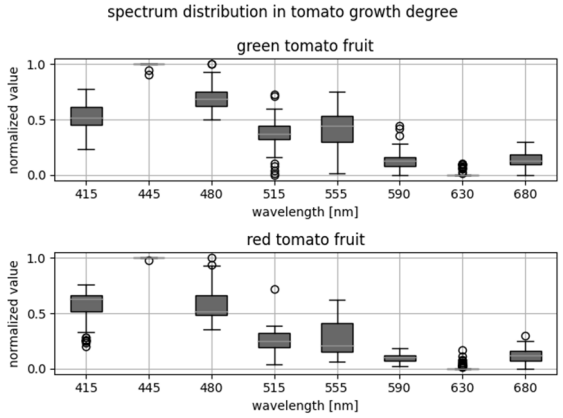

Figure 3. Show Spectral Distribution in Tomato growth degree.

Note: The units for both the vertical axes are [nm].

Furthermore, by incorporating an automatic exposure control algorithm, the system automatically adjusts the measurement range for each measurement, despite changes in exposure conditions, ensuring that measurements do not fall below the lower limit or reach saturation. This also ensures consistency in the relative distribution of light across wavelengths without contradictions. Traditionally, farmers have relied on visual cues, such as color, to judge fruit ripeness. However, using this quantified indicator can significantly aid new farmers. Additionally, the accumulated quantified data sets can be utilized for machine learning models to classify the optimal harvest timing based on future data samples.

III. METHODOLOGY

In this section, we explain how to use machine learning to tell different growth stages of tomatoes apart by analyzing how their light spectral patterns differ. We aim to construct a classification model for a group of data representing the distribution of wavelengths measured and normalized by light spectrum, which is plotted as multiple classes of point clouds on a two-dimensional plane. The method used to create this model is the Support Vector Machine (SVM). Geometrically, it is formulated by maximizing the non-interference zone (margin) M, which separates each point cloud from the boundary line that classifies the point clouds (equation). Furthermore, the data used in this study consists of sampled data, which is treated separately as training data and test data. The point x plotted on a two-dimensional plane and the parameters W and b are formulated in vector form (2) as follows:

For vectors

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})^T$$
$$\boldsymbol{w} = (w_{i1}, w_{i2})^T \quad (2)$$

$$W^T X_i + b = 0 \quad (i = 0,1,2,\cdots,N) \quad (3)$$

The separating hyperplane in n-dimensional space is expressed by equation (3).

Next, this paper applies a classification model that allows

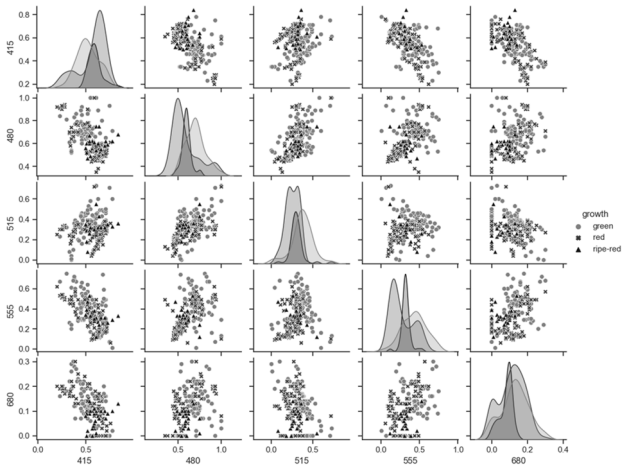

Figure 2. Show Spectral Distribution of Selected between Light Wavelength in Visible light regin as Pairplot.

Note: The units for both the vertical and horizontal axes are [nm].

$$\max_{w,b} M, \frac{t_i(W^T X_i + b)}{\|W\|} \gg M \ (i = 0,1,2,\cdots,N) \quad (4)$$

$$t_i(W^T X_i + b) \geq 1 - \xi_i \xi_i = \max\{0, M - \frac{t_i(W^T X_i + b)}{\|W\|}\} \quad (5)$$

A. Linear Machine Learning

Recently, artificial intelligence (AI) has become common in many fields. It can learn from past data automatically and make decisions or classifications without human help. Machine learning is a type of AI where computers learn from data to create models that can recognize patterns.

ξ_i units inside the margin equation (5). This means that it also permits points to exist in the opposite region of the classification boundary. This is because there is a certain degree of variation in the fruit measurements sampled during the crop growth process. This variation may arise from measurement errors of the sensors in the actual measurements or from light scattering coming from different directions contaminating the sensor readings. Here, ξ_i represents the degree to which the training data x_i is allowed to protrude inside the margin, serving as a parameter when optimizing the SVM, represented by the following equation:

$$\min_{W,\xi} \left\{ \frac{1}{2} \|W\|^2 + C \sum_i^N \xi_i \right\}$$
$$t_i(W^T X_i + b) \geq 1 - \xi_i \xi_i = \max\{0, M - \frac{t_i(W^T X_i + b)}{\|W\|}\} \quad (6)$$

In this research, we take measurements of light spectra from tomatoes, normalize the data, and then plot these data points as clusters in a two-dimensional space. We then build a machine learning model to classify these points. The method we use is called Support Vector Machine (SVM). SVM finds a boundary that separates the different groups of points while keeping the

largest possible margin — the space between the boundary and the nearest points from each group.

The data includes a training set (used to teach the model) and a test set (used to check accuracy). The points are represented mathematically as vectors, and the boundary (or hyperplane) is described by a specific equation. Because plants grow under natural conditions, measurements can fluctuate slightly due to sensor errors or scattered light. So, the SVM allows some points to be on the wrong side of the boundary, within a margin of error called ξ. The model balances fitting the data well with allowing some errors to avoid overfitting — which is controlled by a parameter called C. A small C makes the model more flexible but less strict, while a large C makes it stricter and less tolerant of errors.

The second term of the equation includes the objective function. A smaller cost parameter C allows ξ_i to be larger, resulting in looser constraints. Conversely, a larger C restricts ξ_i from being large, preventing the training data from existing inside the margin or crossing the classification boundary into the opposite region. C becomes a hyperparameter that determines the performance of the SVM.

B. Adopt RBF-Kernel Method

In the previous example, it seems that classification can be achieved using a linear SVM, but better classification is expected if a curved boundary is used.

In the previous example, it appears that classification is possible using a linear SVM, but using a curved boundary can be expected to achieve better classification. Linear SVM seeks to find an optimal boundary for sample data in each of two regions to be classified by a straight line.

Looking at the sample data in each region, for example, samples located near one another tend to be similar. Specifically, if the reflectance at the same frequency in spectral measurements is similar, the characteristic values of crop growth obtained as a result of those measurements will also tend to be similar. If we think of this as similarity, we can expand on it as follows: When two samples with two-axis parameters are located nearby, the similarity of the samples will differ depending on whether the slopes of their indices, as a single index, tend to be the same or completely different. Therefore, we will reconsider this from the perspective of regression analysis. If the mean of each sample is set to 0, the covariance of the samples can be expressed as the correlation coefficient multiplied by the standard deviation. In this case, the larger the positive or negative the value, the higher the correlation; conversely, the closer the value is to 0, the lower the correlation.

This covariance can also be thought of as an inner product operation where each sample is multiplied together. If the similarity between samples is evaluated not only by the dot product between them but also by the distance between them, it becomes easier to understand if we change the idea of inverse proportion, where the smaller the sample distance, the higher the similarity. Therefore, by introducing a Gaussian kernel (RBF kernel), we can determine that the larger its value, the higher the similarity. This can be expressed mathematically as equation (9). This is exactly what is meant by finding a nonlinear classification boundary. Several methods have been proposed to find such a nonlinear classification boundary using a curve. In equation (2), we consider converting a two-dimensional vector into a higher-dimensional vector using a certain function φ.

$$\emptyset(\boldsymbol{x}_i) = (\varphi_1(\boldsymbol{x}_i), \varphi_2(\boldsymbol{x}_i), \cdots, \varphi_m(\boldsymbol{x}_i))^T$$
$$\emptyset(\boldsymbol{x}_j) = (\varphi_1(\boldsymbol{x}_j), \varphi_2(\boldsymbol{x}_j), \cdots, \varphi_m(\boldsymbol{x}_j))^T \quad (7)$$

$$x_i = (x_{i1}, x_{i2})^T$$
$$\emptyset(\boldsymbol{x}_i) = (x_{i1}, x_{i2}, x_{i1}^2)^T \quad (8)$$

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \quad (9)$$

The good thing about the above formula (8) is that it means that the formula is expressed as a three-dimensional linear function. Although there is no guarantee that this method can be easily generalized, classification using linear problems that did not work well in the original dimensions may be possible in higher dimensions. This method can be said to enable linear classification by converting the dimensions of the data to higher dimensions by selecting the function φ appropriately.

IV. EXPERIMENT

Using the UECS environment as an experimental field, we set up a tomato cultivation site. Every summer around the Obon holiday, we rebuild the beds to conduct new experiments with different crops. The beds are made using coconut fibers as a medium, which acts as a root-supporting solution that allows roots to absorb water and liquid fertilizer. Since pesticides are applied weekly, we used a prototype to measure the light spectra of fruits and leaves just before pesticide spraying.

In this experiment, tomato seedlings grew over a meter tall, and about five weeks after planting, fruits developed on the lower and middle parts of the plants. These fruits changed color from green (immature) to orange-yellow (almost ripe) and then to red. Based on their color, we measured their light spectra (see Figures 5). We collected about 70 samples of fruits for each color. Over time, the green fruits turned orange-yellow, and the orange-yellow ones turned red, indicating that the fruits continue to grow and ripen. The spectral data for green and red fruits showed that the normalized reference wavelengths ranged from a maximum of 445 nm to a minimum of 630 nm. The differences at other wavelengths reflected variations in the fruits' growth stages. Notably, during the transition from green to red, significant changes were observed at wavelengths 415 nm, 515 nm, 555 nm, and 680 nm. The 415 nm wavelength corresponds to the blue part of visible light, and its intensity increased as the fruits ripened.

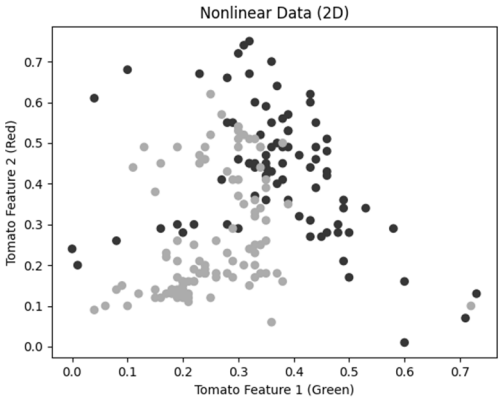

Fig. 4. Show 2D-Spectral Distribution of Selected between Light Wavelength (515-555 [nm]) in Visible light regin as scattered.

Note: The units for both the vertical and horizontal axes are normalized relative values.

On the other hand, wavelengths of 515 nm and 555 nm, corresponding to the yellow-green region, and 680 nm, in the red region, all showed decreasing light intensities. This suggests that measuring the multispectral light reflected by the crops could be linked to how farmers visually judge fruit ripeness. Traditionally, wavelengths in the near-infrared range (above 900 nm), along with 680 nm, have been used to calculate the Normalized Difference Vegetation Index (NDVI), which indicates plant vitality. Our discussion focuses on grouping visible light wavelengths and examining how the light distribution changes during plant growth. However, this is a complex topic because analyzing light wavelengths related to growth over time in a simple 2D or 3D space is difficult, especially considering the effects of time and growth stages.

In our experiment, we used tomato plants grown in a controlled environment. From the eight visible-spectrum wavelengths measured, we selected two wavelengths—515 nm and 555 nm—as features for classification (see Figure 6). We applied a Support Vector Machine (SVM) to classify the plant growth stages based on these features. The model was trained using specific parameters: a penalty parameter C set to 100 and a training set proportion T of 0.2, which controls the margin width.

Before training, we performed data preprocessing by removing outliers—data points outside the interquartile range (IQR) in a box plot—to reduce measurement variability. This preprocessing made it easier to find optimal classification boundaries compared to using raw data directly (see Figure 4). To evaluate the classification performance, we used a confusion matrix. Although some false positives and false

negatives remained, the accuracy stayed above 50%,

indicating decent classification ability (see Figure 6). The results also imply that outliers in the training data, if present in other datasets, could hinder classification results in a 2D space, pointing to limitations of this approach. Next, we explain our new separation method using dimensionality expansion. Linear SVMs sometimes fail to achieve linear separation using a

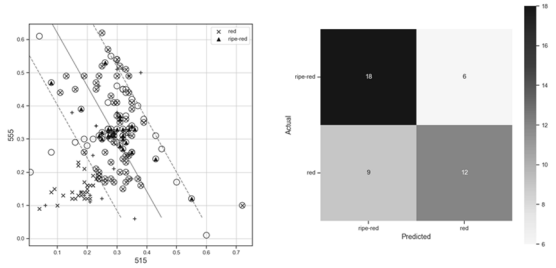

Fig. 5. Show Spectral Distribution by Linear-SVM (left) and Confusion Matrix between growth.

Note: The units for both the vertical and horizontal axes are [nm]. Each axis represents the frequency, and the numbers in the boxes represent the degree of fit. Between red and ripe-red of Tomato are compared.

straight line (or hyperplane). This occurs, for example, when data intermingle on a two-dimensional plane, like an interference zone on the negative side of the ideal margin, making it impossible to classify data in this zone. This is referred to as linear separation being impossible.

In contrast, nonlinear methods can be interpreted as mapping features to higher dimensions. This is equivalent to adding a height axis to the interference differences that exist within the plane. In other words, differences in height can be used to differentiate the data intermingled in the interference zone. It is easy to imagine that the hyperplane found in this way will be a plane that separates three dimensions, with a normal vector that is not parallel to either the plane axis or the height axis. Figure (5) shows a graph of the hyperplane found for the data plot shown in Figure (4).

When the hyperplane separated in this way is projected onto a two-dimensional plane, the boundary is curved, as shown in Figure (6). Although data plots projected onto a plane are

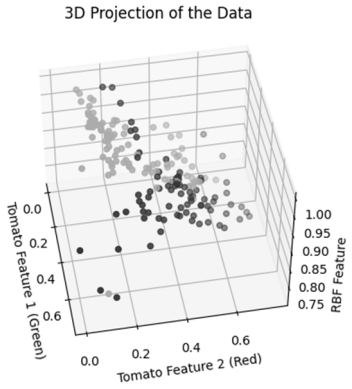

Fig. 6. Show 3D-Spectral Distribution of Selected between Light Wavelength (515-555 [nm]) in Visible light regin using RBF kernel.

Note: Each axis represents 2d- Light Wavelength and result of models using RBF-argolism (case parameter at C=1.0, γ=0.5).

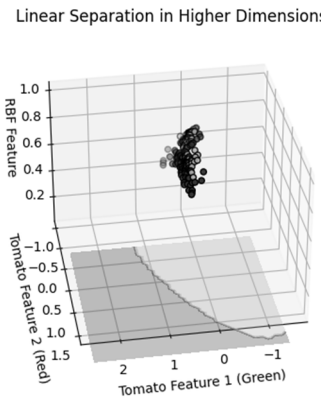

Fig. 7. Show Linear Separation in Higher Dimension between growth red Tomato and ripe-red Tomat with Spectral Distribution for model suitability after applying RBF kernel.

Note: Each axis represents the frequency, and the numbers in the boxes represent the degree of fit.

partially intersecting, they are linearly separable on a spatial hyperplane. In this method, a Gaussian function (Gaussian kernel) is used as the high-dimensional function, as shown in equation (9). The cost parameter C determines the degree of tolerance for misclassification. The smaller the kernel parameter γ, the simpler the decision boundary; the larger the value, the more complex the decision boundary.

In this study, we aimed to determine the growth level of vegetation. We demonstrated how to handle and analyze data sets containing outliers, which are issues that arise when classifying measured data. Furthermore, the SVM-based machine learning used in this study allowed us to tune parameters to the tolerance for outliers and the complexity of the boundary in classification. This demonstrated that even if linear separation is not possible, classification using a high-dimensional hyperplane is possible by expanding the dimensionality.

V. CONCLUSION

In this study, we used a prototype measuring device to conduct experiments to quantify the degree of growth of agricultural crops during the growth process through visible light multispectral (light wavelength) analysis. The growth and sweetness of agricultural crops are influenced not only by photosynthesis by sunlight, but also by complex conditions such as sunshine hours and irrigation timing. In the past, this work relied mainly on experience and subjectivity, but by quantitatively evaluating the distribution and changes of the visible light band, farmers can more accurately grasp the growth status of crops. When applying machine learning using conventional support vector machines (SVM) to classify growth levels and optimize classification, data preprocessing was effective for raw data that may contain outliers, but we attempted classification using a nonlinear separation method that allows

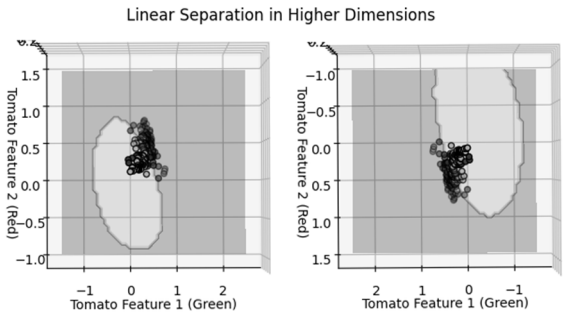

Fig. 8. Show 2D projection of Linear Separation in Higher Dimension between growth red Tomato and ripe-red Tomat with Spectral Distribution. (Left figure: RBF case RBF γ=1.5, Right igure: Hyper-parameter case C=10)

Note: Each axis represents the frequency, and the numbers in the boxes represent the degree of fit.

high-dimensional analysis and demonstrated its effectiveness visually. In the future, it will become increasingly important to evaluate classification characteristics from the differences in classification due to different learning algorithms and tuning of learning parameters. It is expected that quantified light spectrum distribution information will be used as additional information for automatic sorting machines and harvesting robots.

REFERENCES

[1] Murakami :"Morphological Control of Plant Growth in Protected Cultivation by Changing Optical Radiant Environment", Journal of the Illuminating Engineering Society, "Plant cultivation techniques using artificial light", Vol.79, No.4 p.149-154 (1994)

[2] Takashi Kameoka, Ito, Shin. Kameoka, Hashimoto : "Development of lettuce freshness measurement method using a multi-spectral sensing", The 31th Annual Conference of the Japanese Society for Artificial Intelligence, 2E3-OS-36a-3, (2007).

[3] Miura, Watanabe, Asai : "Sensing of laser Normalized Difference Vegetation Index(NDVI) for vegetation lidar", laser-sensing Society, (2019)

[4] Tao Li, Qichang Yang : Advantages of diffuse light for horticultural production and perspectives for further research, frontiers in Plant Science, (2015).

[5] "Detection of Ripe and Raw Tomatoes using Internet of Things", C.N. Vanitha, 7th International Conference on Computing Methodologies and Commnications (ICCMC-2023)

[6] "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] "Classification and Forecasting of Water Stress in Tomato Plants Using Bioristor Data", MANUELE BETTELLI, NICOLA COPPEDÈ , ANDREA ZAPPETTINI, FILIPPO VURRO1, RICCARDO PECORI, MICHELA JANNI,ANDDANIELE TESSERA, IEEE Access, pp.34795-34807, 2023.

[9] "Tomato Quality Classification Based on Transfer Learning Feature Extraction and Machine Learning Algorithm Classifiers", HASSAN SHABANI MPUTU, pp.8283-8295, IEEE Access, 2024.