Imperial College London

Department of Life Sciences

Broccoli Drought and Heat Complex Stress Detection and Shelf Life Prediction Based on Spectrometry and Machine Learning

AUTHOR:

XIAOSHENG LUO

CID:

01627437

SUPERVISOR:

Dr. OLIVER WINDRAM

August 27, 2019

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Research at Imperial College London

Formatted in the style of Methods in Ecology and Evolution

Submitted for the MRes in Computational Methods in Ecology and Evolution

Declaration

The images of the broccoli head on the conveyor belt used to construct and training the neural network was originally provided by Nathan E. Barlow (Imperial College London), and then the optimized image video was collected by the author and the supervisor, Dr. Oliver Windram (Imperial College London). Besides, Dr. Oliver Windram was mainly responsible for setting up the direction of this project. Acquisition of experimental data, data cleaning, data analysis, method modification, model training, parameter tuning and writing were exclusively performed by the author himself.

Abstract

As a non-destructive, and high-efficiency technology, spectroscopy has developed rapidly in crop management for precision agriculture. Many studies on spectroscopic detection of crop stress have made effective progress, including disease detection and water stress, etc. Various vegetation indices have also been established to indicate crops growth status. However, complex growing environment often cause crops to be affected by multiple stresses, which in turn affect crops yield and quality.Few studies have been conducted on this complex stress, especially the two physiological closely related abiotic stresses, heat and drought. Here we took broccoli as our research object, and collected the hyperspectral reflectance data of its leaves under heat and drought, as well as their combination by spectrophotometer. Then, we explored the spectral characteristics by various vegetation indices, and the results showed that a single vegetation index could hardly be significant in all treatments. Next, we train different machine learning models to predict these complex stresses, and the highest AUC can reach 0.9494 by logistic regression. Besides, we have developed a tool that can dynamically visualize the significant different wavelength between treatments. And additionally, we also try to combine spectroscopy and deep learning to build a system for predicting the shelf-life of broccoli heads, for now, it can track broccoli heads with ResNeXt to 97.2% accuracy and segment them with Unet to 99.2% accuracy, while the signal related to the shelf-life is still in progress.

Keywords: Broccoli; Machine Learning; Spectral feature; Heat stress; Drought stress

Word count: 2191.

Contents

1 INTRODUCTION

The environment in which crops growth and reproduction is highly dynamic, crops are constantly exposed to various stresses, including abiotic stimuli such as humidity, light intensity and temperature, and biotic stimuli, like pathogens. Correspondingly, to cope with these stresses, plants have evolved a highly dynamic response mechanism, from gene expression regulation to changes within various secondary metabolites, which in turn alter their physiological characteristics, such as enzyme activity, stomatal aperture, photosynthetic rate, transpiration rate and so on. Rapid detection of stresses through these physiological changes of crops is particularly important for obtaining high yield and high quality products. Traditional detection methods usually require skilled growers who are capable to notice the subtle color changes or a slight droop or curl of plants leaves, which indicate the stress. However, it's generally subjective and time-consuming. In contrast, spectral detection and spectral imaging, as a non-destructive, accurate and efficient method, has developed rapidly both in research and practical application.

The spectral reflectance characteristics and mechanism of leaves have been well summarized (Knipling, 1970, Gates et al., 1965). When the leaves receive solar radiation, only part of the incident energy is reflected and the rest is transmitted or absorbed for photosynthesis. The typical reflectance spectrum of a leaf is that in the visible region ($0.4 - 0.7\mu m$), the reflectance of leaves is generally very low, especially in red (around $0.63 - 0.70\mu m$) and blue (around $0.45 - 0.52\mu m$), these absorption is mainly caused by plant pigments. Specifically, the absorption of red primarily contributed from chlorophyll, and the absorption of blue is also involved in carotenes and xanthophylls (Gates et al., 1965, Rabideau et al., 1946). Furthermore, the high reflection in near-infrared ($0.7 - 1.3\mu m$) is caused by the internal cellular structure (Mestre, 1935, Willstätter and Mieg, 1907). Leaf cuticular wax is transparent and hardly reflects solar radiation. Radiation can be transmitted through the epidermis, then dispersed and multiple reflected and refracted in the mesophyll cells and air cavity, where different refracted index between air (1.0) and hydrated cell walls (1.4) account for these effect(Sinclair et al., 1968).

Previous studies on the application of spectral techniques in plant stress detection have made extensive progress, involving various crops and stress. In biotic stresses, plant disease detection is the most studied. Early in 1982, the diffuse reflectance spectra of potato tubers in the visible and near-infrared bands were measured and analyzed in an attempt to detect the presence of disease before its effects were visible(Muir et al., 1982). On top of that, spectral researches also progress in the detection of various plant diseases, such as panicle blast, brown planthopper, the bacterial leaf in rice (Kobayashi et al., 2001, Prasannakumar et al., 2013, Yang, 2010) and yellow rust, powdery mildew in wheat (Bravo et al., 2003, Cao et al., 2013). In abiotic stress, diverse researches focus on water stress. Among them, many studies are based on canopy temperature based Crop Water Stress Index (CWSI) measured from infrared thermometry (Alchanatis et al., 2010, Aladenola and Madramootoo, 2014, Bellvert et al., 2016). The principle of CWSI theory is that transpiration cools the surface of leaves. When soil moisture in the root zone decreases, stomatal conductance and transpiration are weakened, and then leaf temperature increases. Futher, what makes this theory popular is its linear relationship between canopy temperature and insufficient temperature and vapor pressure, as well as the development of empirical methods for quantifying crop water stress (Idso et al., 1981). However, though the canopy temperature is very useful for water stress detection, it still has some physiological concerns. In some plants, the diurnal fluctuation in stomatal conductance make the relationship unclear between canopy temperature and stress levels (Zarco-Tejada et al., 2012). Moreover, leaf temperature does not directly explain other physiological changes, such as photosynthesis pigments or non-stomatal reduction of photosynthesis under water stress (Zarco-Tejada et al., 2013). Therefore, various alternative vegetation indices (VIs) based on the visible and red edge spectral region are developed to capture water stress related signals (Berni et al., 2009, Zarco-Tejada et al., 2013, Wang et al., 2015, Rossini et al., 2013, Panigada et al., 2014, Dangwal et al., 2016).

Although spectroscopic studies of biotic and abiotic stresses can achieve significant detection under different models, the stress they detect is often single, while in practical production, crops tend to suffer from multiple complex stresses during their growth, such as heat and drought and their combinations, especially in the context of global warming. More interestingly, the molecular and physiological mechanisms by which plants respond to heat and drought stress have been extensively studied and they show lots of connection. oth of them can differentially affect the RNA stability, alter the enzyme activity and disrupt the steady-state of metabolic flux, which in most of cases can cause a common response, oxidative damage (Kollist et al., 2018, Suzuki et al., 2012, Mittler et al., 2012, McClung and Davis, 2010). Besides, photosynthesis is an important physiological phenomenon affected by drought and heat stress. Drought can lead to stomatal closure and reduces CO_2 uptake which makes plants more susceptible to photodamage. And also it can induce negative changes in photosynthetic pigments, either increase or decrease chlorophyll content(Lawlor and Cornic, 2002, Anjum et al., 2011, Din et al., 2011). Similarly, exposure to high temperature can also result in a reduction in chlorophyll biosynthesis, thereby disturbing the photosynthetic pigment components (Camejo et al., 2006). Despite many physiological links between plant heat stress and water stress, and it is of great meaningful to detect them in the pratical production, little research have been conducted on the relationship between these two stress spectral characteristics and their detection. So it's a great interest and also useful to see how well we can detect the heat stress and drought and their combination from the crops by data mining their spectral characteristics changes.

Alongside this, methodologically, many previous studies on crops stresses detection used statistical discriminant model, specifically, variety of VIs, as a simple and effective algorithm for quantitative and qualitative evaluation of vegetation cover, growth dynamics, and stress levels. But due to different spectral combinations, instrumentation, platforms, and resolutions used, it's hard to have a unified mathematical expression that defines all VIs, customized algorithms needed to developed and tested against specific application requirements(Xue and Su, 2017). What'more, the prediction is often unsatisfactory and the generalization ability somewhat insufficient. Compared with traditional statistical discriminant model, machine learning methods, which developed rapidly in recent year, can generally improve the speed and accuracy of prediction. The main difference between machine learning and statistics lies in their purpose. Statistical models are more designed to infer the relationship between variables, while machine learning models are intended to make the most accurate prediction possible.

Overall, here we explore the ability to use hyperspectral techniques to detect and differentiate more complex stresses of crops, that is, the combinations of drought and heat, which are highly correlated with each other. Firstly, we calculated some widely known VIs for statistical testing in anticipation of obtaining simple and effective stress-related signals. Then, in order to approach the upper limit of accuracy for detecting different stresses, we apply the machine learning strategy, using linear classifiers, such as Logistic regression (LG), linear support vector machine (SVM) and tree models, like random forest (RF) and XGBoost for training a robust classifier. On top of this, in order to increase the interpretability of the model, great efforts had been made in feature engineering, including dimension reduction, statistical filter method and model-based embedded method, the comparison and selection of features are carried out for model training, and finally achieved considerable prediction accuracy. In particular, during these process, in order to better visualize the statistical differences of hyperspectral features under different stress comparisons dynamically, meanwhile be able to adjust the wavelength width to reduce the redundancy of hyperspectral information and provide a reference for specific band selection, a simple visualization search tool has been developed.

Additionally, in order to optimize the customization of supermarket broccoli products' shelf-life, a computer vision strategy based on deep learning and spectral detection technology is being developed. Now it can track and segment broccoli heads on conveyor belt through convolution neural network, but futher spectral signals realted to shelf-life still need to be excavated.

2 MATERIALS AND METHODS

2.1 Data Collection

Broccolis were grown in Control Environment room and greenhouse. The normal growth temperature is controlled at 23 °C and the humidity is controlled at 60%, with long daylight (16 hours illumination, 8 hours darkness) treatment, and water is poured every 3 days from the tray to the soil. The whole growth cycle of broccoli takes about three months, during which it needs to be transferred to a suitable pot according to the size of broccoli.

Table 1: Apparatus used in the experiment.

Item	Description
Camera	Ximea 1.3 MP NIR Enhanced Camera MQ013RG-ON
Machine vision lens	MVL12M23-12 mm EFL, f/1.4, for 2/3" " C-Mount Format Cameras, with Lock.
Band pass filter	FEL0800: Ø25.0 mm Premium Longpass Filter, Cut-On Wavelength: 800 nm. FBH520: Ø25.0 mm Hard-Coated Bandpass Filters, Blocking Regions (OD >5): 200 - 485 nm, 556 - 1200 nm. FBH650: Ø25.0 mm Hard-Coated Bandpass Filters, Blocking Regions (OD >5), 200 - 611 nm, 690 - 1200 nm. FBH850: Ø25.0 mm Hard-Coated Bandpass Filters, Blocking Regions (OD >5), 200 - 805 nm, 896 - 1200 nm.
LED controller	Intelligent LED Solutions 12-Channel Light Controller
LED	12 Die LED array Full Spectrum 360-955nm

During stress treatment, broccolis were randomly divided into four groups with 8 individuals in each group. They were treated under control, heat stress (27°C), drought stress (without watering) and the combination of heat and drought stress. Leaf reflectance spectroscopy data was collected by the Ocean Optics FLAME-S-XR1 spectrophotometer in a completely dark room, while the spectral image were collected by Ximea cameras with a specific bandpass filter (FEL0800, FBH650-40) (Table 1) under the illumination of the corresponding wavelength of the LED lamp, four days of data were collected until the leaves show a distinct dehydration drooping phenotype. And in the open-air greenhouse, spectral images are collected in a grow tent.

Broccoli heads for shelf-life prediction come from POLLYBELL FARMS LTD. They are divided into two groups, 18 in each, one of which is stored in cold storage for a some time, and the other is harvested freshly. They were placed naturally at room temperature and spectral image data were collected every day through the cameras with bandpass filter (FBH520-40, FBH650-40,FBH850-40) until they decay significantly.

2.2 Vegetation Indices Calculations

The names, abbreviations, calculation formulas and citations of the various vegetation indices used in this study are as follows, mainly includes commonly used remote sensing indices, chlorophyll-related indices, and indices related to water stress indications.

Table 2: Various vegetation indices

Name	Abbrev.	Equation	References
Ratio vegetation index	RVI	R_n / R_r	(Jordan, 1969) (Pearson and Miller, 1972)
Normalized difference vegetation index	NDVI	$(R_{800} - R_{680}) / (R_{800} + R_{680})$	(Rouse Jr et al., 1974) (Tucker, 1979)
Enhanced vegetation index	EVI	$2.5\,(R_n - R_r)\,(R_n + 6 \cdot R_r - 7.5 \cdot R_b + 1)$	(Huete et al., 2002)
Chlorophyll vegetation index	CVI	$R_n \cdot R_r / R_g^2$	(Vincini et al., 2008)
Chlorophyll index - green	CI-G	$R_n / R_g - 1$	(Gitelson et al., 2003)
Chlorophyll index - red edge	CI-RE	$R_n / R_{re} - 1$	(Gitelson et al., 2003)
Photochemical reflectance index	PRI	$(R_{531} - R_{570}) / (R_{531} + R_{570})$	(?)
Water index	WI	R_{900} / R_{970}	(?)
Structure independent pigment index	SIPI	$(R_{800} - R_{445}) / (R_{800} + R_{680})$	(?)

R_λ is the reflectance at wavelength λ; n, re, b, g and r represent NIR (760 – 900 nm), RE (700 – 730 nm), blue (450 – 520 nm), green (520 – 600 nm) and red (630 – 690 nm) respectively.

2.3 Machine Learning

Machine learning generally includes several steps in practical operation, such as data collection and preprocessing, model selection, training, evaluation and repeatedly fine-tuning until a good prediction effect is achieved (Figure 2). In the data preprocessing stage, Z-score standardization is applied to simplify the calculation and the categorical data is one-hot encoded. In the strategy of training algorithms, firstly, logistic regression (LG), support vector machine (SVM), random forest (RF) and XGBoost algorithm were trained to fit the raw data and obtained the baseline score, and then the performance of the models were optimized by feature engineering and parameters adjustment. Most of the code used in this process is based on the API provided by sklearn(Pedregosa et al., 2011).

2.3.1 Model Fitting

Logistic regression: the binomial logistic regression model is a classification model, which is represented by the conditional probability distribution $P(Y|X)$, in the form of parameterized logistic distribution. Here, the value of X is a real number, and the random variable Y takes a value of 0 or 1, then we estimate the model parameters by supervised learning. The binomial logistic regression model is the conditional probability distribution as follows:

$$P(Y=1|x) = \frac{\exp(w \cdot x)}{1+\exp(w \cdot x)}$$

$$P(Y=0|x) = \frac{1}{1+\exp(w \cdot x)}$$

Here, x is the input vector, w is the weight vector and Y is the output vector, $Y \in \{0,1\}$, $x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)}, 1)^T$, $w = (w^{(1)}, w^{(2)}, \ldots, w^{(n)}, b)^T$. By comparing the probability of $P(Y=1|x)$ and $P(Y=0|x)$ can finally determine the category. The cost function of logistic regression can be derived by the method of maximum likelihood estimation, which is known as the average of cross-entropy loss. Meanwhile, in order to prevent over-fitting, the L1 or L2 regularization terms was added during the optimization process.

SVM: the main idea of SVM is to find the decision boundary with the largest classification interval between two different categories, which means that a hyperplane separating classes in the feature space is defined by the principle of maximum margin between the closest different data points, also known as support vectors. For simple linear separability problems, it can be described as an optimization problem by mathematical formulas as follows:

$$\max_{w,b}\left[\min_{x_i}\frac{y_i\,(w\cdot x_i+b)}{\|w\|}\right]$$

The minimized item represents the distance from the support vectors to the decision boundary with sign, known as geometry margin. By scaling w and b so that (x_j, y_j) as the point to get the minimum value, $y_j\left(\omega^T x_j + b\right) = 1$, so the other sample points are naturally greater than or equal to 1. Derived all the way and we can finally got a methematical optimization problem:

$$\begin{aligned}&\min_{w,b}\tfrac{1}{2}\|w\|^2\\&\text{s.t. } y_i\left(w^{\mathrm{T}}x_i+b\right)\geqslant 1,\quad i=1,2,\ldots,m\end{aligned}$$

Further, in order to allow the SVM to ignore some noise, a slack variable ($\xi_i \geq 0$) is introduced to allow some wrong classification, that is, allow some data points' functional margin less than 1, correspondingly, a penalty term is needed to add to the objective function to limit the slack variable, and here is the basic linear separable SVM:

$$\begin{gathered}\min\tfrac{1}{2}||w||^2+C\textstyle\sum_{i=1}^{m}\xi_i\\\text{s.t. } y_i\left(w^T x_i+b\right)\geq 1-\xi_i(i=1,2,\ldots m)\\\xi_i\geq 0\quad(i=1,2,\ldots m)\end{gathered}$$

Finally, the problem can be resolved by Lagrange Duality and SMO algorithm (Platt, 1998). In addition, kernel mapping has also been tried to verify the model's performance under nonlinearly separable conditions.

Ensemble methods: Random Forests and XGBoost(Chen and Guestrin, 2016), both are based on decision tree model, they use the methods of bagging and boosting respectively, which can help to prevent high variance and high bias. Random forest mainly consists of two stochastic processes, random sampling of samples and features to construct many decision trees that are independent of each other. The final prediction results are summarized by voting strategy. As for XGBoost, it's an algorithm developed from gradient boosted decision trees and designed for speed and performance. It's widely used in many competitions and achieved good grades.

In the process of training machine learning algorithms, when multi-classification is performed on logistic regression and support vector machines, "one to the rest" strategy is applied. All the models are validated by 6-fold cross-validation, and ROC_AUC is used as the metric for model evaluation.

2.3.2 Feature Engineering

Generally, data and features determine the upper bound of machine learning, whereas models and algorithms only approximate this upper bound. The purpose of feature engineering is to extract effective features and remove redundant featuresfrom the original data. It basically includes feature extraction, feature construction and feature selection 2. Separately, feature extraction mainly uses dimension reduction methods such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Feature construction is to construct new features based on previous expertise. Feature selection methods can be roughly divided into three types:

- Filter: scoring each feature according to divergence, correlation, etc., and then set a threshold for selection feature.
- Embedded: use some machine learning algorithms and models to train and get the coefficients of each feature, select features according to the coefficient, kind of similar to the filter method, but models are trained to determine the pros and cons of features. Specifically, multi-method ensemble selection (Feilhauer et al., 2015)was modified from the regression problem and later adapted to the classification problem.
- Wrapper: recursive elimination feature method, due to the high computational complexity and the long execution time of the algorithm, it is not adopted here.

In addition, in order to find a suitable bandpass filter for the camera, a search box with specific bandwidth was used to repeatedly and randomly select features, and the importance of features is sorted by simple ANOVA and TukeyHSD significance test. Finally, the graph is plotted by accumulating significant ($p < 0.05$) bandwidth features.

2.4 Computer Vision

The project involves computer vision tasks such as image classification, segmentation, and image alignment. Specifically, image alignment was performed by classic Scale-invariant feature transform (SIFT)(Lowe et al., 1999), Image classification and segmentation are mainly accomplished by the transfer learning of convolution neural network (Figure 1). More specifically, Image classification

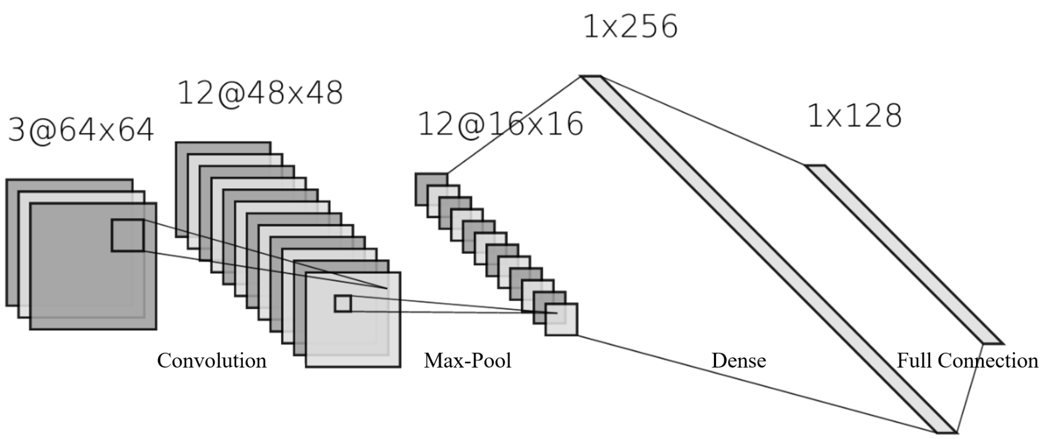

Figure 1: Structure of a simple convolution neural network
An image was taken as input (for example, a RGB image, normally three channels), then through the calculation with the multiple kernels' parameters and activation functions (usually ReLu) in convolution layer, and the downsampling process in pooling layer, can achieve the purpose of weight sharing and parameter reduction. Finally,the results are expanded and classifiedby the fully connected layer and the softmax function.In the figure, the number in front of @ is the number of channels, and the back is the height and width of pixels.

was implemented by ResNeXt (Xie et al., 2016), developed by UC San Diego and Facebook AI Research, while Image segmentation was implemented by Unet (Ronneberger et al., 2015).The training was conducted at P1000 on the HPC of Imperial College London, the optimizer for the neural network is Adam, the cost function is cross-entropy, and cyclical learning rates (Smith, 2017) was used, Most of the code is based on the API provided by Pytorch, fastai library (Howard et al., 2018), and opencv.

2.5 Workflow

The project mainly includes two parts (Figure 2). The laboratory part is to grow broccolis under control conditions and then perform individual and combined stress treatments, collect spectral images and leaf reflectance spectrum data to explore the signals that can effectively distinguish among them and construct a robust machine learning classifier. The application part is to construct a detection system which can predict the shelf life of broccoli on the conveyor belt through computer vision methods and spectral images under specific bandwidth, which is selected based on the results obtained in the laboratory.

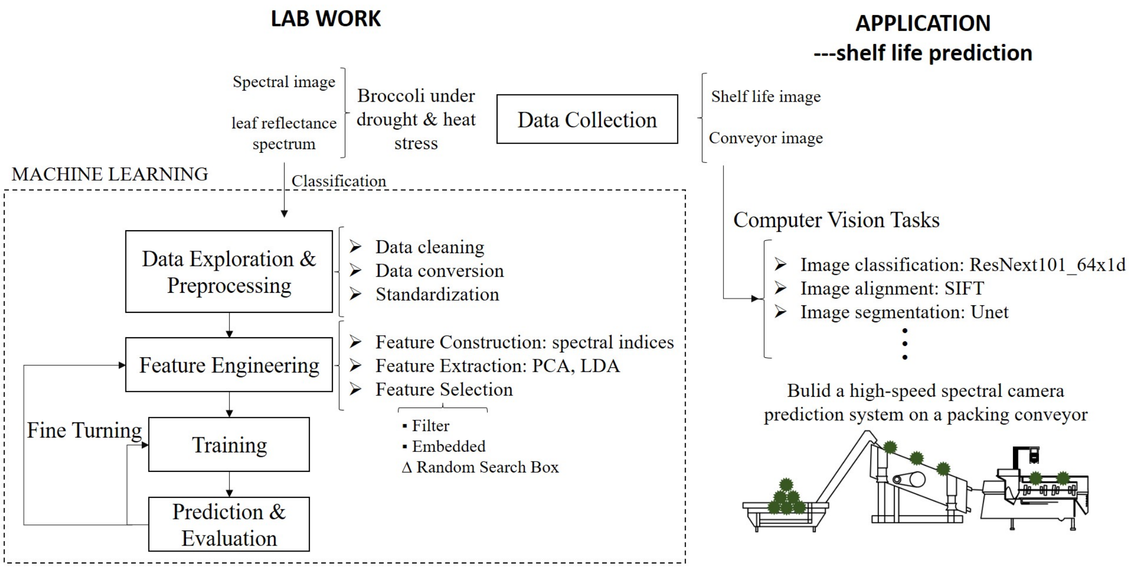

Figure 2: Project workflow

3 RESULTS

3.1 Data

We select the data of the day when the broccoli just appeared phenotype under the stress (Figure 3) to ensure the treatment effect. Under the control and heat conditions, the broccolis have no obvious phenotype, while under drought and combined stress treatment, the broccoli leaves are a little drooping due to dehydration, and the combined stress is slightly more obvious than the drought.

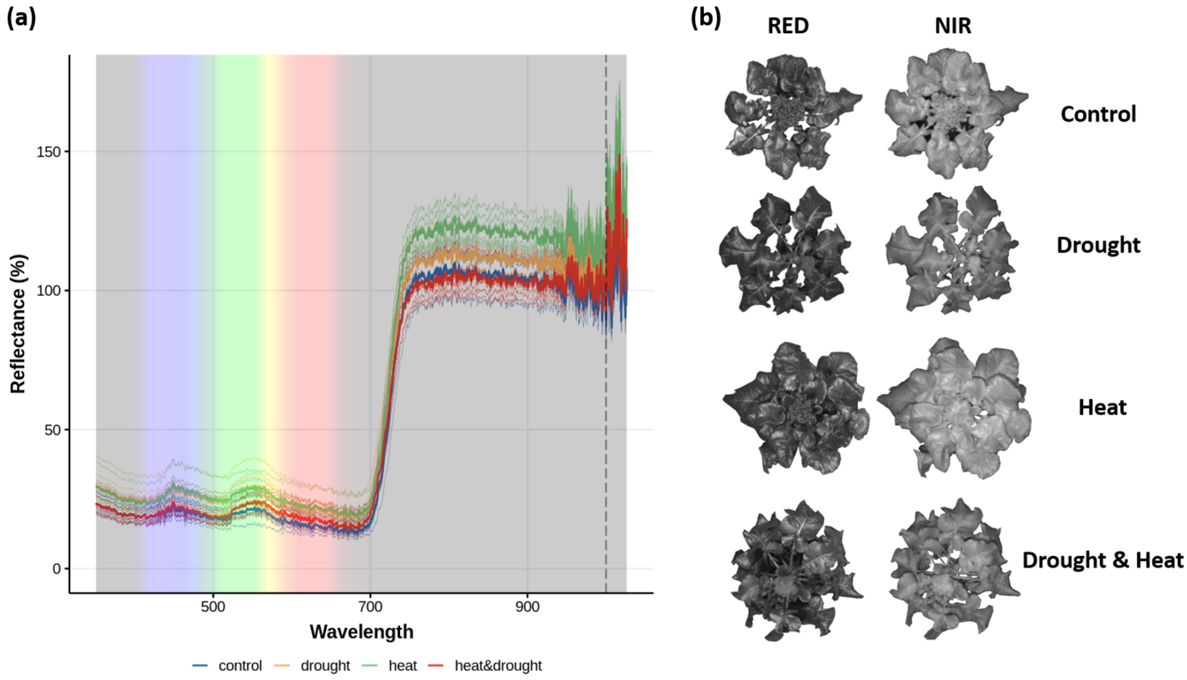

Figure 3: Spectral data of broccoli under heat and drought stress
(a) is the hyperspectral data of broccoli leaves detected by spectrophotometer in dark environment, the vertical axis represent their relative reflectivity.The thin line is averaged by the hyperspectral scan of all the leaves of each sample, and the thick line is the average of all samples in different treatments. The right side of the dashed line was discarded is subsequent processing due to abnormal signal fluctuation. (b) is the spectral images taken by the camera with red bandpass filter (CWL = 650 nm, FWHM = 40 nm) and near infrared bandpass filter (> 800 nm), under the illumination of the corresponding band of LED.

On the other hand, in the hyperspectral data of the leaves (Figure 3 (a)), it is difficult to get a distinct discriminant pattern from the perspective of data distribution, because the samples of different treatments are cross-covered. However, the overall trend of the broccoli leaves reflectance spectrum can still be clearly seen. There are small peaks at the blue and green-yellow junctions in the visible region, and it is well known that a small valley in the red band and strong reflection rate shifting in the near-infrared band. In particular, the average reflectance of the heat treatment appears to be significantly higher than other treatments.

3.2 Vegetation Indices

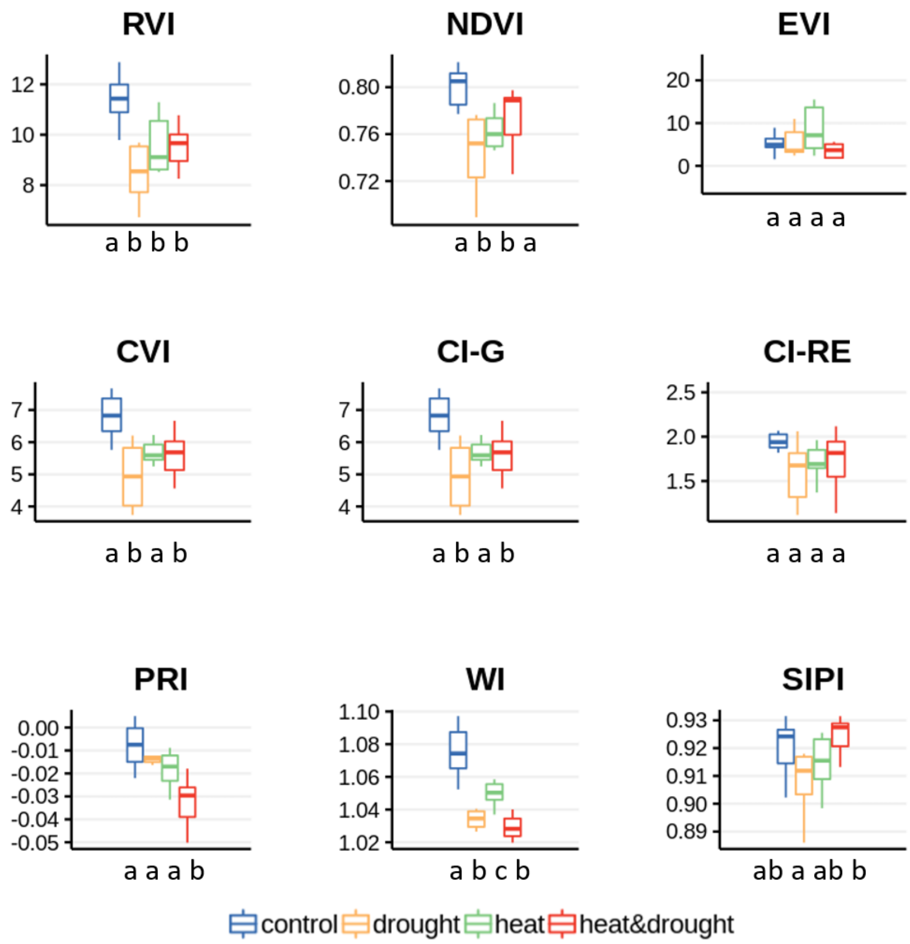

Figure 4: Various vegetation indices under different stresses
There are seven samples per treatment, the equation for calculating vegetation index refers to Tabel **??**.Normality test was implemented by shapiro test, and homogeneity of variance test was implemented by Barlett test before using ANOVA and Tukey's HSD for post-hoc analysis. Different letters under the x-axis represent sigificant differences ($p < 0.05$).

The spectral indices related to vegetation cover and chlorophyll content as well as water stress were calculated under different stress (Figure 4). The results show that significant differences between every two treatments can't be simply obtained from a single index. The results with significant differences also show different forms of band feature calculation methods, which is difficult to establish a unified pattern, suggesting that more complex models are needed.

3.3 Dimensionality Reduction

There may be multi-collinearity between hyperspectral features, that is variables may be correlated. Meanwhile, too many variables may hinder the pattern for model fitting, and it may also involve a lot of redundant information. Therefore, dimensionality reduction was used to reduce variable, speed up computation and extract effective information hidden in the data.

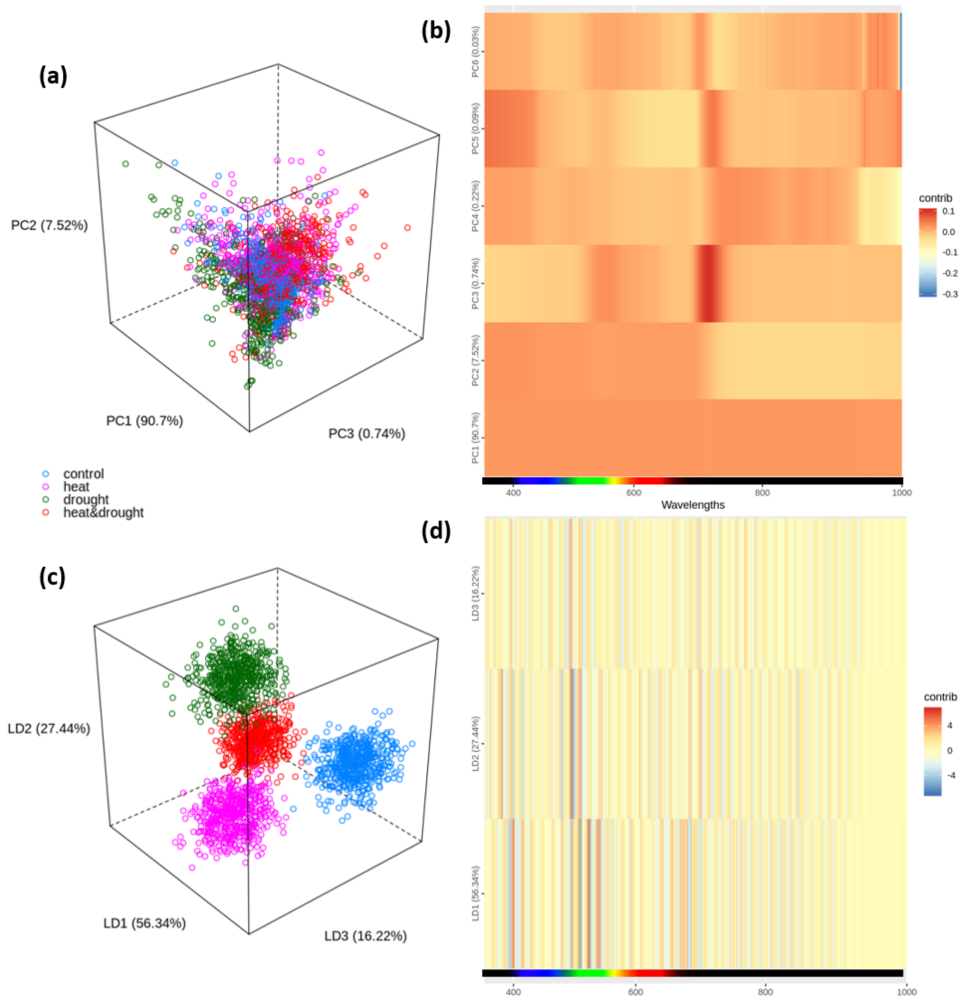

Figure 5: Feature dimensionality reduction by PCA and LDA
(a) and (b) are the distribution of the samples in the first three principal component dimensions, in which the values in the coordinate axis are the explanatory rates of the overall differences; Heat maps (c) and (d) indicate the correlation between each wavelength and each component.

The results of unsupervised dimensionality reduction PCA show that, when the variables are mapped to the linear-independent direction of maximum variance, clustering between different treatment is not effective. Surprisingly, PC1 explains 90.7% of the variance, and the contribution of each wavelength seems to contribute equally to it, possibly due to the systematic errors. In PC2, the visible light region contributes a larger variance, and there are two specific narrow bands in PC3 that are positively correlated with it. On the other hand, LDA, the supervised dimension reduction method, can completely separate the different treatments. In these components, the green (around $520nm$) and red edge (around $680nm$) wavelength may be important to separate each stress treatment in the principal components.

3.4 Feature Selcetion

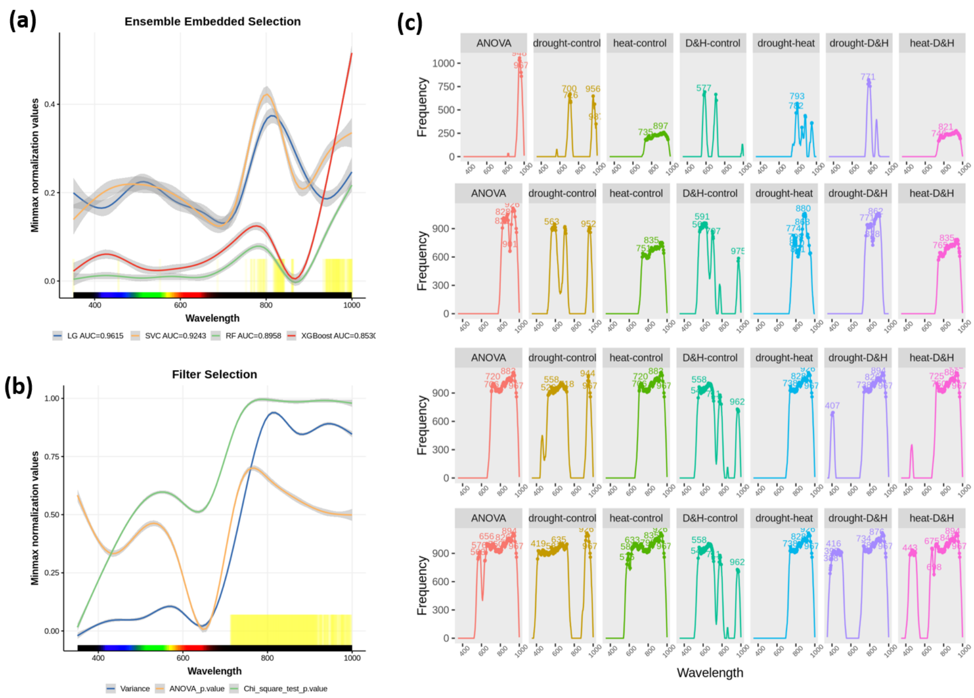

Figure 6: Feature Selection
In (a) and (b), the gray shadows show the confidence intervals, and the yellow vertical lines indicate the important features upon the setting threshold. (c) shows the significant difference of wavelengths between different treatments dynamically, the earlier the peak appears, the more significant it is, the number on the graph represents where the peak is.

Through further data mining, we try to use different methods to select the wavelengths that can effectively distinguish different treatments. Filter and Embedded method are commonly used feature selection methods in machine learning. The results of the Filter show that the variance, chi-square test and the ANOVA of each wavelength are coincident in the NIR region. And through the modified ensemble method, the correlation coefficient or feature importance obtained by fitting the LG, SVM, RF and XGBoost are used to establishing the threshold according to their respective goodness of fitting, results show that two NIR fragments and other fragments in some bands may be important.

However, although the results obtained by the above methods are in good agreement with our existing knowledge, the detail of features importance in the pairwise relationship of different treatment remain unclear, also for embedded method, the generalization ability could be constrainted by the model. Moreover, in practical application, bandwidth is usually limited to a certain value, continuous band selection is more useful. So here, we create a random search box sorting method. Through simple statistical analysis, it can dynamically display the feature importance under specific bandwith, and detail the feature importance between different treatments which can provide a feature engineering basis for more sophisticated and complex models.

3.5 Models

Four kinds of machine learning model were applied to fit the different treatments' hyperspectral data, each treatment has 7 samples, 80 scans per sample and each scan is the average of 10 scans. A total of $4 \times 7 \times 80$ data was used for model fitting, with 6-fold cross-validation and ROC-AUC as an evaluation metric.

Results auc shows that in the fitting of the original data, the linear classifier LG and linear SVM fit well, while tree-based model performance is general, mainly because the number of features is too large. which makes the tree model easier to overfit. After removing the noise from PCA, the performance of the tree model is improved obviously, but the effect of LDA is not ideal. Furthermore, through the result of feature selection in the previous step, the models were trained at $409 \sim 429nm$, $558 \sim 577nm$ and $700 \sim 896nm$ bands, and the performance of the tree model is further improved.

Table 3: ROC_AUC of 4 machine learning models under different data

	Logit	SVM	RF	XGBoost
raw	0.9494	0.9122	0.7794	0.8235
LDA	0.6826	0.6817	0.6710	0.6655
PCA	0.8509	0.8507	0.8285	0.9005
Selection	0.9108	0.8758	0.7438	0.8026
Selection+PCA	0.9155	0.8754	0.8767	0.9096

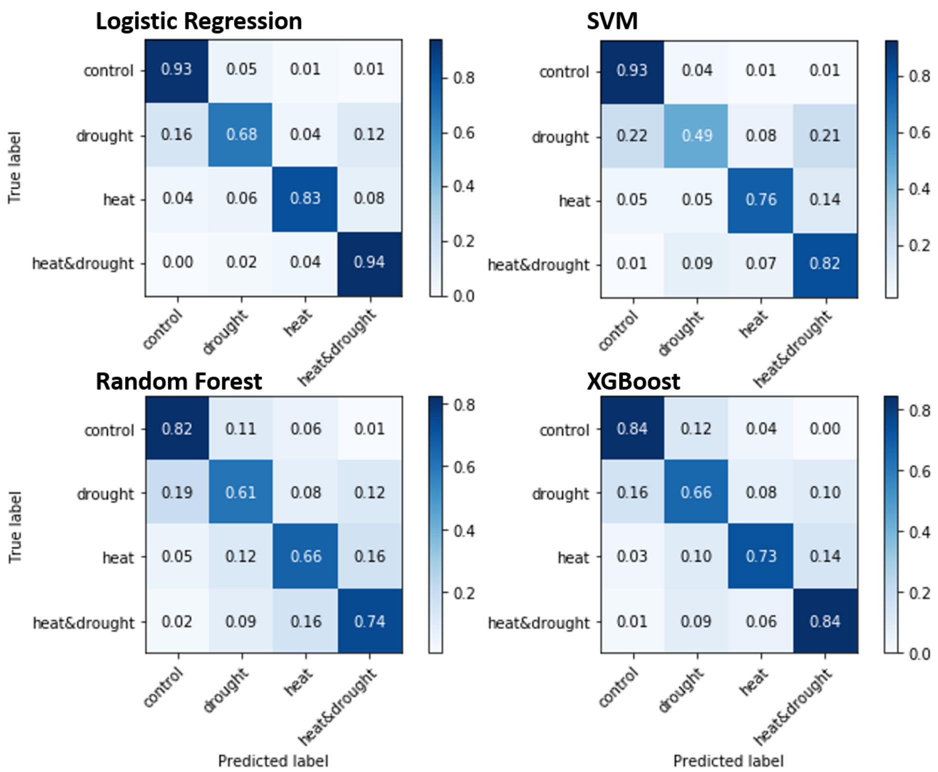

Figure 7: The confusion matrix

Next, we observe their confusion matrix when they have their best AUC. The prediction performance of the four models is similar for different treatments. Among them, the control and combined stress treatments have the best predictive effect, followed by the heat stress, while the drought stress is the worst, and its mispredictions tend to appear more in control and combined stress.

3.6 Shelf Life Prediction

To predict the shelf life on the broccoli packaging conveyor, we collected thousands of broccoli spectral images (Figure 8,(a)) with high-speed band-specific filter cameras and specific band LED lights, and then we marked 5708 broccoli images to tell whether the broccoli head was in the middle of the lens. And through data augmentation, in the case of limited computing resource, a simple classifier can be constructed by transfer learning of ResNext101_64 convolution neural network, which can easily achieve an accuracy of 97.2%. By setting a lower threshold of predicting probability, then we can select the highest probability of broccoli head images between the two valleys to track broccolis. After that, for the sake of removing the influence of the background, we labeled 492 masks for broccoli images segmentation, the Unet was trained to reach 99.2% accuracy. Next, through the SIFT operator, the images of different channels can be aligned for subsequent analysis.

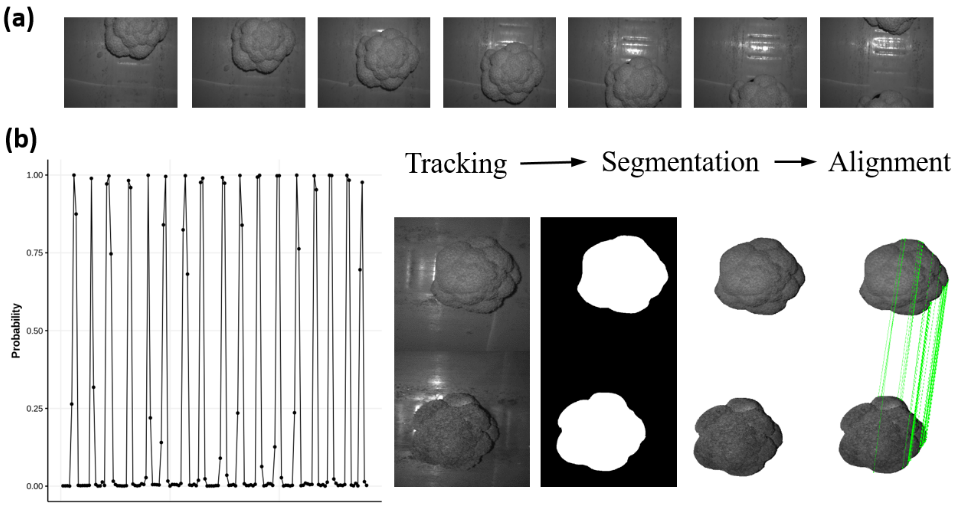

Figure 8: Processing of broccoli head images on conveyor belt
(a) from left to right is a broccoli time series images on a conveyor belt in the near-infrared channel; (b) is the ResNeXt prediction to track whether broccoli is in the middle of the lens. If it did, then segment by Unet and matched by SIFT.

Next, we placed the new and stored broccoli heads at room temperature, let them decay naturally to capture the shelf-life related signals. With time elapsing, the reflectivity of the three selected channels can somewhat reflect the rotting changes. However, our focus is more on the first two days before the broccoli heads showed obvious decay phenotype. Among the three selected channels, the newly harvested broccoli heads could not be effectively distinguished from the stored one. Some vegetation indices were also tried, but more research is needed.

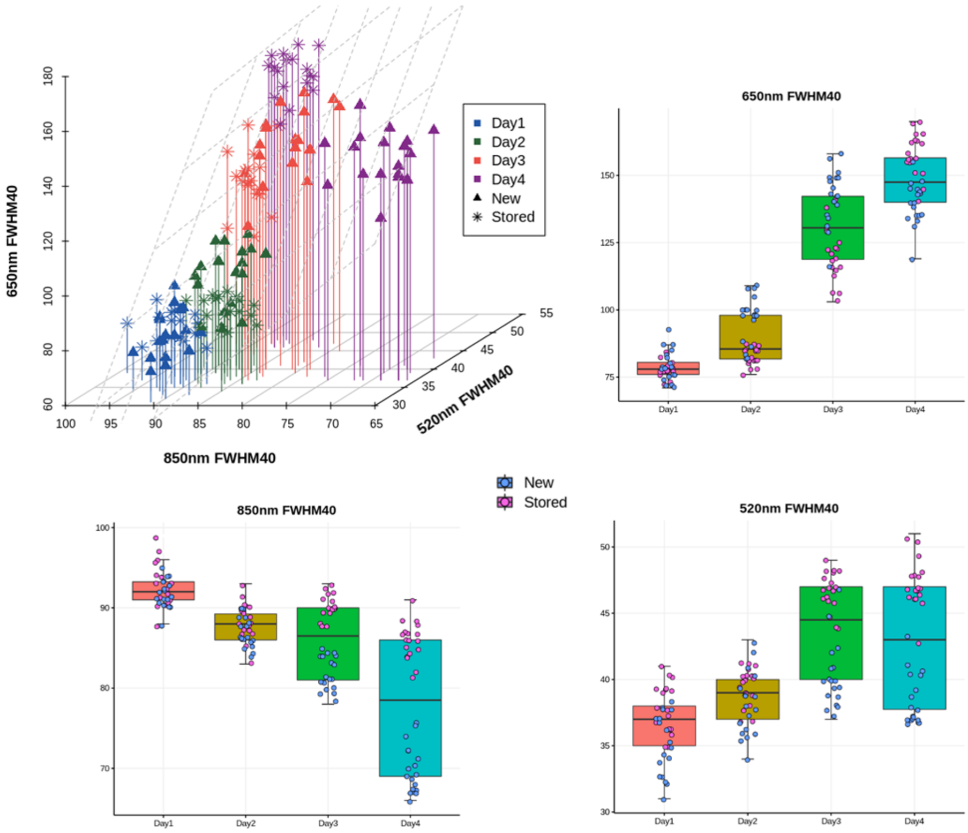

Figure 9: Spectral reflection signal of three channels when the broccoli heads naturally rot

The spectral reflectance of the broccoli head changes significantly with its natural decay. Among the early spectral signals which we are more concerned, 520 nm and 850 nm bands seem to be more effective to distinguish the new and stored heads. However, in general,their non-linear variations in different signals are elusive, and further modeling is needed.

4 DISCUSSION

From the results of various VIs calculated in different treatments, it seems not surprising that not getting a unified index that can effectively distinguish these four treatments, because of the complex relationship among them. However, detailed analysis can still give us some inspiration. Firstly, indices based on red and near-infrared such as RVI, NDVI and EVI, can somewhat reflect the physiological characteristics of plants, however, they are more used in the field of remote sensing for various kinds of local, regional, and global scale models, including general circulation and biogeochemical models(Peterson et al., 1988, Huete et al., 2002). While in CVI, CI-G, CI-RE these indices of leaf chlorophyll content(Gitelson et al., 2003, Vincini et al., 2008), they seem to have a consistent result, except CI-RE, which doesn't consider the green channel. Results show that in the water-deficient environment (drought, heatdrought), the chlorophyll content of broccoli leaves was significantly affected. This water stress phenomenon has been extensively supported, and it's also well understood that lack of water hinders nutrient transport in plants and thus affects photosynthetic pigments synthesis. As for the water index (WI), it reflects water absorption in the mesophyll andd had been shown to have a good indication of water content in many crops(Wang et al., 2015, **?**, **?**). The result here is also very satisfactory, it can be seen that there were significant differences between the two treatments except drought and combined stress. This indicates that in heat stress, the water content of the plant leaves is also affected, even they are well watered, and this effect is not sufficient to superimpose the significance in the combined stress relative to the drought stress. The functional basis of the PRI is related on its sensitivity to rapid changes in carotenoids through the de-epoxidation of the xanthophyll pigments(**?**). It can serve as an indirect means for water stress detection due to the effects of water stress on the efficiency of photosynthesis. Researchers have demonstrated the sensitivity of PRI to short-term crop water stress detection(**??**Zarco-Tejada et al., 2013), and to the long-term change of carotenoid/chlorophyll ratio(**?**). Here we found that PRI is sensitive to the combination of drought and heat stress in broccoli, which may imply the cumulative effect of stress on photosynthetic efficiency and pigments. And Of course, more studies is needed to support these inference.

In the process of machine learning model training for hyperspectral data, the results of feature engineering show that the important features are mainly concentrated in the green and near infrared. In detail, because the hyperspectral bandwidth is small,there will inevitably be a lot of redundant information. Data dimensionality reduction can effectively extract important information, shorten model training time and reduce over-fitting. Here, dimensionality reduction by PCA can effectively de-correlate and remove the linear relationship between dimensions, but it does not consider the classification information. Therefore, after dimensionality reduction, the loss of information will be minimized, but classification may become more difficult. The data points in the graph are not scattered, and from the contribution of each component to the principal component, it is true. It's hard to get useful information. Here, dimensionality reduction by PCA can effectively de-correlate and remove the linear relationship between dimensions, but it does not consider the classification information. Therefore, after dimensionality reduction, the loss of information can reduce to the lowest level, but it may not be helpful in classification **??**. The sample points in the graph are not centralized, and from the contribution of each wavelength to the principal components, it is really difficult to obtain useful information. Another commonly used dimension reduction method is LDA, which seeks to distinguish data points as easily as possible after dimension reduction. After dimensionality reduction, the sample data has the largest inter-class distance and the smallest intra-class variance in the new dimension space, and the data has the best separability in the low dimension space **??**. It can almost reach 100% classification, so that the contribution of each wavelength may better explain the important features for classification, here are the green and red edges.

Then, as for feature selection, several methods of experimentation have a good consistency, that is, infrared band information might berelatively important for classification. It may be suggested that changes in mesophyll cell structure, such as membrane structure, are more likely to affect spectrum reflection in leaves under heat and drought stress, while changes in pigments are less important to distinguish between them. In particular, through the dynamic visualization of the random feature search box, we can more clearly see the importance of features to the relationship between them.The water stress and control group showed the most significant difference around 700 ~ 900nm, which was basically consistent with the WI. Wavelength around 700 ~ 800nm may be important for distinguishing between heat stress and control. As for combined stress, it seems similar to the water stress. And between the combined stress and the individual stress, there is a difference around 420nm.

Finally, the results of the four machine learning models show that linear classifier performs well when the data dimension is relatively large,while the tree model does not. This is understandable because regularization is used in training linear classifiers, which can effectively deal with multiple collinearity problems and reduce the weight of redundant information. The tree model can also be improved after dimensionality reduction.According to the statistical analysis results, we empirically select 409-429, 557-558, 700-896 nm band information for training, so we can see that the performance of the XGBoost model has been improved effectively, its AUC can reach 0.9096. By showing the confusion matrix, it is not surprising that the control group and the combined stress group can achieve the best distinction. But surprisingly, the heat stress group can be more effectively distinguished from other stresses, as opposed to the drought group which may have more phenotype. The erroneous distinction of drought group mostly appears in the difference to control group, which may be explained to some extent that some of the leaves do not reach the threshold at which the drought can be detected.

As for the prediction of broccoli shelf life, due to the increasingly mature computer vision technology based on deep learning, and relatively stable environment and large data generated in production. It is easy to obtain high accuracy through a large number of data labeling and transfer learning. What is important is that for the capture of broccoli shelf life related signals, the more challenging is the signal difference in the early fresh period. Although we can get a signal that changes significantly with the decay of broccoli,how to predict its shelf life in the early stage remains to be further studied.

References

Aladenola, O. and Madramootoo, C. (2014), 'Response of greenhouse-grown bell pepper (capsicum annuum l.) to variable irrigation', *Canadian journal of plant science* **94**(2), 303–310.

Alchanatis, V., Cohen, Y., Cohen, S., Moller, M., Sprinstin, M., Meron, M., Tsipris, J., Saranga, Y. and Sela, E. (2010), 'Evaluation of different approaches for estimating and mapping crop water status in cotton with thermal imaging', *Precision Agriculture* **11**(1), 27–41.

Anjum, S., Wang, L., Farooq, M., Hussain, M., Xue, L. and Zou, C. (2011), 'Brassinolide application improves the drought tolerance in maize through modulation of enzymatic antioxidants and leaf gas exchange', *Journal of Agronomy and Crop Science* **197**(3), 177–185.

Bellvert, J., Marsal, J., Girona, J., Gonzalez-Dugo, V., Fereres, E., Ustin, S. and Zarco-Tejada, P. (2016), 'Airborne thermal imagery to detect the seasonal evolution of crop water status in peach, nectarine and saturn peach orchards', *Remote Sensing* **8**(1), 39.

Berni, J. A., Zarco-Tejada, P. J., Suárez, L. and Fereres, E. (2009), 'Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle', *IEEE Transactions on geoscience and Remote Sensing* **47**(3), 722–738.

Bravo, C., Moshou, D., West, J., McCartney, A. and Ramon, H. (2003), 'Early disease detection in wheat fields using spectral reflectance', *Biosystems Engineering* **84**(2), 137–145.

Camejo, D., Jiménez, A., Alarcón, J. J., Torres, W., Gómez, J. M. and Sevilla, F. (2006), 'Changes in photosynthetic parameters and antioxidant activities following heat-shock treatment in tomato plants', *Functional Plant Biology* **33**(2), 177–187.

Cao, X., Luo, Y., Zhou, Y., Duan, X. and Cheng, D. (2013), 'Detection of powdery mildew in two winter wheat cultivars using canopy hyperspectral reflectance', *Crop Protection* **45**, 124–131.

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', ACM, pp. 785–794.

Dangwal, N., Patel, N., Kumari, M. and Saha, S. (2016), 'Monitoring of water stress in wheat using multispectral indices derived from landsat-tm', *Geocarto International* **31**(6), 682–693.

Din, J., Khan, S., Ali, I., Gurmani, A. et al. (2011), 'Physiological and agronomic response of canola varieties to drought stress', *J Anim Plant Sci* **21**(1), 78–82.

Feilhauer, H., Asner, G. P. and Martin, R. E. (2015), 'Multi-method ensemble selection of spectral bands related to leaf biochemistry', *Remote Sensing of Environment* **164**, 57–65.

Gates, D. M., Keegan, H. J., Schleter, J. C. and Weidner, V. R. (1965), 'Spectral properties of plants', *Applied optics* **4**(1), 11–20.

Gitelson, A. A., Gritz, Y. and Merzlyak, M. N. (2003), 'Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves', *Journal of plant physiology* **160**(3), 271–282.

Howard, J. et al. (2018), 'fastai', `https://github.com/fastai/fastai`.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X. and Ferreira, L. G. (2002), 'Overview of the radiometric and biophysical performance of the modis vegetation indices', *Remote sensing of environment* **83**(1-2), 195–213.

Idso, S., Jackson, R., Pinter Jr, P., Reginato, R. and Hatfield, J. (1981), 'Normalizing the stress-degree-day parameter for environmental variability', *Agricultural meteorology* **24**, 45–55.

Jordan, C. F. (1969), 'Derivation of leaf-area index from quality of light on the forest floor', *Ecology* **50**(4), 663–666.

Knipling, E. B. (1970), 'Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation', *Remote sensing of environment* **1**(3), 155–159.

Kobayashi, T., Kanda, E., Kitada, K., Ishiguro, K. and Torigoe, Y. (2001), 'Detection of rice panicle blast with multispectral radiometer and the potential of using airborne multispectral scanners', *Phytopathology* **91**(3), 316–323.

Kollist, H., Zandalinas, S. I., Sengupta, S., Nuhkat, M., Kangasjärvi, J. and Mittler, R. (2018), 'Rapid responses to abiotic stress: priming the landscape for the signal transduction network', *Trends in plant science* .

Lawlor, D. W. and Cornic, G. (2002), 'Photosynthetic carbon assimilation and associated metabolism in relation to water deficits in higher plants', *Plant, cell & environment* **25**(2), 275–294.

Lowe, D. G. et al. (1999), Object recognition from local scale-invariant features., *in* 'iccv', Vol. 99, pp. 1150–1157.

McClung, C. R. and Davis, S. J. (2010), 'Ambient thermometers in plants: from physiological outputs towards mechanisms of thermal sensing', *Current Biology* **20**(24), R1086–R1092.

Mestre, H. (1935), 'The absorption of radiation by leaves and algae', *Cold Spring Harbor Symposia on Quantitative Biology* pp. 191–209.

Mittler, R., Finka, A. and Goloubinoff, P. (2012), 'How do plants feel the heat?', *Trends in biochemical sciences* **37**(3), 118–125.

Muir, A., Porteous, R. and Wastie, R. (1982), 'Experiments in the detection of incipient diseases in potato tubers by optical methods', *Journal of Agricultural Engineering Research* **27**(2), 131–138.

Panigada, C., Rossini, M., Meroni, M., Cilia, C., Busetto, L., Amaducci, S., Boschetti, M., Cogliati, S., Picchi, V., Pinto, F. et al. (2014), 'Fluorescence, pri and canopy temperature for water stress detection in cereal crops', *International Journal of Applied Earth Observation and Geoinformation* **30**, 167–178.

Pearson, R. L. and Miller, L. D. (1972), Remote mapping of standing crop biomass for estimation of the productivity of the shortgrass prairie, *in* 'Remote sensing of environment, VIII', p. 1355.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Peterson, D. L., Aber, J. D., Matson, P. A., Card, D. H., Swanberg, N., Wessman, C. and Spanner, M. (1988), 'Remote sensing of forest canopy and leaf biochemical contents', *Remote Sensing of Environment* **24**(1), 85–108.

Platt, J. C. (1998), Sequential minimal optimization: A fast algorithm for training support vector machines, Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING.

Prasannakumar, N., Chander, S., Sahoo, R. and Gupta, V. (2013), 'Assessment of brown planthopper,(nilaparvata lugens)[stål], damage in rice using hyperspectral remote sensing', *International journal of pest management* **59**(3), 180–188.

Rabideau, G. S., French, C. S. and Holt, A. S. (1946), The absorption and reflection spectra of leaves, chloroplast suspensions, and chloroplast fragments as measured in an ulbricht sphere.

Ronneberger, O., Fischer, P. and Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.

Rossini, M., Fava, F., Cogliati, S., Meroni, M., Marchesi, A., Panigada, C., Giardino, C., Busetto, L., Migliavacca, M., Amaducci, S. et al. (2013), 'Assessing canopy pri from airborne imagery to map water stress in maize', *ISPRS Journal of Photogrammetry and Remote Sensing* **86**, 168–177.

Rouse Jr, J., Haas, R., Schell, J. and Deering, D. (1974), 'Monitoring vegetation systems in the great plains with erts'.

Sinclair, T. R. et al. (1968), 'Pathway of solar radiation through leaves'.

Smith, L. N. (2017), Cyclical learning rates for training neural networks, *in* '2017 IEEE Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 464–472.

Suzuki, N., Koussevitzky, S., Mittler, R. and Miller, G. (2012), 'Ros and redox signalling in the response of plants to abiotic stress', *Plant, Cell & Environment* **35**(2), 259–270.

Tucker, C. J. (1979), 'Red and photographic infrared linear combinations for monitoring vegetation', *Remote sensing of Environment* **8**(2), 127–150.

Vincini, M., Frazzi, E. and D'Alessio, P. (2008), 'A broad-band leaf chlorophyll vegetation index at the canopy scale', *Precision Agriculture* **9**(5), 303–319.

Wang, X., Zhao, C., Guo, N., Li, Y., Jian, S. and Yu, K. (2015), 'Determining the canopy water stress for spring wheat using canopy hyperspectral reflectance data in loess plateau semiarid regions', *Spectroscopy Letters* **48**(7), 492–498.

Willstätter, R. and Mieg, W. (1907), 'Untersuchungen über chlorophyll; iv. ueber die gelben begleiter des chlorophylls', *Justus Liebigs Annalen der Chemie* **355**(1), 1–28.

Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K. (2016), 'Aggregated residual transformations for deep neural networks', *arXiv preprint arXiv:1611.05431* .

Xue, J. and Su, B. (2017), 'Significant remote sensing vegetation indices: A review of developments and applications', *Journal of Sensors* **2017**.

Yang, C.-M. (2010), 'Assessment of the severity of bacterial leaf blight in rice using canopy hyperspectral reflectance', *Precision Agriculture* **11**(1), 61–81.

Zarco-Tejada, P. J., González-Dugo, V. and Berni, J. A. (2012), 'Fluorescence, temperature and narrow-band indices acquired from a uav platform for water stress detection using a micro-hyperspectral imager and a thermal camera', *Remote sensing of environment* **117**, 322–337.

Zarco-Tejada, P. J., González-Dugo, V., Williams, L., Suárez, L., Berni, J. A., Goldhamer, D. and Fereres, E. (2013), 'A pri-based water stress index combining structural and chlorophyll effects: Assessment using diurnal narrow-band airborne imagery and the cwsi thermal index', *Remote sensing of environment* **138**, 38–50.

5 Supplementary Information

5.1 Experiment Apparatus