Multimodal neural network with clustering-based drop for estimating plant water stress

Computers and Electronics in Agriculture 168 (2020) 105118

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Multimodal neural network with clustering-based drop for estimating plant water stress

Kazumasa Wakamori[a], Ryosuke Mizuno[a], Gota Nakanishi[a], Hiroshi Mineno[a,b,c,*]

[a] Graduate School of Integrated Science and Technology, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan
[b] Research Institute of Green Science and Technology, Shizuoka University, Japan
[c] JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

ARTICLE INFO

Keywords:
Plant water stress
Multimodal deep learning
Time-series modeling
Image processing

ABSTRACT

Decision-making with low-cost data is an attractive approach in the field of agriculture as it aids to solve the difficulty of inheriting advanced cultivation technologies. To provide expertise in the decision-making process for stress cultivation, precision irrigation based on plant water stress is required to steadily produce high-quality fruits. Single low-cost data namely, single-modal data, is used in the traditional approach. However, for advanced cultivation, multimodal data such as physiological and meteorological data is required. In this study, we propose a multimodal neural network with clustering-based drop (C-Drop) for accurate estimation of plant water stress, as it is an index for irrigation decision-making, using plant image and environmental data. Our proposed method extracts temporal multimodal features from leaf wilting features (physiological data) using environmental features (meteorological data) as an attention mechanism of a multimodal neural network that includes long short-term memory layers. Moreover, the proposed neural network with C-drop realizes a novel end-to-end deep learning architecture in consideration of environmental conditions. On evaluating this method against the existing methods, the proposed method was found to improve the accuracy of plant water stress estimation by 21% for mean absolute error and root-mean-squared error, thereby indicating that this method is precise and stable for the plant water stress estimation. The performance of our proposed method to support precision irrigation will allow new-age farmers to produce high-quality fruits steadily.

1. Introduction

Advances in technology enable explicit knowledge modeling of decision-making by the expert. In the field of agriculture, expert farmers produce high-quality crops based on their knowledge and decision-making skills. However, the knowledge of decision-making seems to have been lost owing to the reduced population of farmers and the difficulties involved in technological inheritance. Recently, several studies have reported that determining the factors behind making a decision using information technologies is highly valuable to prevent the loss of sophisticated expert knowledge (Singh et al., 2018). These studies apply data mining, image processing, and machine learning technologies to the images of plants and environmental data to extract variables that are considered as the decision-making factors. Defining decision-making factors of stress cultivation is strongly needed because such cultivation approach can produce high-quality fruits such as high-sugar content tomato.

Stress cultivation requires decision-making for precision irrigation based on the plant water stress. Therefore, a water stress index should be defined for irrigation scheduling as a decision-making factor. The sugar content of fruits increases when the amount of water being provided by restricted irrigation is decreased during the cultivation. Therefore, the total yield decreases but each fruit is much better in quality (Patanè and Cosentino, 2010). However, when irrigation is extremely restricted, and the plants are exposed to high water stress, they will die, and recovery will not be possible. To provide stress cultivation to new farmers who have no expert knowledge, a practical and accurate measurement method of plant water stress is required to support the irrigation scheduling.

Previous studies have proposed several methods to measure plant water stress. In general, water stress changes according to the water potential, which includes the water content of the leaves and stems. Thus, plant water stress is measured accurately by the direct evaluation of the water potential (Boyer, 1967). However, this measurement method cannot be applied to real-time irrigation scheduling because it requires destructive measurement including leaf excision and long

* Corresponding author at: 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan.
E-mail address: mineno@inf.shizuoka.ac.jp (H. Mineno).

https://doi.org/10.1016/j.compag.2019.105118
Received 7 August 2019; Received in revised form 18 November 2019; Accepted 22 November 2019
Available online 03 December 2019

processing time for quantitation. Several studies have proposed non-destructive estimation methods using plant stress responses including leaf vibrations (Sano et al., 2015) and stem diameter variations (Meng et al., 2017; Wang et al., 2017). These methods can estimate plant water stress in real-time and can be applied to real-time irrigation control based on water stress. Water stress diminishes the water content in plant cells and decreases the turgor pressure in leaves and stems. As a response to stress, the leaf tension decreases and the stem diameter shrinks. Thus, the two methods estimate water stress by measuring the leaf turgor pressure using an ultrasonic speaker and microphone or measuring the stem diameter variations using a laser displacement sensor. However, the two methods are inadequate to provide the water stress cultivation method to new farmers because these measuring sensors are expensive and require expert knowledge to install and measure the water stress.

Plant images and environmental data are also used for estimating plant water stress (Sánchez-Molina et al., 2015; Takayama and Nishina, 2007; Guo et al., 2017; Kaneda et al., 2017). These methods cost low and provide ease of measurement. The measurement devices (RGB cameras) and environmental sensors have become inexpensive owing to the popularity of IoT devices that can measure plant data without any contact. These methods can substitute specialty sensors including laser displacement sensors that require expert knowledge and are expensive with more powerful sensors such as cameras and environmental sensors. Therefore, image or environmental-data-based methods can be applied to deliver water stress cultivation to new farmers who cannot use specialty sensors. Traditional methods used single-modal data (plant image or environmental data). In contrast, a water stress estimation method (Kaneda et al., 2017) uses both plant images and environmental data as the multimodal data. This method estimates water stress using multimodal neural network, including CNN, and sliding window-based support vector regression (SW-SVR) (Kaneda and Mineno, 2016). In the experimental results, this method has proven to be more accurate in estimating the plant water stress.

Meanwhile, there are two issues in the existing method. First, the temporal information is not considered, despite which the input features and dependent variables have a temporal dependence relationship. Recently, recurrent neural network (RNN) has shown state-of-the-art results in certain time-series applications by extracting temporal features optimized for the main problem (Benabderrahmane et al., 2018; Veličković et al., 2017). However, the existing water stress estimation method concluded that RNN should be applied in the future work because the combination of CNN and RNN requires large amounts of network parameters. The trade-off relationship between temporal feature extraction by RNN and the increase in the number of network parameters should be resolved. The second problem is insufficient consideration of environmental characteristics. The existing method assumed that the plant images and environmental data have equal importance in water stress estimation. However, the leaf wilting in plant images and environmental data have a different role on plant physiology water stress estimation (Wakamori and Mineno, 2019). The environmental data can define the importance of leaf wilting for the estimation. However, the existing method processed the two data equally in the neural network and SW-SVR. The dynamic attention mechanism based on environmental consideration optimizes the features for estimating water stress.

This article proposes a water stress estimation method called multimodal neural network with clustering-based drop (C-Drop). As compared to the existing method (Kaneda et al., 2017), our method calculates the variations in stem diameter in the same way as the water stress index but with improved multimodal neural network and input data. Our multimodal neural network includes long short-term memory (LSTM) layers, which is an RNN. The proposed method adopts optical-flow-based feature extraction instead of CNN-based feature extraction that was used in the existing methods. Considering that leaf wilting is defined by the movement of the leaf, optical flow, which can quantify the movement of objects in an image, extracts the wilting features, thereby contributing to the estimation of water stress index without using CNN. Moreover, the leaf wilting and environmental data have a temporal dependence relationship with the stem diameter variations, which should be resolved. Using LSTM layers that can solve the temporal dependence, the neural network is expected to build a highly accurate water stress estimation model. The multimodal neural network appropriately trains the water stress estimation model with the proposed C-Drop as a dynamic attention mechanism on environment data.

The remaining of this article is as follows: Section 2 presents the details of the plant water stress and related work. The detailed proposal is described in Section 3. In Section 4, the proposal is evaluated using actual cultivation data. Finally, the conclusions are presented in Section 5.

2. Preliminaries

2.1. Plant water stress

Plant water potential and stress are generally controlled by irrigation scheduling and physical and chemical natures of the substrate or soil conditions. In this study, we intend to distinguish dynamic regulation factors and static regulation factors for water stress. The dynamic factors include climatic environment and irrigation scheduling that dynamically changes the water stress due to diurnal or seasonal variations. The static factors are the physical and chemical nature of the substrate or soil. In conventional farming, the static factors are defined when cultivation is initiated and not changed during cultivation. In this study, we focus on the dynamic regulation factors for plant water stress caused as leaves' transpiration speed exceeds the water absorption root speed.

In a greenhouse (a general cultivation environment), environmental factors such as temperature, relative humidity, and brightness continuously change due to the strength of solar radiation. Thus, the cultivation systems in a greenhouse cannot accurately control the environmental factors that depend on the weather and the seasons without using large capital investments such as a fully controlled plant factory. A simple environmental control that includes shade and ventilation is used in greenhouse, which is not highly accurate and cannot regulate the transpiration speed of plants. However, any cultivation environment that utilizes a greenhouse can control the irrigation timing. Therefore, several plant water stress estimation methods have been proposed to control irrigation and regulate plant water status (Sano et al., 2015; Meng et al., 2017; Wang et al., 2017; Sánchez-Molina et al., 2015; Takayama and Nishina, 2007; Guo et al., 2017). These methods are categorized into two types: plant response utilization method and environmental data utilization method.

The plant response utilization method observes the changes in plants due to water stress such as leaf vibrations (Sano et al., 2015) and stem diameter variations (Meng et al., 2017; Wang et al., 2017). In addition, image-based method observes leaf wilting extracted from plant images (Takayama and Nishina, 2007; Guo et al., 2017). These all methods use plant response. Therefore, they can directly express water stress. Moreover, the plant image-based method has two advantages: ease of measurement and low-cost as compared to the leaf vibration and stem-diameter-based methods. The measuring device, an RGB camera, can measure without establishing contact. It is less expensive than other sensors. Using camera devices based on mature and widespread IoT technologies can reduce the hardware cost in practical application. These methods enable new farmers to inherit water stress cultivation without expert guidance and are low-cost. The water stress index is calculated by an estimation method to express water stress regardless of camera angle or location. Conversely, different commercial cultivation environmental conditions, such as the type of farming facility or camera location, are a cause for variation in index of previous method. In this sense, these "plant image-based" methods can be further improved to

quantify water stress and leaf wilting.

Methods that use environmental data apply the knowledge from climate data related to control transpiration speed to indirectly express water stress (Sánchez-Molina et al., 2015). Environmental data-based methods have the same two advantages of image-based methods: the ease of measurements and its low-cost. In addition, environmental data can express the water stress levels because it indirectly relates to transpiration speed, a water stress index. The use of environmental data to estimate water stress could be further improved by combining it with plant response data such as leaf wilting, information which is not available in environmental data. Furthermore, environmental data has the potential to support leaf wilting interpretation extracted from plant images, transforming it into an absolute index of water stress by the relationship modeling of the optical-flow-based analysis between leaf wilting and stem diameter variations (Wakamori and Mineno, 2019).

To investigate this hypothesis, a multimodal method (Kaneda et al., 2017) has been proposed. The method estimates the variations in stem diameter as an absolute index of water stress from plant images and environmental data. The estimation is performed by a regression model that calculates the stem-diameter-based absolute water stress index from plant images and environmental data with a neural network. To measure the stem diameter, it replaces expensive sensors for cheap ones including an RGB camera and an environmental sensor. Details of this method are described in Section 2.3. Nevertheless, the relationships between the three collected data (leaf wilting extracted from plant images, environmental data, and stem diameter variations) are mentioned below. Leaf wilting and environmental data perform distinct roles in the estimation of stem diameter variations to calculate the water stress index in plants. Leaf wilting data directly expresses stem diameter variations because both are related to water transport caused by transpiration in day-time (Fricke, 2017). Most of the water is transported from roots to leaves according to the plant water potential gradient. When the leaf osmotic pressure, which defines the water potential in leaves decreases due to transpiration, water is transported from roots to leaves through the stems. In this sense, leaf wilting directly correlates with stem diameter variations with a water transport temporal dependence. The transpiration speed is defined by environmental conditions. Light is a cause for stomatal opening. Plant transpiration is based on the difference in vapor pressure of the air outside the stomata. Therefore, the variations in the stem diameter as an absolute index of water stress can be appropriately estimated by plant images and environment data. The importance of leaf wilting increases in specific environmental conditions of high transpiration speed.

2.2. Multimodal neural network

Multimodal neural network continues to become more sophisticated (Veličković et al., 2017; Duan et al., 2018). These neural networks learn how to fuse multimodal input features with their deep architecture (Ngiam et al., 2011) and can roughly resolve the relationship between multimodal features and the dependent variables. Recent studies have proposed network architectures suitable for multimodal data. As a typical architecture, multi-stream architecture has been proposed (Duan et al., 2018) that has independent streams, which extract independent modal features from each modality of data. This is followed by combining the independent features with the neural network and extracting multimodal features that optimize the main problem. In addition, cross-modal LSTM (X-LSTM) has been proposed as state-of-the-art architecture for time-series multimodal modeling (Veličković et al., 2017). X-LSTM extracts independent features from their respective streams and cross-modal features using cross-connection. The cross-connection, inspired by biological cross-modal systems, allows the information flow between multimodalities. It is implemented by branching each stream and connecting them to other modal streams. By using the cross-connect, the X-LSTM demonstrated higher accuracy than the existing methods in time-series modeling using time-series healthcare data. Our proposal includes input feature design and C-Drop. The architecture of the proposed neural network is not limited to a specific architecture. In this study, we evaluate the proposed concept using two architectures: multi-stream and X-LSTM, as typical architecture and state-of-the-art architecture, respectively.

2.3. Multimodal sliding window-based support vector regression

Multimodal SW-SVR (Kaneda et al., 2017) has been proposed for predicting water stress and it demonstrated the precision of multimodal data for water stress modeling. This method predicts future stem diameter variations as a plant water stress index from the plant images and environmental data using a multimodal neural network and SW-SVR. The neural network extracts multimodal features from the input data, then the SW-SVR predicts plant water stress using the extracted features. This method adopts a two-stream architecture as the multimodal neural network. One stream extracts the wilting features from the plant images by CNN, and the other stream extracts environmental features using fully connected layers. For input images, a preprocessing method has been proposed known as remarkable moving objects detected by adjacent optical flow (ROAF). In ROAF image, the non-wilting area is masked based on optical flow and the wilting area is emphasized. By masking the unnecessary area, the CNN can extract leaf wilting features efficiently from the image. These streams are then fused and the network extracts multimodal features that are given as input to the SW-SVR. The SW-SVR is an ensemble learning algorithm based on feature clustering. Its basic theory is to build weak learners based on feature clustering, and dynamic weighting for the inference values of weak learners. First, the SW-SVR builds weak learners for each feature condition such as different seasons and weather. The specific feature condition is defined by the cluster center calculated by k-means. Each weak learner is allocated to each cluster center. The weak learners are then trained using neighborhood data from the cluster center. To predict future values, weak learners collect training data by dynamic-short distance data collection (D-SDC) that selects effective data for the specific condition by considering movement using ‘k’ neighborhood data for their cluster center on Euclidean distance, which is the feature variation to predict horizons. The SW-SVR proposes the future predictions. It is also used for estimating the current by training weak learners using current dependent variable. Finally, the inference values of SW-SVR take the changing characteristics of testing data into account by dynamically prioritizing them. The SW-SVR dynamically determines the weights of weak learners that are based on the similarities between the input feature and each corresponding weak learner. Here, when the first clustering result defined the environmental conditions, the water stress was theorized with high accuracy by weak learners trained in similar environmental conditions.

Meanwhile, considering the relationship between leaf wilting, environmental data and stem diameter described in Section 2.1, two issues were found in the existing method. First, the existing method does not consider temporal information in modeling. However, leaf wilting and stem diameter variation, as water stress index, have temporal dependency. To solve this problem, research on the estimation of water stress using temporal information was conducted (Brillante et al., 2016). However, in (Brillante et al., 2016), the temporal information had a fixed length and a dynamic time feature corresponding to the changes in a plant could not be considered. Thus, an estimation method that considers dynamic temporal information is required. Recently, RNN has shown state-of-the-art results in several time-series modeling because it can extract temporal features optimized for the main problem but the existing method (Benabderrahmane et al., 2018; Veličković et al., 2017) does not include RNN. RNN uses a large amount of network parameters for the recurrent connection. Therefore, the combination of CNN and RNN has larger parameters to solve spatio-temporal problems than just CNN. The requirements of the training data and computational resources increase according to the increase in

the number of network parameters. For this reason, the existing study has focused on only using CNN and mentioned that RNN will be applied in future work. When the pre-extraction of image features without CNN explaining the plant water stress, there is room for applying RNN-based multimodal neural network. By extracting multimodal temporal features using the RNN-based neural network, water stress will be estimated with high accuracy considering complex temporal dependency. For the first issue, we use optical flow in the pre-extraction of leaf wilting, then we propose a novel water stress estimation method using RNN-based multimodal neural network. The second problem is that the existing method cannot consider environmental conditions end-to-end. The multimodal neural network and SW-SVR are trained independently in the existing method. If an environmental condition is lost in the multimodal neural network training, SW-SVR cannot consider the lost environmental conditions because clustering cannot find the lost conditions. We assumed that leaf wilting and environmental data have a different role in estimating stem diameter variations in Section 2.1. Environmental data not only explains water stress indirectly but also defines the importance of leaf wilting for estimating stem diameter variations (Wakamori and Mineno, 2019). Therefore, end-to-end training that prevents loss of environmental conditions will improve the accuracy of water stress estimation using the two types of data. For the second issue, we propose a novel neural network modeling method, called C-Drop, which can dynamically control the network based on environmental data. The proposal improves the accuracy of water stress estimation by temporal multimodal feature extraction and end-to-end consideration of environmental conditions.

3. Proposed method

3.1. Overview

We propose a novel plant water stress estimation method using a multimodal neural network with C-Drop to support decision-making of irrigation in stress cultivation (Fig. 1). The proposed method considers the leaf wilting and environmental features as the key input features. Other common features support the interpretation of wilting and environmental features for water stress. The wilting features express the water stress directly. The environmental features are related to transpiration, which is the cause of stress. Given that wilting and environmental features have a different role for estimating plant water stress, water stress is expressed multilaterally by fusing these features in a neural network as an attention mechanism. The water stress estimation model is built by the multimodal neural network that includes LSTM layers, with RNN being one of the layers. Multimodal neural networks are suitable for data with temporal dependence relationships that explain multimodality such as plant water stress. In the proposed method, the neural network extracts multimodal features to estimate water stress by network fusion while interpreting temporal dependence between the input features and water stress using LSTM layers. LSTM can solve more complex temporal dependency than traditional RNN by using memory units instead of normal units. In the memory unit, the memory cell memorizes information. The input, output and forget operations of the memory cell are controlled by three gates. The gate control solves the gradient explosion or vanishing problem of traditional RNN. Thus, LSTM can explain complex temporal dependency by using the memory unit. In addition, it is claimed that multiple LSTM layers can improve the accuracy of time-series modeling (Graves et al., 2013). Therefore, the proposed method uses a multimodal neural network including multiple LSTM layers to solve the complex and temporal dependence relationship between input features and water stress. In the existing method (Kaneda et al., 2017), CNN has been used to extract leaf wilting features. However, our method replaces CNN with pre-extracting wilting features to inhibit the increment of the parameters of the neural network, even if LSTM layers are included. In other words, our neural network focuses on temporal feature extraction by LSTM instead of spatial feature extraction by CNN. CNN is extremely effective in the identification of various objects that are present in an image such as general object detection (Liu et al., 2016). However, we assume that CNN has much higher expressiveness to recognize leaf wilting. Given that the leaf wilting is defined by the angle of leaves, optical flow (traditional image processing) can express the wilting from time-series plant images. In addition, as CNN has high expressiveness, it causes overfitting in training images that have low angular diversity owing to fixed point measurement of plant images. By replacing CNN with optical-flow-based feature pre-extraction, network parameters are reduced, and overfitting can be prevented by the network. Thus, we apply the pre-extraction of wilting features using optical flow to prevent overfitting and to focus temporal feature extraction in the multimodal neural network. In our approach, the architecture of multimodal network is not limited to a particular architecture. Various architectures of multimodal neural networks have previously been proposed (Veličković et al., 2017; Duan et al., 2018) but our proposed method, including designed input features and C-Drop, can be applied in these architectures. C-Drop promotes a multi-modal neural network to fuse features effectively via end-to-end consideration of environmental conditions. C-Drop is a neural network modeling method based on environmental features clustering and generates multiple sub-networks in a neural network based on the clustering result. Because the subnetworks train and infer each assigned specific data, each subnetwork becomes an estimation model specialized for each environmental condition.

As leaves move due to water stress in 3-dimensional (3D), this movement information could improve the estimation model. We attempted to obtain image depth information using the RGB-D camera (RealSense D435, Intel Corporation). Although the error is in the range of a few centimeters from the installation site, the leaf wilting motion is also in the range of a few centimeters. Therefore, this study uses optical flow data as 2D leaf movement that can be measured from a few millimeters with a high degree of accuracy at the present time. If the depth data was more accurate, the proposed method could be applied using such multimodal data.

3.2. Design of input features

Input features consist of wilting features, common features, and environmental features, as listed in Table 1. Wilting features and environmental features are key modalities to multilaterally explain water stress. The common features support the interpretation of these two modalities in a multimodal neural network. Wilting features express the movement of the leaf between two time-points. They are extracted by the following procedures: optical flow and masked optical flow. The process of extracting wilting features is shown in Fig. 2. First, an optical flow (a motion quantitation method) is used to calculate the leaf wilting motion. Optical flow quantitates the motion of objects based on the spatiotemporal variation between images taken at two time-points. The motion is calculated in pixels and each pixel motion is expressed by optical flow vectors that have angle and magnitude. The proposed method uses DeepFlow (Weinzaepfel et al., 2013), an optical flow algorithm, similar to the existing study (Kaneda et al., 2017). DeepFlow calculates the dense optical flow and the motion of non-rigid objects such as plant leaves. Thus, DeepFlow can determine the leaf wilting motion easily and robustly from the plant image. Second, we use excess-green (ExG) basedmasking for optical flow image to remove noise from the outside of the plant area. ExG is a general segmentation method for the plant area in an image (Jiang et al., 2018). The plant image has a complex background and optical flow, so, noise will be detected in the background. Therefore, we applied the ExG based mask to the optical flow and created a masked optical flow that reduced the background noise. Finally, we calculated the 11-dimensional wilting features from the masked optical flow, which consist of histogram features (6 dimensions) and statistical features (5 dimensions). The histogram

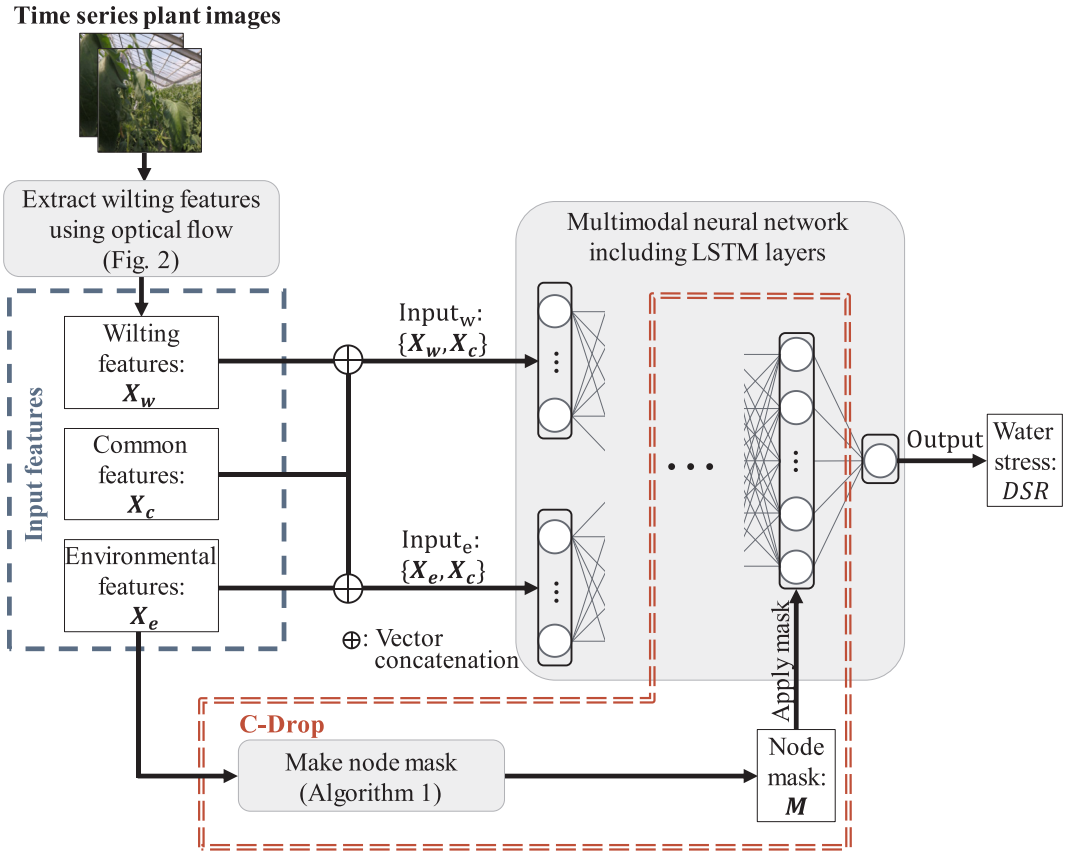

Fig. 1. Overview of proposed method.

Table 1
Input features for estimating plant water stress.

Type	Feature	# of dimensions
Wilting features: X_w	Histograms of oriented optical flow (HOOF)	6
	Mean of optical flow angle	1
	Standard deviation of optical flow angle	1
	Mean of optical flow magnitude	1
	Standard deviation of optical flow magnitude	1
	Optical flow detection ratio	1
Common features: X_c	Elapsed time from sunrise	1
	Irrigation flag	1
Environmental features: X_e	Temperature	1
	Relative humidity	1
	Vapor pressure deficit	1
	Scattered light	1

feature is defined using the histograms of oriented optical flow (HOOF) (Chaudhry et al., 2009), which was proposed for human action recognition. The HOOF is calculated based on the angle and the magnitude of optical flow vectors of all pixels. In our method, the plant image-based HOOF can differentiate between leaf wilting (downward motion) and leaf recovering (upward motion). The frequencies of HOOF change according to the distance between the leaves, the camera and the wilting area. Thus, we used the statistic features as reference values for HOOF. The statistic features include the average and standard deviation (SD) of the angle and magnitude, and the optical flow detection ratio obtained by dividing the number of detected optical flows by the number of pixels. The average and SD values are considered as the reference values for the distance between leaves and camera, and the

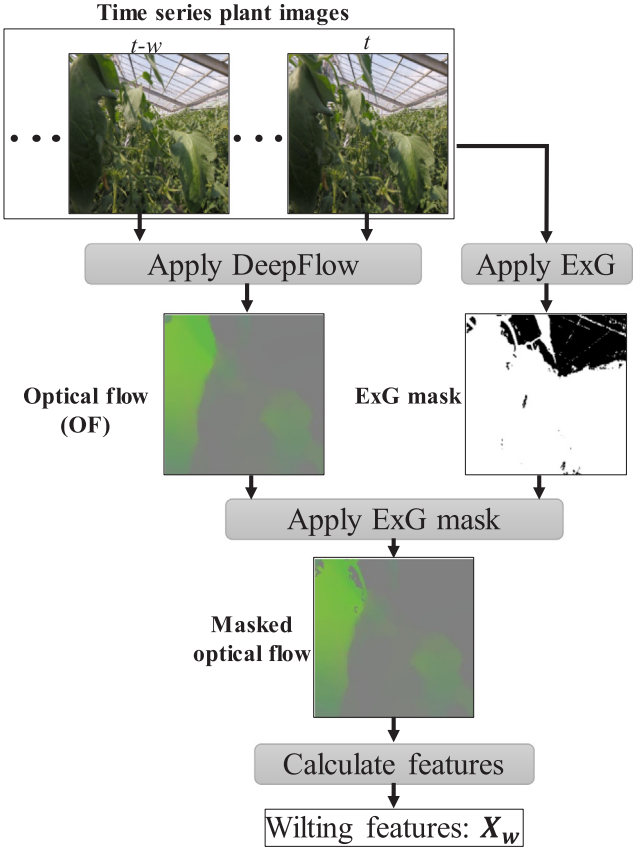

Fig. 2. Process of wilting feature extraction.

optical flow detection ratio will be the reference value for the wilting area.

Environmental features characterize climatic conditions and explain the plant transpiration speed, a water stress factor (Chanseetis et al., 2005). The transpiration rate is a key factor for water stress because it defines the easiness to lose water of plants. In this study, we used temperature, relative humidity, vapor pressure deficit (VPD) and scattered light as environmental feature inputs to explain transpiration rate. The transpiration speed of plants is controlled by stomatal opening and the leaf-air vapor pressure difference (LVPD) (Jolliet and Bailey, 1992; Nereu, 2003). LVPD is the difference between the leaf vapor pressure of a plant and the atmosphere. When the stoma is opened and the leaf vapor pressure is higher than the atmosphere vapor pressure, the plant experiences a high transpiration rate. The scattered light and VPD explain stomatal opening. Temperature, relative humidity, and VPD can express the vapor pressure of the atmosphere. Additionally, leaf vapor pressure could be expressed by the irrigation flag (described in the next paragraph) because leaf vapor pressure is controlled by water absorption from irrigated culture. Thus, environmental features could express the transpiration rate. This environmental data is collected by a wireless scattered light sensor node (Ibayashi et al., 2016). Scattered light is the amount of sunlight unaffected by the shadows of the steel pipes in a greenhouse (Oishi, 2016). The scattered light sensor node is cube-shaped, coated with vinyl chloride, except for one surface, and a silicon photodiode (S1133-14, Hamamatsu Photonics, K.K.) which measures the scattered light in the cube. The silicon photodiode can measure the light scattered around a plant, not the direct sunlight affected by the steel pipes.

The common features are not categorized into wilting or environmental modality. They express the plant response. Plants have a response related to the brightness throughout the day called diurnal variation or circadian rhythm (Meng et al., 2017; Moriyuki and Fukuda, 2016). The time elapsed after the sunrise can explain the diurnal variation that cannot be expressed by the wilting and environmental data. The irrigation flag is a binary variable, which denotes if the plants are getting irrigated at each time point. The common feature compensates the diurnal variations and the irrigation that is not extracted from the wilting and environmental features.

3.3. C-Drop

We proposed a new modeling method for neural networks: C-Drop for end-to-end consideration of environmental conditions. C-Drop creates subnetworks in the neural network by masking the nodes based on environmental features followed by the specialization of each subnetwork for a specific environmental condition. The proposed method extracts temporal multimodal (such as physiological and meteorological) features from leaf wilting features (physiological data) by using environmental features (meteorological data) as an attention mechanism of a multimodal neural network that includes LSTM layers.

The basic algorithm is composed of defining node masks based on environmental features and creating subnetworks by applying the masks (dotted double red[1] line area in Fig. 1). The wilting features and environmental features used in the proposed method have a different role in estimating the plant water stress, as described in Section 2.1. The environmental features can define the importance of wilting features on the water stress estimation. In the leaf wilting environment, the importance of wilting features should be increased, but in the environment without leaf wilting, the importance of wilting features should be ignored as noise. The specialized modeling of water stress in each environmental condition is necessary. Each subnetwork created by C-Drop extracts multimodal features specialized in each specific environmental condition, followed by each subnetwork accurately estimating the water stress by considering the dynamic importance of wilting features. The details of the algorithm for node masks to create subnetworks, which is a key algorithm in the C-Drop, are shown in Algorithm 1.

First, C-Drop performs clustering including preprocessing on the environmental features to find the latent environmental conditions (lines 1–2 in Algorithm 1). C-Drop transforms environmental features $\boldsymbol{X_e}$ using kernel approximation (Rahimi and Recht, 2007) and principal component analysis (PCA) as preprocessing. Next, the k-means clustering (MacQueen, 1967) generates clusters from these transformed features. The improved estimation performance by using C-Drop is related to the feature space for clustering. If the clustering result cannot find latent environmental cluster effective for water stress estimation, then the estimation accuracy will be reduced. The environmental features space must be tuned to enable the clustering algorithm to find the latent clusters because the environmental features (temperature, relative humidity, and solar radiation) have complex relationships with each other. In C-Drop, the kernel approximation and PCA generate the transformed feature space. Kernel approximation generates a new feature space in higher dimension and converts a linear algorithm to a nonlinear algorithm with low computational complexity. By combining kernel approximation and PCA, which is a linear algorithm, a new nonlinear feature space is created that consists of the features effective for clustering, while preventing an increase in computational complexity. Kernel PCA (Scholkopf and Smola, 1998) is also a nonlinear feature mapping method that can be applied as a preprocessing in the C-Drop. However, it has a large computational complexity in the training phase because it uses a general kernel function such as radial basis function (RBF) kernel. The training time in neural networks is higher, therefore, low computational complexity is required in the preprocessing. In this research, we adopted the combination of kernel approximation and PCA as a realistic example for calculating the complexity in the training phase. After the preprocessing, k-means extracts clusters in the transformed environmental features $\boldsymbol{X'_e}$. The k-means is one of the most popular non-hierarchical clustering algorithms and it can classify data faster under multiple clusters as compared to other clustering algorithms. The k-means extracts cluster centers $\boldsymbol{g_i}(i = 1. ..k)$, where each center represents an environmental condition. The inverse of Euclidean distance between all cluster centers is used as the similarity index of specific environmental characteristics in C-Drop.

Algorithm 1 (*Definition of node mask vectors for one layer*).

Input:
Environmental features: $\boldsymbol{X_e}$
Number of units: u
Drop ratio: r
Number of clusters: k
Output:
Mask vectors: $\boldsymbol{M} = \{\boldsymbol{M_1}, \boldsymbol{M_2}, \cdots, \boldsymbol{M_k}\}$
Definition of node mask vectors:
1. $\boldsymbol{X'_e} \leftarrow$ fit kernel approximation and PCA to $\boldsymbol{X_e}$ ▷ Preprocessing for environmental features
2. $\boldsymbol{g_i} \leftarrow$ each center of k-means $(\boldsymbol{X'_e})$, $i = 1 \ldots k$
3. $l \leftarrow \frac{u}{k}$ ▷ l: Number of initial active nodes allocated per cluster
4. $p \leftarrow \frac{u*(1-r)}{l}$ ▷ p: Number of clusters sharing active nodes to satisfy drop ratio r
5. **For** $c = 1$ to k do ▷ Assign initial active nodes for each cluster
6. $\quad \boldsymbol{M_c'} = \left\{ m_1, m_2, \cdots, m_u \middle| \begin{matrix} m_i = 1, i > l*(c-1) \quad and \quad i \le l*c \\ m_i = 0, i \le l*(c-1) \quad or \quad i > l*c \end{matrix} \right\}$
7. **For** $c = 1$ to k do ▷ Generate a node mask $\boldsymbol{M_c}$ for each cluster to satisfy drop ratio r
8. $\quad$ *compute distances between* $\boldsymbol{g_c}$ *and all* $\boldsymbol{g_i}$, $i = 1 \ldots k$
9. $\quad$ *sort the computed distances*
10. $\quad \boldsymbol{I_c} \leftarrow$ *select index of p neighbor clusters for the center of* $\boldsymbol{g_c}$
11. $\quad \boldsymbol{M_c} \leftarrow$ *compute logical OR in all* $\boldsymbol{M_i'}(\boldsymbol{i} \in \boldsymbol{I_c})$

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

Next, C-Drop determines to drop or not for each node based on the clustering result. The definition process requires two parameters: number of target nodes u and drop ratio r, in the addition cluster centers $\boldsymbol{g_i}(i = 1. ..k)$. The details of the definition process are described as follows. First, the process calculates l, the number of initial active nodes allocated per cluster, and p, the number of clusters sharing active nodes to satisfy drop ratio r (lines 3–4 in Algorithm 1). Then, initial active nodes are assigned for each cluster (lines 5–6 in Algorithm 1). When the drop ratio r is not satisfied, active nodes are stored in neighbor clusters. The environmental conditions neighbor clusters require similar features for water stress estimation. Therefore, the final masks for each cluster are generated by sharing active nodes in p neighbor clusters (lines 7–11 in Algorithm 1). The sharing of active nodes enables the neural network to extract similar features efficiently between p resembling environmental conditions. In addition, the number of clusters and drop ratio r are tuned into hyperparameters, and the C-Drop creates effective subnetworks and node sharing for high accuracy water stress estimation.

Although Dropout (Nitish Srivastava et al., 2014) creates multiple subnetworks in a neural network by masking nodes like C-Drop, the purpose and the inference process are different. The purpose of the Dropout is to prevent a trained model from overfitting. Dropout masks nodes randomly with the drop ratio r_d in the training phase and outputs of the nodes are multiplied with r_d^{-1} without masking in the inference phase. This improves the generalization performance of each node and prevents overfitting in the neural network. In addition, the C-Drop masks nodes in the training and the inference phase by the same process, considering the environmental conditions end-to-end. The masks in C-Drop are generated algorithmically based on environmental features without randomness. Each subnetwork created by C-Drop uses weights trained for each environmental condition and can perform high accuracy inference suitable for the environmental condition. Moreover, to prevent overfitting in each specialized subnetwork, C-Drop, which creates the subnetworks, can use Dropout to prevent the overfitting of a neural network.

In addition, the existing attention mechanism (Luong et al., 2015; Xu et al., 2015) improves RNN by selectively focusing on important data points in the temporal direction by calculating a weighted average of the hidden layer in the past and the current state. In addition, C-Drop focuses on the modal direction in the input feature to select important input modality that dynamically changes based on the environmental conditions. Thus, C-Drop is designed to adjust the importance of the features based on the environmental conditions at a specific point of time as a more versatile attention mechanism.

3.4. Definition of water stress index

The proposed method uses stem-diameter-based water stress index as the dependent variable in the training phase. As water stress decreases the amount of water in the cells of a plant, it shrinks the stem diameter. The plant water stress can then be quantitated based on the stem diameter variations by a laser displacement sensor, which is measured over time with non-disruptive measurement for plants. Therefore, we can collect true data of plant water stress for machine learning in actual cultivation using a laser displacement sensor. We adopted stem-diameter-based water stress index as the dependent variable in neural network training.

The stem diameter represents the plant water stress. However, we cannot use the diameter as a water stress index directly because the diameter changes with the growth of the plant and the diurnal variation. For this issue, we defined the difference in stem diameter calculated using the most recent irrigation (DSR) as a water stress index. The DSR value $dsr_t$is calculated as follows:

$$dsr_t = max(stem_{t-n}, stem_{t-n+1}, \cdots, stem_t) - stem_t$$

where t is the current time and n is the time elapsed since the recent irrigation. DSR is a value calculated by subtracting the current stem diameter from the maximum stem diameter since recent irrigation. The variations based on plant growth and diurnal variation are almost removed because the index is based on the irrigation timing, which is repeated several times a day.

In DSR based irrigation control, a threshold value is defined for the DSR. When the DSR value exceeds the threshold, an irrigation system irrigates the plants. After this, the stem diameter increases owing to water uptake, and the DSR is maintained at 0. Then, the stem diameter decreases due to water stress and the DSR increases and exceeds the threshold again. Therefore, DSR based irrigation scheduling can control irrigation based on plant water stress status.

4. Evaluation

4.1. Dataset

We evaluated multimodal neural network with C-Drop using actual cultivation datasets. To construct the datasets, we collected plant images, environmental data (temperature, relative humidity, VPD, and scattered light), and stem diameter data from three pinched tomato plants (*Solanum lycopersicum* L. cultivar Frutica) in a dense cultivation. Each plant was planted in a rockwool culture (Yasaihana-pod, Nippon Rockwool Corporation), and the plant density was 148 plants/m^2. Because the rockwool culture was shaped like a small cube (6 cm × 6 cm × 6 cm), the roots of the plant were restricted in the cube. The data was collected at commercial greenhouses in Fukuroi, Japan, from Dec. 22, 2017 to Jan. 8, 2018, Apr. 7, 2018 to May 24, 2018 and June 23, 2018 to July 19, 2018. In each period, we collected plant images, environmental data and stem diameter data of the three tomato plants. To measure this data, RGB cameras (GoPro HERO5 Session, GoPro Inc.), wireless scattered light sensor nodes (Ibayashi et al., 2016), and laser displacement sensors (HL-T1010A, Panasonic Corporation) were installed for each target plant, as shown in Fig. 3(a) and (b). Each sensor device measured the data in same installation conditions for data collection periods. The RGB cameras were installed at a location with the highest number of leaves captured in an image. The laser displacement sensors measured stem diameter variations between the 9th and 10th nodes of each target plant, and the measured signals were logged by a data logger (midi LOGGER GL840, GRAPHTEC Corporation). Wireless scattered light sensor nodes were installed above each target plant to collect environmental data that included temperature, relative humidity, VPD, and scattered light.

In addition, plant images were collected from 4 a.m. to 7p.m., and the environmental and stem diameter variations data were collected for 24 h. These sensors were similarly installed in all three cultivations of the tomato plant and collected continuously every minute until the end of the experiment because tomatoes require frequent irrigation every day. However, owing to defects such as sensor failure, there are several points when data was lost, and the number of datasets differed for each cultivation of tomato plants. Using the collected data from 7 a.m. to 6p.m. in bright sunshine, we defined three datasets for cross-validation, as shown in Fig. 3(c)–(e), and Table 2. In each dataset, training/validation data and testing data are independent of the target plants and the day of measurement. Specifically, the data measured on the same day are not included in both training/validation data and testing data. The images of the same target plant have high similarity even if their days of measurement are different. The plant images of different target plant should be used for the training/validation and testing data. The environmental data measured on the same day has similar variations even if the target plants are different. Therefore, the environmental data for different days should be used for the training/validation and testing data. Separate target plants and measurement days are required in training/validation and testing data to generalize this evaluation. The evaluation in the existing study did not considered this independence (Kaneda et al., 2017) but we designed new datasets

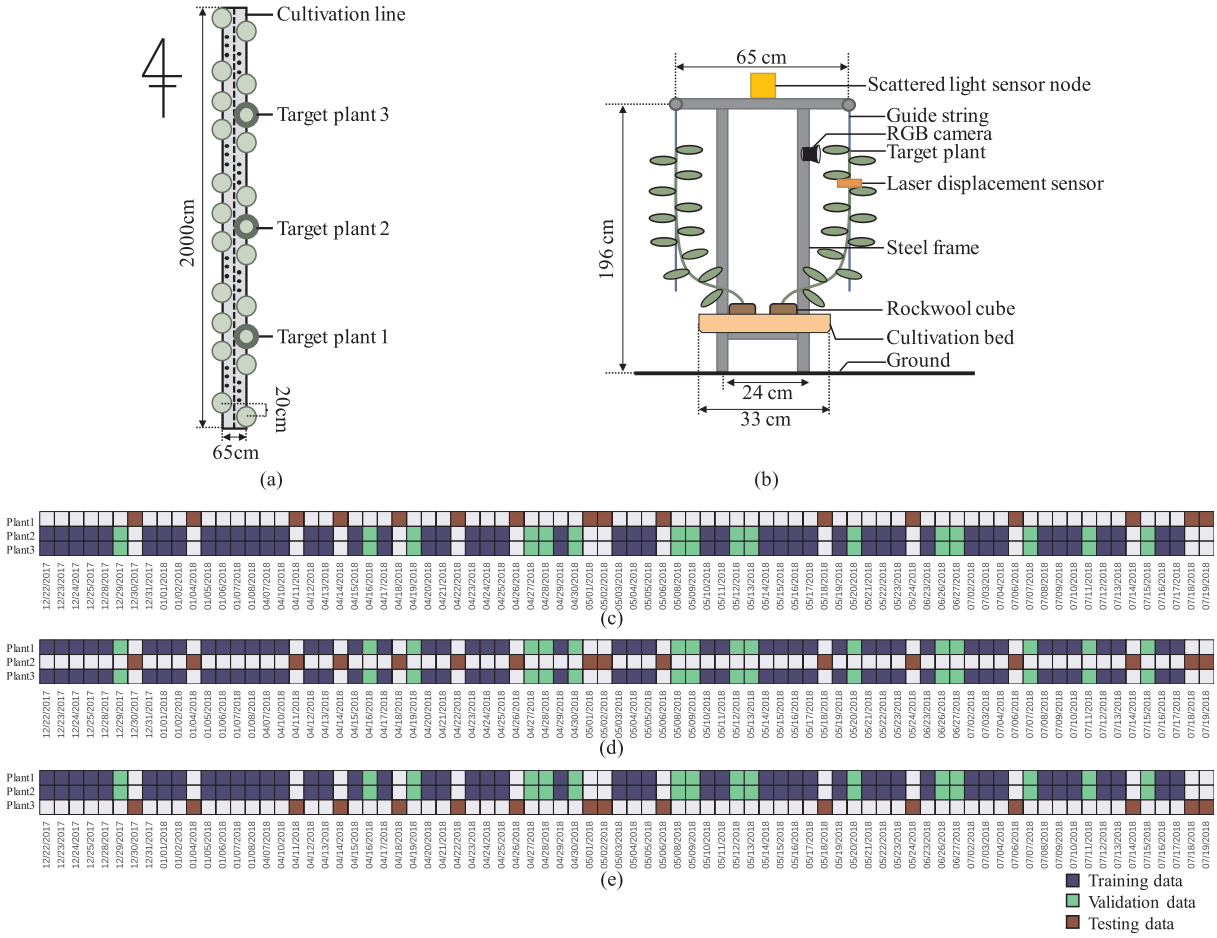

Fig. 3. Dataset for the evaluation. (a) and (b) show data collection environment, (a) overhead view of the cultivation line where data was collected, (b) layout of measurement sensors for a target plant. (c), (d) and (e) show datasets which test the plants 1, 2 and 3, respectively.

considering the independence for this evaluation. We then defined the data types: training, validation, and testing data, for each day based on the daily average scattered light to prevent season and weather bias. We assumed that the proposed estimation model is being used in real cultivation that is affected by seasonal and weather characteristics similar to the training data. Therefore, we defined the data type of each day based on the daily average scattered light representing the seasonal and weather characteristics. Approximately 60%, 20% and 20% of all days were defined as training, validation, and evaluation data types, respectively, in order of the daily average scattered light. Thus, we evaluated the proposed method assuming actual cultivation application by using the datasets that considered independence and data bias.

For data augmentation, we used cropped plant images for training/validation data. A plant image includes the area without leaf wilting, thus, general image crop methods such as center and random crop are not appropriate. With these methods, the leaf wilting area in the original image may be lost. Thus, we cropped the images based on optical flow that detects leaf wilting and crops it from the original image. The cropped area is determined to maximize the daily average optical flow magnitude. We calculated optical flow using DeepFlow in the same algorithm with wilting feature extraction. Consequently, the optical-flow-based crop generated new images in which the wilted part of the target plant was enlarged. The cropped area was determined daily based on the target plant because the area with wilted leaf depends on the view angle and the plant growth stage. The size of the cropped area was a quarter of the original image. Finally, we resized the original and cropped images to the same size (144 × 144) and extracted the leaf wilting features as shown in Fig. 2.

4.2. Experimental settings

We evaluated the performance of the proposed method through two experiments: comparison experiment and ablation experiment. In both experiments, we applied cross-validation using three datasets shown in Fig. 3(c)–(e), and compared the average of the testing score. In addition, we demonstrated the neural network architectures, details of the comparison method, and details of the hyperparameters used in this evaluation in Fig. 4, Table 3, and Table 4, respectively. In Fig. 4, $\text{Input}_{\text{img}}$ denotes the RGB image, Input_{w} denotes the leaf wilting modality consisting of wilting features and common features, and Input_{e}

Table 2
Number of under training, validation and testing data.

Dataset	# of training data (after augmentation)	# of validation data (after augmentation)	# of testing data
dataset 1 (Fig. 3(c))	61,873 (123,746)	21,420 (42,840)	9,864
dataset 2 (Fig. 3(d))	61,970 (123,940)	21,706 (43,412)	10,382
dataset 3 (Fig. 3(e))	61,549 (123,098)	21,298 (42,596)	10,474

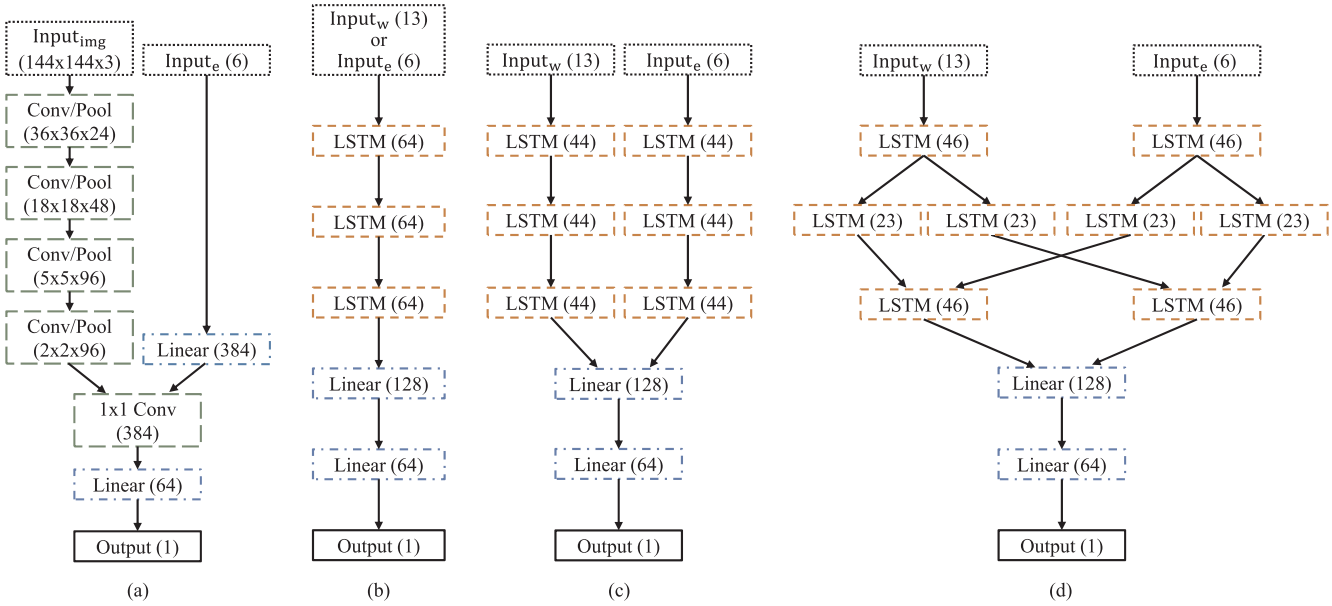

Fig. 4. Neural network architectures, (a) existing deep neural network (DNN) (Kaneda et al., 2017) (b) single modal network based on LSTM (LSTM), (c) multimodal neural network named two stream LSTM (2sLSTM), (d) multimodal neural network named cross-modal LSTM (X-LSTM). The numbers in parentheses mean number of dimensions.

denotes the environmental modality composed of environmental and common features. Fig. 4(a) is the existing deep neural network (DNN) that includes CNN proposed in (Kaneda et al., 2017). Input$_{img}$ is used in the existing DNN. Fig. 4(b) shows a single-modal network with LSTM layers to evaluate the performances of each modality (wilting modality and environment modality). Fig. 4(c) and (d) show the multimodal neural networks with LSTM layers that evaluate the multimodal input data and C-Drop. As a method of combine the two LSTMs (Fig. 4(a) and (b)), multiple networks were combined into one network by simply combining the last layers of the two networks. Fig. 4(c) is a typical two-stream architecture that extracts each modal feature through respective streams. In this study, the network extracts the leaf wilting temporal features and environmental temporal features through respective LSTM-based streams. Fig. 4(d) shows a recent state-of-the-art architecture for multimodal time-series data known as X-LSTM (Veličković et al., 2017). We applied the proposed method in the two networks (Fig. 4(c) and (d)) in this evaluation. Our proposed method focused on designing the input features for LSTM-based neural networks and C-Drop algorithm for estimating plant water stress. Thus, neural network is not limited to a specific architecture. We evaluated the proposed method in typical network architecture (Fig. 4(c)) and state-of-the-art architecture (Fig. 4(d)). In all neural networks, parametric rectified

Table 3
Details of the evaluation.

(a) Settings of the comparison experiment between the proposed and the existing methods.

Method	Input data			
	RGB image	Wilting features	Common Features	Environmental features
XGBoost (WILT, ENV)		✓	✓	✓
DNN (ORGIMG, ENV) (Fig. 4(a))	✓ (Original)		✓	✓
DNN (ORGIMG, ENV) (Fig. 4(a)) w/SW-SVR	✓ (Original)		✓	✓
DNN (ROAFIMG, ENV) (Fig. 4(a))	✓ (ROAF)		✓	✓
DNN (ROAFIMG, ENV)w (Fig. 4(a)) w/SW-SVR	✓ (ROAF)		✓	✓
2sLSTM (WILT, ENV) (Fig. 4(c)) w/C-Drop		✓	✓	✓
X-LSTM (WILT, ENV) (Fig. 4(d)) w/C-Drop		✓	✓	✓

(b) Settings of the ablation experiment of the proposal.

Method	Input data		
	Wilting features	Common Features	Environmental features
LSTM (WILT) (Fig. 4(b))	✓	✓	
LSTM (ENV) (Fig. 4(b))		✓	✓
2sLSTM (WILT, ENV) (Fig. 4(c))	✓	✓	✓
2sLSTM (WILT, ENV) (Fig. 4(c))w/SW-SVR	✓	✓	✓
2sLSTM (WILT, ENV) (Fig. 4(c)) w/C-Drop (w/o preprocessing)	✓	✓	✓
2sLSTM (WILT, ENV) (Fig. 4(c)) w/C-Drop	✓	✓	✓
X-LSTM (WILT, ENV) (Fig. 4(d))	✓	✓	✓
X-LSTM (WILT, ENV) (Fig. 4(d)) w/SW-SVR	✓	✓	✓
X-LSTM (WILT, ENV) (Fig. 4(d)) w/C-Drop (w/o preprocessing)	✓	✓	✓
X-LSTM (WILT, ENV) (Fig. 4(d)) w/C-Drop	✓	✓	✓

Table 4
Details of hyperparameters used in evaluation. Hyperparameters with multiple values are tuned using grid-search.

Hyperparameter	Value(s)
(a) Hyperparameters for neural network.	
Learning rate	0.01
Batch size	1024
Dropout ratio	0.3, 0.5, 0.7
(b) Hyperparameters for LSTM-based neural network.	
Sequence length	60
(c) Hyperparameters for C-Drop.	
Drop ratio	0.3, 0.5, 0.7
Gamma of kernel approximation	0.1, 1, 10
Component number of kernel approximation	100
Cumulative contribution rate of PCA	99%
Number of clusters	64
(d) Hyperparameters for SW-SVR.	
Cost: C	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
Tube: ε	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
Prediction weight	0.5, 1, 3
Number of estimators	10, 100, 300
Data collection size: k	100, 1000
(e) Hyperparameters for XGBoost.	
Learning rate	0.05, 0.1, 0.3, 0.6, 0.9
Max depth of a tree	3, 5, 10
Number of estimators	10, 50, 100, 150, 200, 250, 300
Subsample ratio of the training instances	0.8, 0.85, 0.9, 0.95
Subsample ratio of columns in each tree	0.3, 0.5, 1.0

linear unit (PReLu) (He, 2015), normalization processing, and Dropout (Nitish Srivastava et al., 2014) are applied to obtain the output of each hidden layer. For the normalization processing, we used batch normalization (Sergey Ioffe, 2015) in the linear or convolutional layer and layer normalization (Ba et al., 2016) in the LSTM layer. All weights of neural networks were initialized by He initialization (He, 2015), and the neural networks were trained using Adam (Diederik and Kingma, 2015) as the optimizer.

In the comparison experiment, we compared our proposed method and the existing method (Kaneda et al., 2017), as shown in Table 3 (a). "2sLSTM (IMG, ENV) w/C-Drop" and "X-LSTM (IMG, ENV) w/C-Drop" are our proposed methods which were composed using our proposed input features, multimodal neural network, and C-Drop. The two proposed methods use 2sLSTM and X-LSTM (Fig. 4(c) and (d), respectively). C-Drop was applied to obtain the output of the two linear layers to construct subnetworks based on features extracted via LSTM layers. The other DNN-based methods have been proposed in the existing study (Kaneda et al., 2017), which includes four types of combinations based on with or without preprocessing the input image and fine-tuning by SW-SVR. According to the existing study, the SW-SVR is applied to the 64-dimension features extracted from the last linear layer of the DNN (Fig. 4(a)). In the existing study, the DNN-based method was evaluated to predict future water stress. However, the irrigation control system is realized by estimating only the current water stress. From a practical perspective, we evaluated the performance of the estimated current water stress as an experimental setting. In the DNN-based method with SW-SVR, SW-SVR uses D-SDC to extract training data considering the prediction horizons for each weak learner as described in Section 2.3. In this experiment, to estimate the current water stress, we built each weak learner using the neighbor data k from the weak learner, and the data collection size of k was tuned as a hyperparameter. Furthermore, in XGBoost (Chen and Guestrin, 2016), the estimation was made using the feature with the same length as the sequence length of LSTM used in the proposed method as an explanatory variable.

Next, we evaluated the performances of the proposed input features and C-Drop in the ablation experiment. Table 3(b) shows the details of the ablation experiment. We compared the accuracy of estimation on single-modal approaches using the single-modal network (Fig. 4(b)) and multimodal approaches using multimodal neural networks (Fig. 4(c) and (d)) to evaluate the superiority of multimodality in input features. We then compared the C-Drop and other methods such as without C-Drop, SW-SVR and C-Drop without preprocessing to evaluate the performance of C-Drop including preprocessing. SW-SVR is a machine learning algorithm ensemble that includes feature clustering by k-means, like C-Drop. Thus, SW-SVR can contribute to build specialized models considering environmental conditions just like C-Drop. However, SW-SVR will not extract more effective features than C-Drop because SW-SVR cannot apply end-to-end training in a neural-network. When SW-SVR is applied to train neural networks, it cannot consider environmental characteristics. Thus, latent features and clusters may get lost while training the neural network. Moreover, as C-Drop performs end-to-end neural network training, it can extract effective features and clusters for improving the accuracy of estimation, as compared to SW-SVR. The preprocessing of C-Drop finds more latent clusters and improves the accuracy because the it clarifies the relationship between the environmental features. Latent clusters are composed of data located at large distances in the original feature space. The preprocessing can reduce these distances by transforming the space. After the preprocessing, the k-means algorithm finds the latent clusters in the clarified environmental features. Thus, we compared with or without C-Drop method, the SW-SVR and the C-Drop without preprocessing to evaluate the performance of C-Drop with preprocessing. The network parameters are numbered as 104 k, 102 k and 102 k for single-modal network (Fig. 4(b)), 2sLSTM (Fig. 4(c)) and X-LSTM (Fig. 4(d)), respectively. The three networks have almost equivalent accuracy and these parameters are less than the number of augmented training data (123 k) listed in Table 2.

The list of hyperparameters is shown in Table 4. The parameters with multiple values were tuned using grid-search in two experiments. The evaluation metrics are coefficient of determination (R^2), mean absolute error (MAE) and root-mean-squared error (RMSE). Because the hyperparameters of all models were tuned using the validation data, the models that had the lowest RMSE were selected. In this evaluation, the source code was implemented using Python2.7. We used Chainer 1.22 and scikit-learn 0.19.1 to implement our proposed method and the comparison methods.

4.3. Results and discussion

Fig. 5 shows the results of the comparison experiment. The results demonstrate that the proposed methods 2sLSTM (WILT, ENV) w/C-Drop and X-LSTM (WILT, ENV) w/C-Drop provided more accurate estimation than the existing methods for all evaluation metrics. The estimated performance of X-LSTM (WILT, ENV) w/C-Drop was the best and we confirmed the effectiveness of our proposed method for X-LSTM, which is a state-of-the-art multimodal neural network. In the existing methods, the ROAF image was used as an input image that improved the accuracy of estimation as compared to the original image. In addition, applying SW-SVR to DNN further improved the accuracy. The performances of image preprocessing and SW-SVR coincided with the results of the existing research (Kaneda et al., 2017). The effectiveness of SW-SVR, which is a clustering-based algorithm, supports the applicability of clustering to the features of water stress modeling. However, R^2 was 0.00 in the existing DNN method (ROAFIMG, ENV) w/ SW-SVR and XGBoost, which is a state-of-the-art regression model. The score denotes that the existing method approximately outputs the average of true values in three datasets. DNN and SW-SVR were difficult to extract effective features to estimate appropriate plant water stress. Moreover, even if it used the state-of-the-art regression method like XGBoost, it could not make estimations with an accuracy higher than

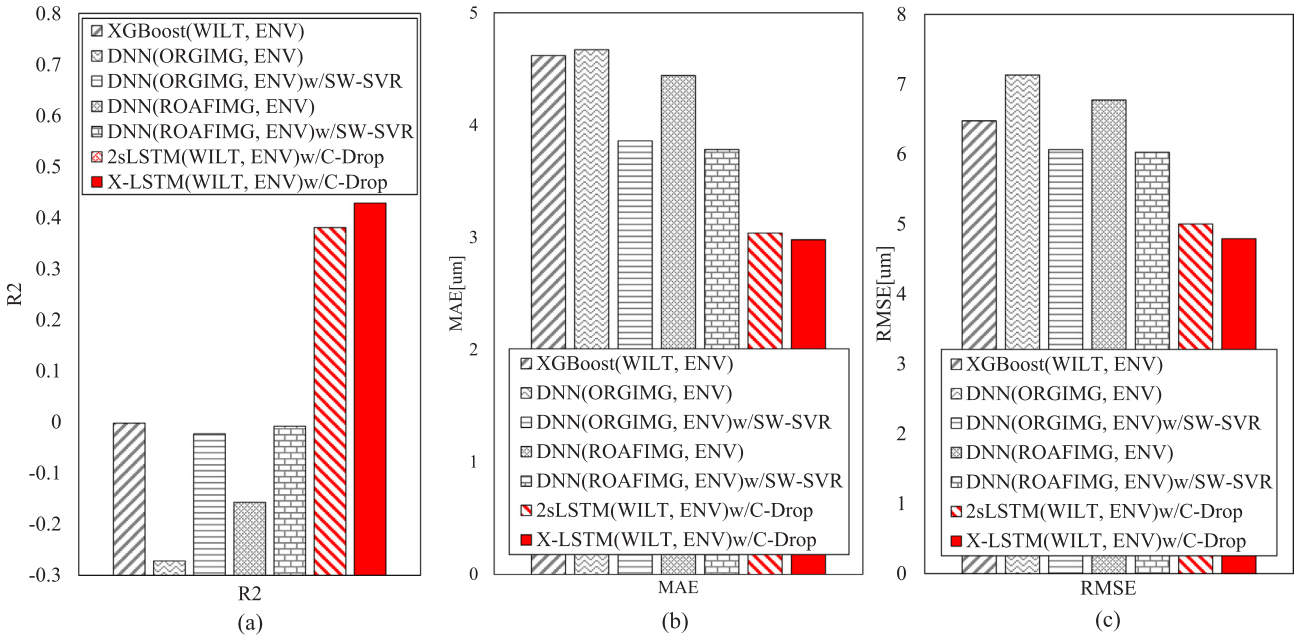

Fig. 5. Results of the comparison experiment between the proposed and existing methods: (a) R^2, (b) MAE, and (c) RMSE.

the proposed method only with the input feature quantity. Thus, the existing method had room for improvement. On the contrary, our proposed methods improved the accuracy of estimation for all metrics. The R^2 score was 0.381 and 0.429 in 2sLSTM (WILT, ENV) w/C-Drop and X-LSTM (WILT, ENV) w/C-Drop, respectively. The X-LSTM (WILT, ENV) w/C-Drop consists of the best algorithm and features that reduce the estimation errors of MAE and RMSE by approximately 21% each, as compared to DNN (ROAFIMG, ENV) w/SW-SVR. In this evaluation, the dataset was independent of the target plant and the day of measurement for training/validation and testing data, as shown in Fig. 3(c)–(e). The proposed methods improved the accuracy of estimation in the independent testing data. Therefore, it was suggested that the proposed methods have higher robustness as compared to the existing method for estimating plant water stress. Although, the dataset used for the evaluation belonged to one breed (*Solanum lycopersicum* L. cultivar Frutica) and three individuals (plants 1, 2, and 3), in future, there is a plan for verification of generality by increasing the number of breeds and individual data.

Fig. 6 shows the difference between the estimated and the true DSR in time-series. Hourly average values of estimated and the true DSR are shown since original data have a high frequency for time-series visualization. In 2sLSTM (WILT, ENV) with C-Drop and X-LSTM (WILT, ENV) with C-Drop (Fig. 6(c)), the multimodal characteristics enable the estimated DSR to change by more than approximately 30 μm and follow approximately 10 μm. In particular, X-LSTM (WILT, ENV) w/C-Drop demonstrated high accuracy during the period 2017/12/30 7 a.m. to 2018/7/19 2 p.m., where DSR exhibited a large value during the period 2017/12/30 to 2018/1/4 and a small value during the period 2018/1/4 to 2019/7/19, thereby indicating that it can follow the true value. In contrast, DNN (ORGIMG, ENV) (Fig. 6(a)) and DNN (ROAFIMG, ENV) (Fig. 6(b)) demonstrated a large error during the period 2017/12/30 to 2018/1/4 and slightly false value of the period between 2017/1/4 and 2018/7/19.

In Fig. 7, it is shown the results of the ablation experiment. Each multimodal approach demonstrated high accuracy compared to single-modal approaches (LSTM (WILT) and LSTM (ENV)). The results demonstrate that the estimation performance of X-LSTM (WILT, ENV) with C-Drop is the highest for the evaluation metrics. The difference in the accuracy of estimation are related to different feature performances. Thus, this result indicates that the combination of leaf wilting and environmental features improve the accuracy of estimation by complementarity explanation of plant water stress. The effect of the environmental features alone has not been evaluated in the existing study (Kaneda et al., 2017). However, our result confirmed the performance of the input features and the effectiveness of multimodalities in input features. Moreover, each metric score indicates higher performance of the proposed methods with C-Drop including preprocessing algorithms in each neural network (2sLSTM and X-LSTM). The results support the effectiveness of C-Drop on multimodal neural networks to improve the accuracy of estimation in both 2sLSTM, a typical architecture, and X-LSTM, a state-of-the-art architecture, of the multimodal neural network.

We determined that C-Drop can improve a neural network to estimating the water stress independent of the detailed network architecture. When the preprocessing was not applied to C-Drop, the accuracy of estimation decreased in both multimodal neural networks. Thus, the network that can implement environmental characteristics using kernel approximation and PCA was required to improve the accuracy of estimation in these neural networks. As the environmental features have complex relationships, only k-means could not find latent clusters for effective modeling of plant water stress. In this study, we used kernel approximation and PCA as preprocessing in C-Drop to inhibit increasing computational complexity. Besides, other space mapping methods such as kernel PCA can be used in C-Drop preprocessing as well. Therefore, the trade-off relationship between the computational complexity and the accuracy of estimation should be analyzed in future research (whereas, our results showed the necessity of preprocessing in C-Drop). SW-SVR was able to improve R^2 and RMSE score compared to that without SW-SVR in the two multimodal neural networks. In MAE, SW-SVR improved the accuracy in 2sLSTM. The effectiveness of SW-SVR for estimating plant water stress can be evaluated using the existing method. The effectiveness of C-Drop and SW-SVR, which have a similar process such as feature clustering, supports the effectiveness of clustering-based learning to estimate plant water stress. However, the combination of multimodal neural network and C-Drop provides more the accuracy of estimation than SW-SVR. It is assumed that C-Drop provides effective feature extraction and modeling by end-to-end training in a neural network, which is impossible with SW-SVR. Thus, it appears that applying SW-SVR to the trained neural network could not extract effective features because effective features have been reduced owing to the data imbalance in the training phase of the neural

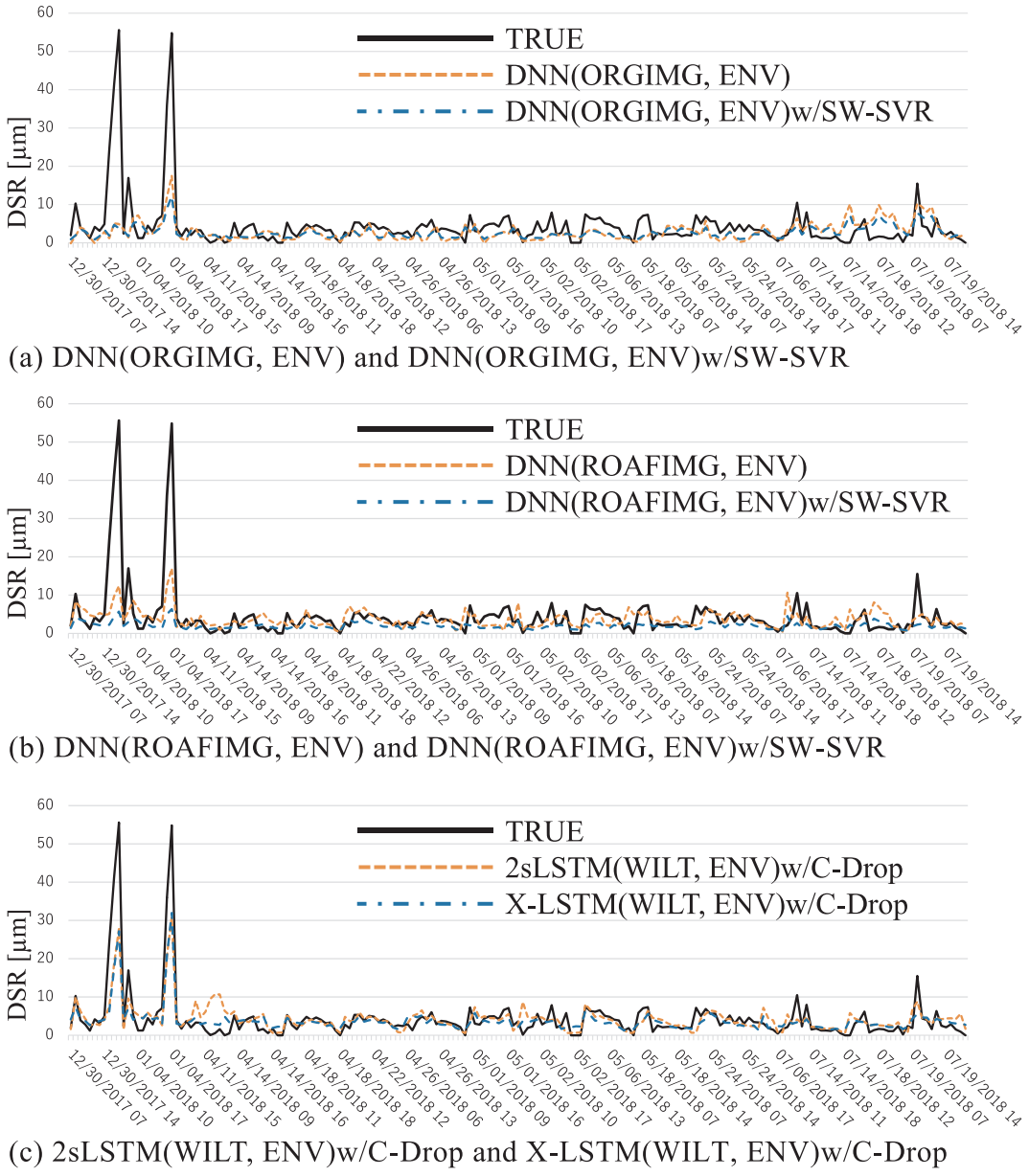

Fig. 6. True values and estimated values with DNN (ORGIMG, ENV), DNN (ROAFIMG, ENV), 2sLSTM (WILT, ENV) and X-LSTM (WILT, ENV) w/C-Drop.

network. Based on these advantages of C-Drop, we confirmed its effectiveness including preprocessing on multimodal neural networks based on typical architecture or state-of-the-art architecture.

4.4. Inference time

We measured the inference time of training model to verify that the proposed method is applicable to the real world. We used Intel core i7-6700 k, Nvidia GTX960 GPU, DDR4-2133 32 GB memory, and Ubuntu 14.04 as the operating system for verification and measured the average processing time after 100 inferences. Consequently, it took 64.94 ms processing time for each point. Therefore, assuming an application that estimates in a one-minute cycle, it is possible to satisfy the cultivation control of 923 locations with the PC we used for verification. Thus, our proposed method is proven to be worthy of application to the real world.

5. Conclusion

We proposed a novel plant water stress estimation method using a multimodal neural network with C-Drop to support the decision-making of precision irrigation during stress cultivation. The multimodal neural network includes multiple LSTM layers. The leaf wilting and environmental features are used as the key input features. In plant physiology, the leaf wilting and environmental features have a different role in the estimation of water stress. Therefore, our method extracts effective multimodal features for estimating plant water stress by combining the above-mentioned features in the neural network. In addition, we proposed a neural network modeling method, named C-Drop, to promote end-to-end consideration of environmental conditions, then build a high accuracy estimation model. C-Drop trains the neural network on the basis of environmental conditions that not only have an indirect relationship with water stress but also control the importance of wilting features in water stress estimation. C-Drop creates subnetworks in the neural network based on environmental features algorithmically, and each subnetwork can specialize in specific environmental conditions

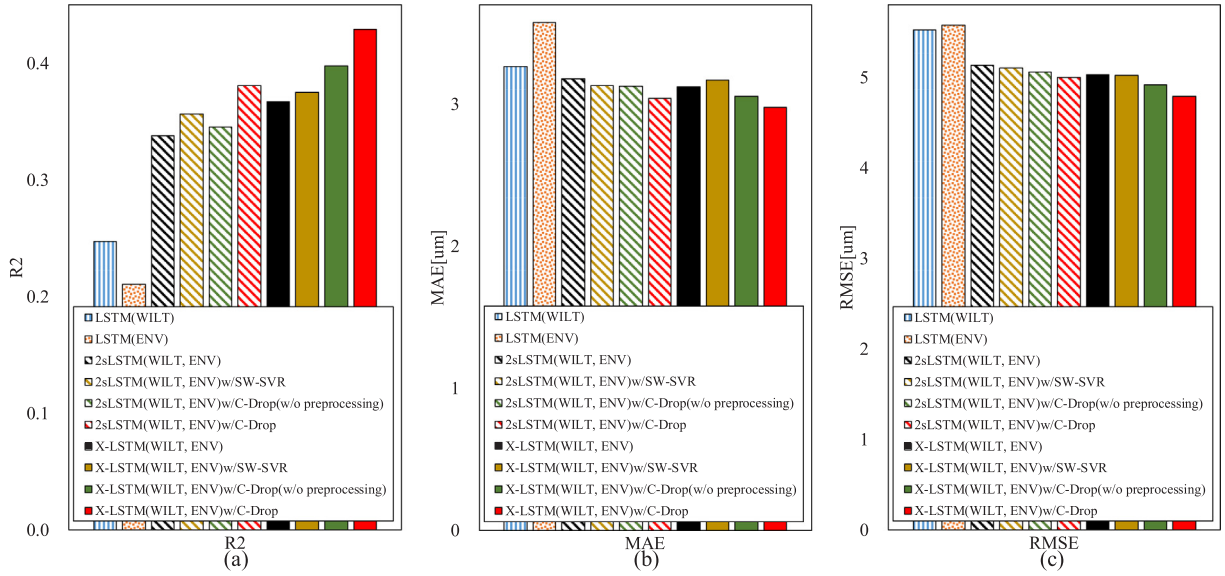

Fig. 7. Result of the ablation experiment with the following metrics: (a) R^2, (b) MAE and (c) RMSE.

and estimate water stress with high accuracy. We evaluated the proposed method using actual cultivation dataset of tomato. The result of the comparison experiment demonstrated that the proposed method estimates the plant water stress with 21% lower MAE and RMSE than the existing method (Kaneda et al., 2017). In the ablation experiment, the result showed the effectiveness of multimodalities in the input features. Moreover, it was confirmed that applying C-Drop to the multimodal neural network improves the accuracy of estimation as a more versatile attention mechanism.

There are four limitations and future work related in this study. Firstly, 3D movement of leaves should be evaluated in water stress estimation. Because the 3D leave movement characteristics has the potential to increase the estimation accuracy, we would examine the applicability of the RGB-D camera for leaf condition quantification in future study. Secondly, the time information such as the elapsed days for a specific growth event should be considered with the proposed method. Time information through cultivation is expected to improve the estimation accuracy by considering plant growth. Thirdly, the detailed performance of C-Drop should be evaluated. Because of the characteristics of C-Drop and of neural networks, there is a possibility that different applications of C-Drop together with neural network may affect the accuracy of estimation. We evaluated C-Drop in a limited experimental condition. Thus, its applicability and related concerns have not been clarified. In future work, it is necessary to clarify the effect of the mask created by C-Drop on the weighting of neural network. In addition, an exhaustive performance of C-Drop using more general dataset such as meteorological data should be evaluated. To conclude, we should investigate fruit quality cultivated by controlling irrigation with the proposed method. The performance of the current accuracy of water stress estimation should be evaluated by fruit quality compared with irrigation by true DSR and estimated DSR.

CRediT authorship contribution statement

Kazumasa Wakamori: Conceptualization, Methodology, Writing - original draft. **Ryosuke Mizuno:** Formal analysis, Investigation. **Gota Nakanishi:** Software. **Hiroshi Mineno:** Supervision, Writing - review & editing, Project administration.

Acknowledgments

This work was supported by Japan Science and Technology Agency (JST), PRESTO Grant Number JPMJPR15O5. We greatly appreciate Mr. Makoto Miyachi (Happy Quality Co., Ltd., Japan) and Mr. Daigo Tamai (Sun Farm Nakayama Co., Inc., Japan) for providing an environment for data collection and experiment.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2019.105118.

References

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer Normalization, ArXiv Prepr. ArXiv1607.06450.

Benabderrahmane, S., Mellouli, N., Lamolle, M., 2018. On the predictive analysis of behavioral massive job data using embedded clustering and deep recurrent neural networks. Knowledge-Based Syst. 151, 95–113. https://doi.org/10.1016/j.knosys.2018.03.025.

Boyer, J.S., 1967. Leaf water potentials measured with a pressure chamber. Plant Physiol. 42, 133–137.

Brillante, L., Mathieu, O., Lèveque, J., Bois, B., 2016. Ecophysiological modeling of grapevine water stress in burgundy terroirs by a machine-learning approach. Front. Plant Sci. 7, 1–13. https://doi.org/10.3389/fpls.2016.00796.

Chanseetis, C., Shinohara, Y., Maruo, T., Takagaki, M., Hohjo, M., 2005. An estimation of tomato transpiration for effective fertigation management system using integrated solar radiation and vapor pressure deficit. Environ. Control Biol. 43, 105–112. https://doi.org/10.2525/ecb.43.105.

Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R., 2009. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE Int. Conf. Comput. Vis. pp. 1932–1939.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. https://doi.org/10.1145/2939672.2939785.

Diederik, J.L.B., Kingma, P., 2015. Adam a method for stochastic optimization. Int. Conf. Learn. Represent.

Duan, J., Wan, J., Zhou, S., Guo, X., Li, S.Z., 2018. A unified framework for multi-modal isolated gesture recognition. ACM Trans. Multimed. Comput. Commun. Appl. 14, 1–16. https://doi.org/10.1145/3131343.

Fricke, W., 2017. Water transport and energy. Plant Cell Environ. 40, 977–994. https://doi.org/10.1111/pce.12848.

Graves, A., Mohamed, A., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: IEEE Int. Conf. Acoust. Speech Signal Process. pp. 6645–6649.

Guo, D., Juan, J., Chang, L., Zhang, J., Huang, D., 2017. Discrimination of plant root zone water status in greenhouse production based on phenotyping and machine learning techniques. Sci. Rep. 7, 8303. https://doi.org/10.1038/s41598-017-08235-z.

He, J.S., Kaiming, Zhang, Xiangyu, Ren, Shaoqing, 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: IEEE Int. Conf. Comput. Vis. https://doi.org/10.1109/ICCV.2015.123.

Ibayashi, H., Kaneda, Y., Imahara, J., Oishi, N., Kuroda, M., Mineno, H., 2016. A reliable wireless control system for tomato hydroponics. Sensors 16, 644. https://doi.org/10.3390/s16050644.

Jiang, Y., Li, C., Robertson, J.S., Sun, S., Xu, R., Paterson, A.H., 2018. GPhenoVision: a ground mobile system with multi-modal imaging for field-based high throughput

phenotyping of cotton. Sci. Rep. 8, 1213 doi:10.1038fmultimodaldee/s41598-018-19142-2.
Jolliet, O., Bailey, B.J., 1992. The effect of climate on tomato transpiration in greenhouses: measurements and models comparison. Agric. For. Meteorol. 58, 43–62.
Kaneda, Y., Mineno, H., 2016. Sliding window-based support vector regression for predicting micrometeorological data. Expert Syst. Appl. 59, 217–225. https://doi.org/10.1016/j.eswa.2016.04.012.
Kaneda, Y., Shibata, S., Mineno, H., 2017. Multi-modal sliding window-based support vector regression for predicting plant water stress. Knowledge-Based Syst. 134, 135–148. https://doi.org/10.1016/j.knosys.2017.07.028.
Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: single shot multibox detector. In: Eur. Conf. Comput. Vis. pp. 21–37.
Luong, M.T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. Conf. Empir. Methods Nat. Lang. Process.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proc. Fifth Berkeley Symp. Math. Stat. Probab. 1. pp. 281–297.
Meng, Z., Duan, A., Chen, D., Dassanayake, K.B., Wang, X., Liu, Z., Liu, H., Gao, S., 2017. Suitable indicators using stem diameter variation-derived indices to monitor the water status of greenhouse tomato plants. PLoS One 12, e0171423. https://doi.org/10.1371/journal.pone.0171423.
Moriyuki, S., Fukuda, H., 2016. High-throughput growth prediction for Lactuca sativa L. seedlings using chlorophyll fluorescence in a plant factory with artificial lighting. Front. Plant Sci. 7 1–8. https://doi.org/10.3389/fpls.2016.00394.
Nereu, A., 2003. Stomatal response to water vapor pressure deficit: an unsolved issue. Rev. Bras. Agrociência. 9, 317–322. https://doi.org/10.18539/cast.v9i4.649.
Ngiam, J., Khosla, A., Nam, M., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: Int. Conf. Mach. Learn. pp. 689–696.
Nitish Srivastava, R.S., Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Oishi, N., 2016. Non-disruptive evaluation of leaf area index using diffused light sensor for tomato cultivation. Shokubutsu Kankyo Kogaku. 28, 125–132. https://doi.org/10.2525/shita.28.125.
Patanè, C., Cosentino, S.L., 2010. Effects of soil water deficit on yield and quality of processing tomato under a Mediterranean climate. Agric. Water Manag. 97, 131–138. https://doi.org/10.1016/j.agwat.2009.08.021.
Rahimi, A., Recht, B., 2007. Random features for large scale kernel machines. Adv. Neural Inf. Process. Syst. 1–8 doi: 10.1.1.145.8736.
Sánchez-Molina, J.A., Rodríguez, F., Guzmán, J.L., Ramírez-Arias, J.A., 2015. Water content virtual sensor for tomatoes in coconut coir substrate for irrigation control design. Agric. Water Manag. 151, 114–125. https://doi.org/10.1016/j.agwat.2014.09.013.
Sano, M., Nakagawa, Y., Sugimoto, T., Shirakawa, T., Yamagishi, K., Sugihara, T., Ohaba, M., Shibusawa, S., 2015. Estimation of water stress of plant by vibration measurement of leaf using acoustic radiation force. Acoust. Sci. Technol. 36, 248–253. https://doi.org/10.1250/ast.36.248.
Scholkopf, B., Smola, A., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319. https://doi.org/10.1162/089976698300017467.
Sergey Ioffe, C.S., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. Folia Zool. 37, 448–456.
Singh, A.K., Ganapathysubramanian, B., Sarkar, S., Singh, A., 2018. Deep learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci. 23, 883–898. https://doi.org/10.1016/j.tplants.2018.07.004.
Takayama, K., Nishina, H., 2007. Early detection of water stress in tomato plants based on projected plant area. Environ. Control Biol. 45, 241–249. https://doi.org/10.2525/ecb.45.241.
Veličković, P., Karazija, L., Lane, N.D., Bhattacharya, S., Liberis, E., Liò, P., Chieh, A., Bellahsen, O., Vegreville, M., 2017. Cross-modal recurrent models for weight objective prediction from multimodal time-series data. Int. Conf. Pervasive Comput. Technol. Healthc. 178–186. https://doi.org/10.1145/3240925.3240937.
Wakamori, K., Mineno, H., 2019. Optical flow-based analysis of the relationships between leaf wilting and stem diameter variations in tomato plants. Plant Phenomics. https://doi.org/10.34133/2019/9136298.
Wang, X., Meng, Z., Chang, X., Deng, Z., Li, Y., Lv, M., 2017. Determination of a suitable indicator of tomato water content based on stem diameter variation. Sci. Hortic. (Amsterdam) 215, 142–148. https://doi.org/10.1016/j.scienta.2016.11.053.
Weinzaepfel, P., Harchaoui, Z., Schmid, C., Weinzaepfel, P., Harchaoui, Z., Schmid, C., Harchaoui, Z., Schmid, C., 2013. DeepFlow: large displacement optical flow with deep matching. In: IEEE Int. Conf. Comput. Vis. pp. 2818–2826.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Bengio, Y., 2015. Show attend and tell: neural image caption generation with visual attention. Int. Conf. Mach. Learn. 2048–2057.