scientific reports

Check for updates

OPEN

Dissection of hyperspectral reflectance to estimate nitrogen and chlorophyll contents in tea leaves based on machine learning algorithms

Hiroto Yamashita[1,2], Rei Sonobe[1]✉, Yuhei Hirono[3], Akio Morita[1] & Takashi Ikka[1]✉

Nondestructive techniques for estimating nitrogen (N) status are essential tools for optimizing N fertilization input and reducing the environmental impact of agricultural N management, especially in green tea cultivation, which is notably problematic. Previously, hyperspectral indices for chlorophyll (Chl) estimation, namely a green peak and red edge in the visible region, have been identified and used for N estimation because leaf N content closely related to Chl content in green leaves. Herein, datasets of N and Chl contents, and visible and near-infrared hyperspectral reflectance, derived from green leaves under various N nutrient conditions and albino yellow leaves were obtained. A regression model was then constructed using several machine learning algorithms and preprocessing techniques. Machine learning algorithms achieved high-performance models for N and Chl content, ensuring an accuracy threshold of 1.4 or 2.0 based on the ratio of performance to deviation values. Data-based sensitivity analysis through integration of the green and yellow leaves datasets identified clear differences in reflectance to estimate N and Chl contents, especially at 1325–1575 nm, suggesting an N content-specific region. These findings will enable the nondestructive estimation of leaf N content in tea plants and contribute advanced indices for nondestructive tracking of N status in crops.

Nitrogen (N) is the most demanded element for photosynthetic function and growth in plants. N fertilization has become a major yield-enhancing technique in modern crop production[1–3]. However, excessive N fertilization is known to be a source of water and air pollution. Groundwater contamination by nitrate–N (NO_3-N) from excess N fertilizer is a serious problem in many countries[4]. Furthermore, N fertilizers are important sources of nitrous oxide (N_2O), which is involved in destruction of the atmospheric ozone layer[5,6]. Furthermore, excess N fertilizer application increases management costs, even in modern large-scale agricultural production. Therefore, optimizing the amount of N fertilization by tracking N status is necessary for crop nutrient status and to reduce the environmental impact of agricultural management.

N is a structural element of chlorophyll (Chl), affecting leaf greenness and Chl accumulation[7–9]. The proportion of leaf N allocated to the chloroplast is approximately 75%[10,11]. The leaf N content and Chl content in plant green leaf are positively correlated. This has been reported for numerous plant species, and nondestructive and rapid N status estimation has been conducted using Chl meters on most major crops, including rice (*Oryza sativa* L.)[12,13], wheat *(Triticum aestivum* L.)[14], maize (*Zea mays* L.)[15], and others[16,17]. Nondestructive N status estimation in the leaf and canopy using hyperspectral sensing has also been applied to many plant species for nutritional diagnosis[18–25]. In most cases, the hyperspectral indices for Chl estimation, namely a green peak and red edge (500–800 nm) in the visible region, have been identified and used owing to their multicollinearity, with the leaf N content closely related to Chl content in plant green leaves, as mentioned above[12,22,25,26]. However, decreased Chl content can be caused by various factors, such as herbicide injury, that are not necessarily related to N deficiency[27]. Therefore, estimations of Chl and N content must be decoupled from remote sensing data to assess various stresses and pathogens[26].

[1]Faculty of Agriculture, Shizuoka University, Shizuoka, Japan. [2]United Graduate School of Agricultural Science, Gifu University, Gifu, Japan. [3]Division of Tea Research, Institute of Fruit Tree and Tea Science, National Agriculture and Food Research Organization (NARO), Shimada, Japan. ✉email: sonobe.rei@shizuoka.ac.jp; ikka.takashi@shizuoka.ac.jp

Machine learning techniques are powerful tools for estimating agricultural indices from hyperspectral remote sensing data[28]. Among the main advantages of machine learning algorithms is their ability to autonomously solve large nonlinear problems using datasets from multiple variables, and provide a powerful and flexible framework not only for data-driven decision making, but also for incorporating expert knowledge into the algorithms[29]. This methodology also shows potential for analyzing hyperspectral reflectance data with a large number of bands, working with not only variables such as derived spectral indices, but also all spectral information[30]. Previous spectral indices have depended on a small number of available spectral bands and, therefore, do not use all information conveyed by the spectral trace[29]. Machine learning techniques can assess the features that are more informative for high-accuracy prediction modelling[29,31].

Tea plants (*Camellia sinensis* L.) are mainly cultivated in Asia for the production of green, oolong, and black teas, which are among the most popular beverages worldwide. Tea is a leaf-harvested crop, and N is the most important nutrient for improving the yield and quality, such as free amino acid contents, of tea leaves[32]. Therefore, to meet these criteria, tea fields, especially in Japan, tend to receive higher rates of N fertilization, such as with ammonium sulfate, than other crops, sometimes exceeding 1000 kg N ha^{-1} $year^{-1}$[33]. Heavy N fertilization in tea fields often causes problems such as increased NO_3-N levels in surrounding water systems[34] and high N_2O emission levels[32,33,35,36]. N_2O emission rates in tea fields are much higher than those in other upland fields and paddy fields[33,37]. Reducing NO_3-N leaching and N_2O emissions from tea fields would be a significant step towards decreasing the environmental impact of agricultural N management. In addition, tea plants with characteristic leaf colors, such as yellow (or white)[38–42] and purple[43], have been studied extensively. Mutant (bud-sport) branches with albino yellow leaves due to a lack of chlorophyll are often found in tea gardens, with albino-induced tea leaves generally containing higher amino acid contents than conventional green tea leaves[38–42]. Therefore, the possibility that albino tea leaves might have an N status that does not reflect the Chl status was considered.

This study mainly aimed to assess differences in hyperspectral reflectance to estimate N and Chl contents and enable nondestructive estimation of leaf N contents in tea plants, which require large amounts of N nutrition. Initially, the dataset of N and Chl contents and visible and near-infrared hyperspectral reflectance with variations derived from green leaves with various N nutrient conditions and albino yellow leaves was obtained. A regression model was then constructed using several machine learning algorithms and preprocessing techniques. Data-based sensitivity analysis based on high-performance models using integrating datasets from green and albino yellow leaves identified clear differences in reflectance to estimate N and Chl contents.

Results

Data distribution of nitrogen and chlorophyll contents. To obtain the dataset of N and Chl contents with variations, green leaves (GL) with various N nutrient conditions from hydroponic and shading tests (Exp. 1 to Exp. 3) and albino yellow leaves (YL; Exp. 4) were tested (Fig. 1). In all experiments, the N and Chl contents were in the range of 164.8–732.5 and 0.61–118.3 mg cm^{-2}, respectively (Figs. 2A,B). The dataset for subsequent modelling was divided into DatasetA (n = 181), comprising only the GL data (Exp. 1 to Exp. 3), and DatasetB (n = 227), comprising the GL and YL data (Exp. 1 to Exp. 4) (Fig. 1). A significant positive correlation was observed between N and Chl content in DatasetA, but not in DatasetB (Fig. 2C,D).

Regression model performance. Original reflectance (OR) data at 1-nm steps across the entire wavelength domain from 400 to 2500 nm was obtained from the leaf samples. Five preprocessing methods, namely first derivative reflectance (FDR), continuum-removed (CR), standard normal variate (SNV), multiplicative scatter correction (MSC), and de-trending (DT) (Supplementary Fig. S1), were applied to the OR data to compare regression model performance. The following five regression methods were performed: Random Forest (RF), Support Vector Machine (SVM), Cubist, Stochastic Gradient Boosting (SGB), and Kernel-based Extreme Learning Machine (KELM). Model performance was evaluated using the ratio of performance to deviation (RPD) values and robustness over 100 repetitions. For N content modelling, Cubist and KELM models indicated high performance and robustness when OR, DT, and SNV were applied as hyperspectral data both in DatasetA and DatasetB, as most RPD values over 100 repetitions were greater than 1.4, representing a fairly acceptable prediction level (Fig. 3A). For Chl content modelling, the same pattern of results as above was observed (Fig. 3B), and the model performance of DatasetB was higher than that of DatasetA, with most RPD values over 100 repetitions being above 2.0, representing an accurate prediction level (Fig. 3B). These results were also supported by the coefficient of determination (R^2) and root mean square error (RMSE) values as model performance indices (Supplementary Figs. S2 and S3).

Detection of important hyperparameters by data-based sensitivity analysis. Data-based sensitivity analysis (DSA) was performed to detect important hyperspectral parameters in models to estimate N and Chl contents. The results of DSA in models applying OR hyperspectral parameters are shown in Fig. 4A. For Chl content, both DatasetA and DatasetB showed peaks of importance at 525–725 and 1875–1925 nm in all models (Fig. 4A). For N content, peaks of importance were observed at 675–725 and 1325–1575 nm, with the latter peak enhanced in DatasetB compared with DatasetA, especially for RF, Cubist, and SGB (Fig. 4A). The DSA results of models applying DT hyperspectral parameters are shown in Fig. 4B. For Chl content, both DatasetA and DatasetB showed peaks of importance at 475–725 nm in all models, and 1975–2075 nm for RF, Cubist, and SGB (Fig. 4B). For N content, both DatasetA and DatasetB showed peaks of importance at 2125–2275 nm, especially for RF, Cubist, and SGB (Fig. 4B).

Typical OR and DT spectra, and their important regions based on DSA of GL and YL with different N statuses, are shown in Fig. 5. The OR spectra of GL showed a clear response to N and Chl contents at 500–800 nm in the green peak and red edge region (Fig. 5A). The OR spectra of YL showed high reflectance at around 550 nm in the

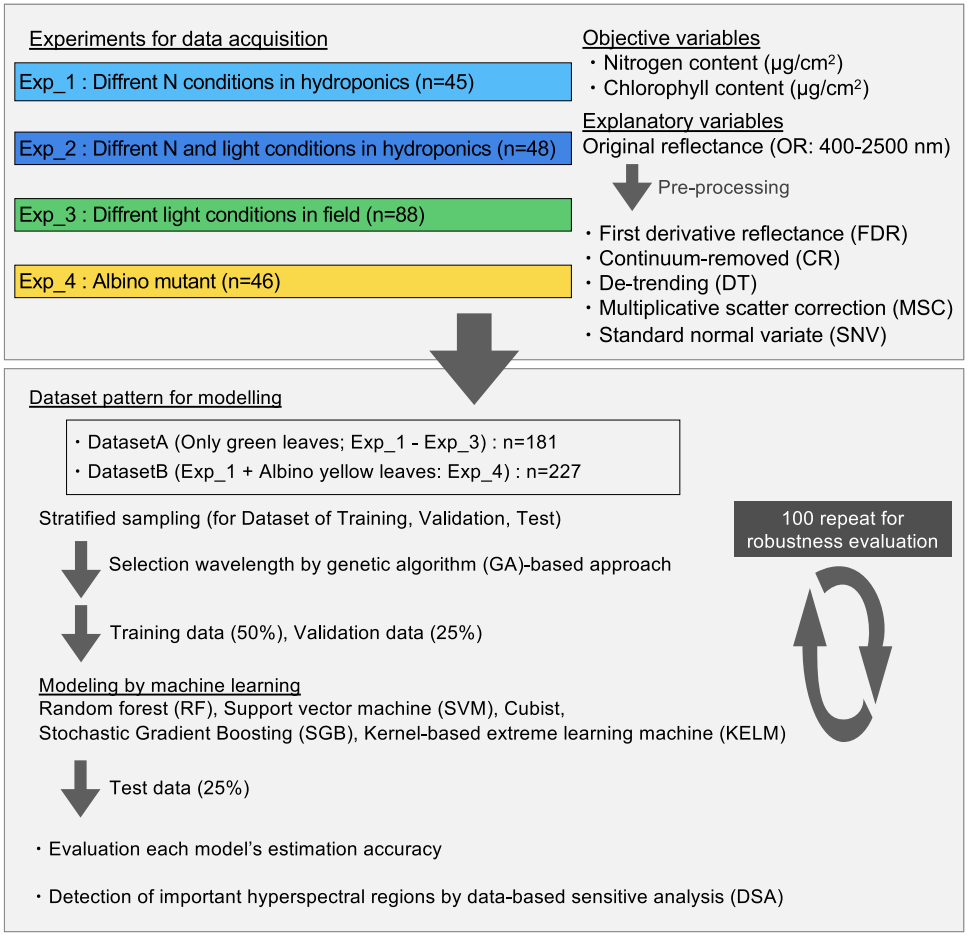

Figure 1. Experiment and modelling designs in this study.

green peak region, independent of N status (Fig. 5B). A shift dependent on N status toward shorter wavelengths in the 700–780 nm region (blue shift) was observed in GL, but not YL (Fig. 5A,B). In the OR spectra of GL and YL at 1300–1700 nm, a clear response to N content was observed (Fig. 5A,B). Furthermore, DT spectra of GL and YL showed a clear response to N and Chl contents at 500–800 nm in the green peak and red edge region (Fig. 5C,D). At 2100–2300 nm in the DT spectra of GL and YL, a clear response to N content was observed, especially for GL (Fig. 5C,D).

Discussion

N fertilization, which is directly related to yield and quality, is indispensable in modern agriculture to achieve stable food production[1–3,32]. However, improving the efficiency of N nutrition is necessary for both agricultural management and global environmental conservation[4,6,33]. In green tea cultivation, these environmental impacts in agricultural N management have been notably problematic[33–35]. Therefore, the impact of N management can be maximized in tea cultivation. Nondestructive estimation techniques, such as hyperspectral reflectance, are effective tools for tracking crop status[18–25]. In previous studies, hyperspectral techniques have been used to estimate N status, such as for Chl in leaves[12,22,25]. Therefore, this study aimed to determine differences in hyperspectral reflectance to accurately estimate N and Chl contents in GL and YL using machine learning algorithms, and enable the nondestructive estimation of leaf N content in tea plants.

Datasets of N and Chl contents were obtained from GL under various N nutrient conditions in hydroponic and shading tests (Exp. 1 to Exp. 3) and YL (Exp. 4) (Fig. 2A,B), and DatasetA (n = 181; only GL) and DatasetB (n = 227; GL and YL) were constructed for subsequent modelling (Fig. 1). A reflectance in YL clearly showed a different appearance to that in GL (Fig. 5). The reflectance of the green peak region in YL was high and broad, merging together with the start point of the red edge region (Fig. 5B), as a characteristic of albino leaves reported by Baldini et al. (1997)[44]. In DatasetB, the positive correlation between N and Chl contents was not observed (Fig. 2D). This result suggested that modelling using DatasetB did not suffer from multicollinearity between the N and Chl contents, allowing the selection of explanatory variables specific to each.

In our modelling, machine learning methods, especially Cubist and KELM, with OR data showed high-performance and robustness both for N and Chl contents, ensuring accuracy thresholds based on the RPD values (Fig. 3). Cubist can generate so-called committee models that consist of a set of consecutive rule-based models to

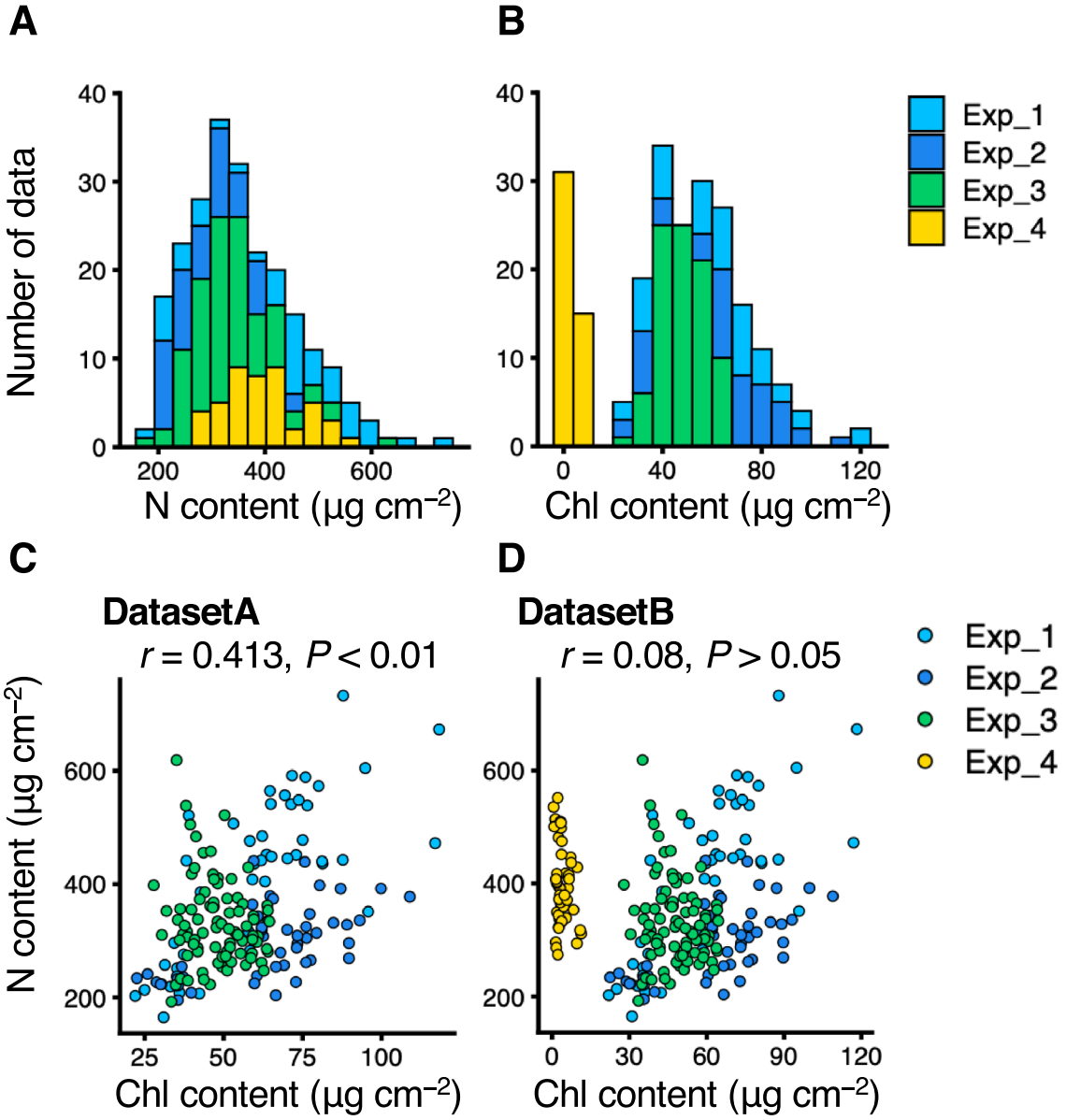

Figure 2. Data distributions of nitrogen (N) and chlorophyll (Chl) contents in tea leaves. Histograms of (**A**) N and (**B**) Chl contents in all experiments. Correlation plots of N and Chl contents in (**C**) DatasetA and (**D**) DatasetB. Figures were visualized by using the R package "ggplot2" ver. 3.3.2. Statistical results for correlation were shown on the figure.

correct the predictions of previous member models[44]. Furthermore, Cubist is computationally efficient[44] and well-suited to big data analytics[45]. Cubist has shown potential as an efficient model method for various agricultural targets using reflectance data, such as soil physical properties[46], vegetation stress[47], leaf area index[48], and crop yields[49]. Although KELM and SVM are both kernel-based algorithms, SVM was clearly inferior to KELM in this modelling. Furthermore, the variance of the kernel function parameters of KELM was apparently smaller than that of SVM (Supplementary Table S1). Generally, the ranges of the kernel bandwidth of SVM (s) were wider than those of the kernel parameter of KELM (Kp), implying that the Bayesian optimization sometimes resulted in local solutions for optimizing SVM hyperparameters. Indeed, quite low RPD values (less than 1.0) for SVM were confirmed using all preprocessing techniques (Fig. 3). The incorrect selection of hyperparameters related to kernel function has been reported to reduce the estimation accuracy[50], and local solutions of the Bayesian optimization led to worse estimation accuracies for SVM. Furthermore, KELM has fewer optimization constraints[51], which is advantageous in regression applications[52]. The results of this and previous studies indicate that the Cubist and KELM methods are suitable for modelling using reflectance data in plants. Some preprocessing techniques, especially DT, also showed similar results to those using OR data. Barnes et al. (1989)[53] reported that DT accounted for the variation in baseline shift and curvilinearity from the reflectance spectra of powdered or densely packed samples using second-degree polynomial regression. DT was an effective preprocessing technique for estimating the N and Chl contents, and the potential of DT for targeting plants was also confirmed in this study. Comparing

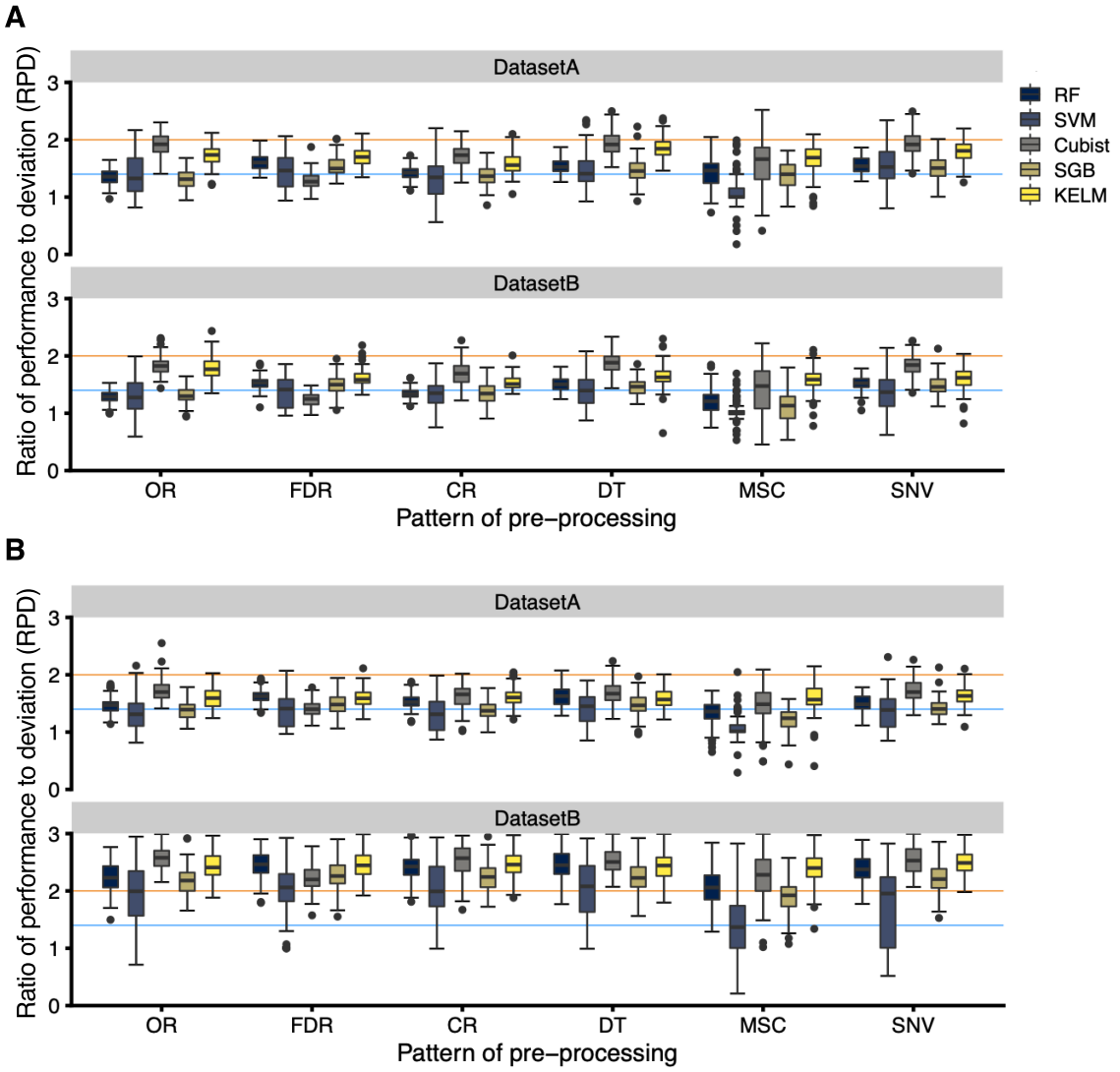

Figure 3. Model performance and robustness for each pre-processing of reflectance. Modelling with (**A**) nitrogen and (**B**) chlorophyll contents as the objective variables was performed using the explanatory variables of DatasetA and DatasetB. The ratio of performance to deviation (RPD) was applied to evaluate the accuracy of each model. A stratified sampling approach for modelling was repeated 100 times to obtain robust results. Figures are plots of the RPD values in each repeat. Orange and blue lines indicate RPD values of 1.4 and 2.0, respectively, as accuracy thresholds. Figures were visualized by using the R package "ggplot2" ver. 3.3.2.

DatasetA and DatasetB, the model performance for N content was similar (Fig. 3A). However, for Chl content, the model performance of DatasetB was higher than that of DatasetA, with most RPD values above 2.0 for 100 repetitions, representing an accurate prediction level (Fig. 3B). These results suggested that the inclusion of YL, which had the lowest Chl content in the dataset, functioned well at estimating the Chl content, but not the N content. Furthermore, these results indicated that the approach using hyperspectral reflectance and machine learning algorithm allowed not only the Chl content, as previously reported by Sonobe et al. (2020)[54], but also the N content to be estimated nondestructively in tea leaves. For green tea production, especially in Japan, N fertilizer is mainly applied in the autumn, winter, and at the bud-opening stage before the first crops in spring, to improve the quality[55]. This is because N absorbed into the source organs, such as mature leaves, contributes to new shoots for the first crops[55]. Therefore, modelling was also conducted to estimate the N content in mature leaves, using only GL from this experiment design (Supplementary Fig. S4). This modelling approach provided the N content of mature leaves with high performance and robustness (Supplementary Figs. S5 and S6), but was limited by the possibility of multicollinearity with the Chl content (Supplementary Fig. S4C).

The important hyperspectral parameters in models to estimate N and Chl content were detected using DSA. For the Chl content in both DatasetA and DatasetB, peaks of importance were observed at 525–725 and 1875–1925 nm in all models (Fig. 4A). Chlorophyll absorbs energy strongly in the ultraviolet (200–400 nm), blue (400–500 nm), and red (650–690 nm) regions, and shows weak reflectance and transmittance[56]. Therefore, most indices for Chl content in plant leaves and canopies were selected wavelengths, as a few narrow bands, ranging from 400 to 860 nm[57–59] or the red edge (680–750 nm)[60,61]. Previously, our analysis of Chl content in

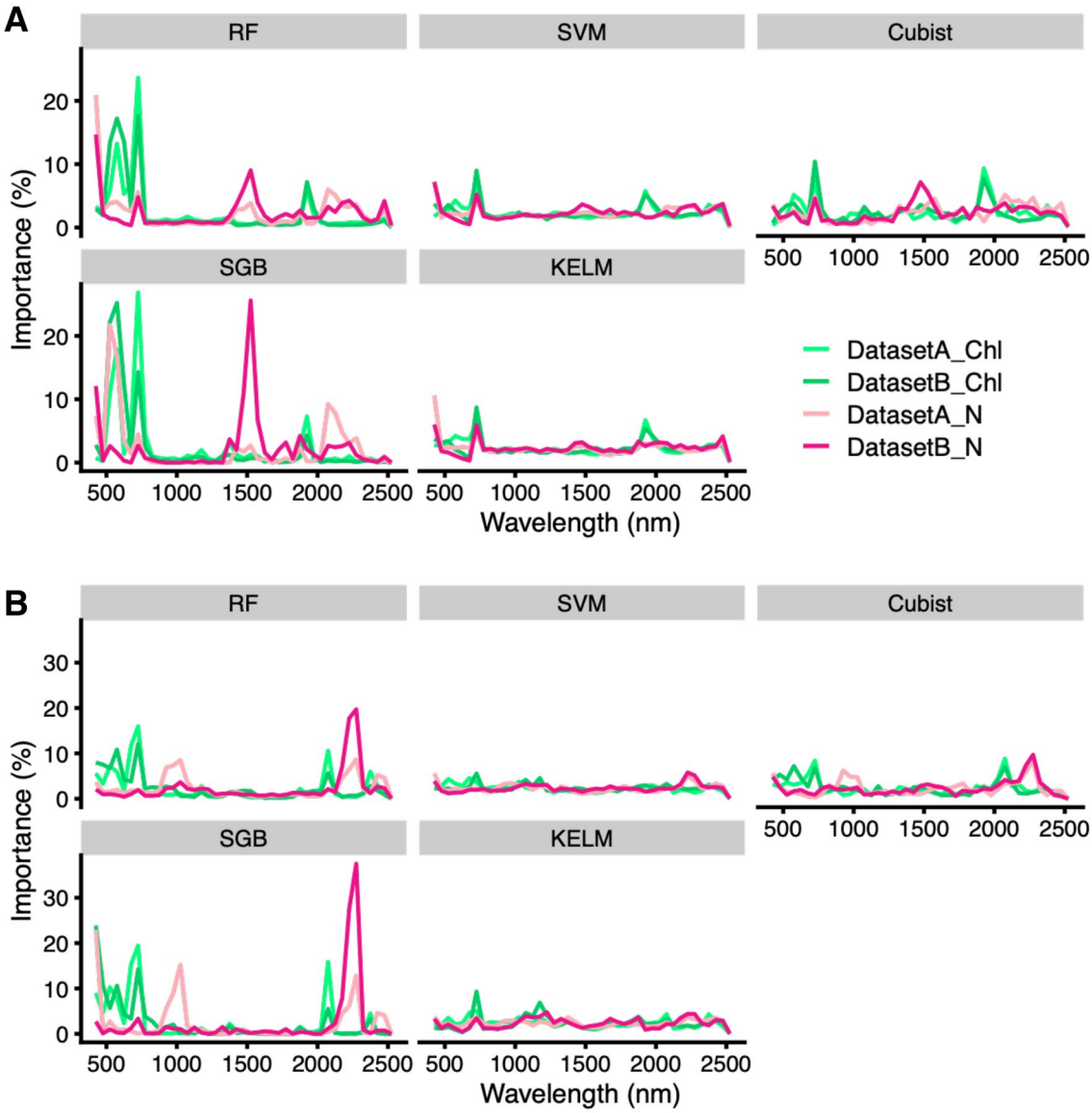

Figure 4. Detection of important hyperspectral parameter regions as model variables by data-based sensitive analysis (DSA). DSA results when (**A**) original reflectance or (**B**) de-trending (DT) were visualized as the explanatory variables in each model. Figures were visualized by using the R package "ggplot2" ver. 3.3.2.

tea leaves also indicated that the values of highest importance were confirmed at 700–750 nm (the red edge region) when OR and FDR were used[54]. For N content, the peaks of importance were observed at 675–725 and 1325–1575 nm, with the latter peak in the near-infrared region enhanced in DatasetB compared with DatasetA, especially for RF, Cubist, and SGB (Fig. 4A). These results suggested that this near-infrared reflectance region, at 1325–1575 nm, was N content-specific, independent of the Chl content. Previous studies have shown that plant water status affects shortwave-infrared and near-infrared domains, showing that these domains are sensitive to water absorption[62–65]. No correlation was observed between the leaf water and N contents in the dataset of this study (Supplementary Fig. S7), which supported that the near-infrared reflectance region of 1325–1575 nm detected in this analysis was an N content-specific region. Furthermore, DSA with some preprocessing techniques also detected the different hyperspectral regions to estimate N and Chl content (Supplementary Fig. S8). Therefore, preprocessing techniques might be a tool for the specific estimation of N and Chl contents using hyperspectral data.

The results of this study suggest that hyperspectral reflectance data and machine learning techniques show good potential for estimating leaf N and Chl contents in tea plants. Remote sensing techniques using unmanned aerial vehicles will also enable the high-throughput estimation of N and Chl statuses in canopy-scale tea gardens. Furthermore, the inclusion of YL, which had the lowest Chl content in the dataset, contributed to the detection of N-content-specific hyperspectral regions, independent of Chl content, in the near-infrared reflectance region. These findings will contribute advanced indices for the nondestructive tracking of crop N status. In the future, these techniques could be applied to improve irregularities in fertilizer and the real-time diagnostics of physiological status changes in large farms.

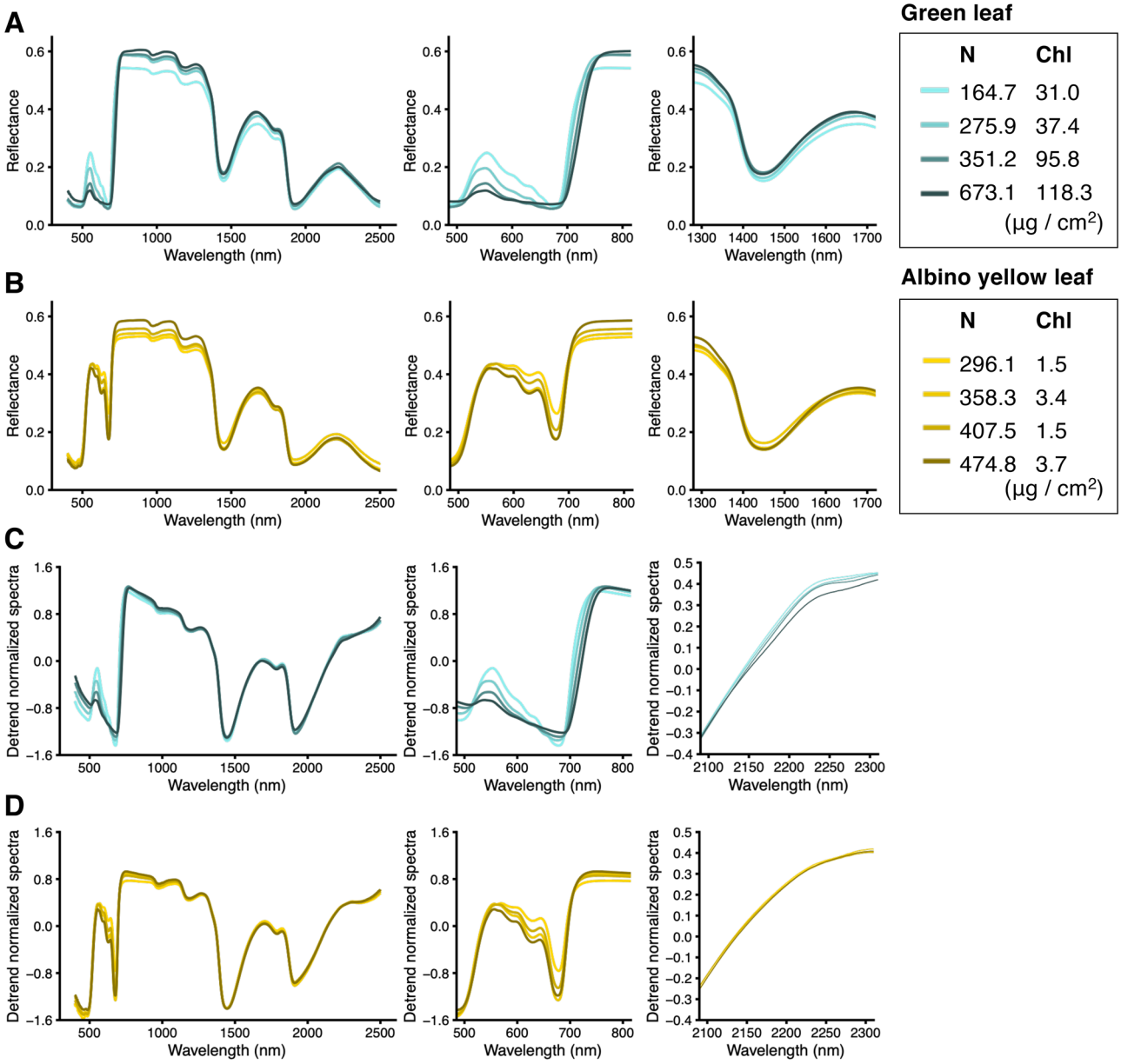

Figure 5. Typical original reflectance and de-trending spectra of green and albino yellow leaves with different nitrogen (N) statuses. Reflectance spectra of (**A**) green and (**B**) albino yellow leaves. De-trend normalized spectra of (**C**) green and (**D**) albino yellow leaves. N and chlorophyll contents in each spectrum are shown with different colors to the right of the figure. Each figure consists of all measured wavelengths (left), green peak and red edge regions (500–800 nm; middle), and important ranges identified by DSA (right). Figures were visualized by using the R package "ggplot2" ver. 3.3.2.

Methods

Plant materials for datasets. To obtain the dataset of N and Chl contents with variations, the following four experiments (Exp. 1 to Exp. 4) were conducted.

Exp. 1 and Exp. 2 were conducted using hydroponic nutrient tests. The hydroponic culture of tea plants was conducted under sunlight conditions in the greenhouse at Shizuoka University (Shizuoka, Japan) using a slight modification of the method described by Konishi et al. (1985)[66]. One-year-old rooted tea cuttings of cv. 'Yabukita', a leading Japanese green tea cultivar, were transplanted with three individuals in Wagner pots (1/5000 a) containing tap water (3 L) adjusted to pH 4.2 using H_2SO_4, and continuously aerated. After 1 week, standard nutrient solutions containing N 40 mg L^{-1} (1 × N), 10 mg L^{-1} NO_3-N, and 30 mg L^{-1} NH_4-N at pH 4.2[66] were supplied stepwise for 1 week each at 1/5, 1/2, and 1/1 strength to adapt the plants to the hydroponic system. Subsequently, the following experiments were performed.

In Exp. 1, plants were transferred to nutrient solutions adjusted to the following six N levels: –N, 0.01 × N, 0.1 × N, 1 × N, 2 × N, 4 × N. Each experiment was conducted using three to five biological replicates. Solutions were

renewed every week. After approximately 6 months, new leaves, the second leaf from bottom in four-leaf-stage new shoots, and mature leaves were plucked from two to four spots in one replicate, and then leaf reflectance was measured. After measuring reflectance, the leaves were scanned at 300 dpi (CanonScan LiDE 210 JP scanner, Canon Inc, Tokyo, Japan) to calculate the leaf area, weighed in fresh condition, and then reweighed in dry condition after freeze-drying. Dry samples were grounded into a fine powder and stored at room temperature in a desiccator prior to nitrogen and chlorophyll measurements.

In Exp. 2, when new buds were developed, plants were transferred to nutrient solutions adjusted to the following four N levels: –N, 0.1 × N, 1 × N, 4 × N. The plants at each N level were grown under normal (control) and low-light (shading) conditions. Low-light conditions were set up with 85% coverage by synthetic black cloth (85P, Dio Chemicals, Tokyo, Japan). Each experiment was conducted using three biological replicates. The solutions were renewed every week. After 23 days, new leaves, the second leaf from bottom in four-leaf-stage new shoots, and mature leaves were plucked from one to two spots in one replicate, respectively, and measured as mentioned above.

In Exp. 3, new shoots were plucked from mature tea ridges of cv. Yabukita at Shizuoka University (Shizuoka, Japan) under open and low-light conditions with 85% shading. After developing the two leaves of almost new shoots in the first crop season (around the end of April), shaded cultivation was conducted. These tea ridges were shaded by 85% coverage with black cloth (85P, Dio Chemicals, Tokyo, Japan) for 20 days of the first crop season of spring in Japan. Shading materials directly covered the tea canopies, as described by Sano et al. (2018)[67]. After shade cultivation, new shoots and mature leaves were plucked from approximately 15 spots in open and shaded tea ridges, and measured as described above. New leaves were then plucked from the third, fourth, and fifth leaves in five-leaf-stage new shoots. Tea ridges were managed by conventional methods in Japan and cultivated under the same soil and environment conditions. Nitrogen fertilizer was applied at 400 kg-N ha^{-1} $year^{-1}$.

In Exp. 4, new shoots were plucked from 7-year-old rooted tea cuttings of cv. 'Koganemidori', previously known as "Morokozawa"[38], which had been bred from the natural albino mutant, and hydroponically cultivated under sunlight condition in the greenhouse at Shizuoka University (Shizuoka, Japan), as described above. This albino cultivar "Koganemidori" has a yellow color leaf and very low chlorophyll content (Supplementary Fig. S9)[38]. In the first crop season, new shoots were plucked from approximately 20 spots in one replicate and measured as described above. New leaves were then plucked from the second, third, and fourth leaves in four-leaf-stage new shoots.

Reflectance measurements. An ASD FieldSpec4 unit (Analytical Spectral Devices, USA) was used to obtain reflectance data from leaf clippings (φ, 10 mm). This device contained three detectors, namely, visible and near-infrared (VNIR), short wave infrared (SWIR) 1, and SWIR 2. The splice correction function of ViewSpec Pro Software (Analytical Spectral Devices) was applied to correct differences in the spectral drifts (at 1000 and 1800 nm) caused by inherent variation in detector sensitivities. Finally, reflectance data at 1-nm steps were obtained across the entire wavelength domain from 400 to 2500 nm. Five preprocessing methods were also tested based on their success in previous studies, namely FDR[54,68], CR[69], SNV[70], MSC[71], and DT[53].

Nitrogen and chlorophyll measurement. Total N was measured by dry combustion using an NC analyzer (Vario MAX cube, Elementar, Hanau, Germany). Aspartic acid was used as a standard in total N analysis.

Chlorophylls a and b were extracted from finely ground powder (5 mg) of freeze-dried leaf samples using *N*,*N*′-dimethylformamide (5 mL). After incubation for 24 h at 4 °C under dark conditions to allow complete decolorization, the samples were centrifuged at 2000×*g* for 30 min, and the absorbance of the supernatant was measured at 663.8 and 646.8 nm using a spectrophotometer (UV-1900, Shimadzu, Kyoto, Japan). Chlorophyll a and b contents were calculated using the equation of Porra et al. (1989)[72].

Regression models by machine learning algorithms. The flow of regression model generation was conducted as described by Sonobe et al. (2020)[54], with a slight modification. For modelling, all measurements were divided into three groups, a training dataset (50%), a validation dataset (25%), and a test data dataset (25%), using a stratified sampling approach[73]. To ensure robust results, this approach was repeated 100 times before preprocessing the original reflectance and generating regression models based on machine learning algorithms.

When applying machine learning algorithms, a genetic algorithm (GA)-based approach[74] is applied to select wavelengths using R ver. 3.6.3. that are effective for removing noninformative variables to obtain better and simpler prediction models. Regression models were then created from the selected bands using the following five methods: RF, SVM, Cubist, SGB, and KELM. To optimize the hyperparameters of these machine learning algorithms, Bayesian optimization was applied with the Gaussian process[75,76] using R package "rBayesianOptimization" ver. 1.1.0. The information about the hyperparameters of these machine learning algorithms were shown in Supplementary Table S2.

RF was performed and optimized with R package "randomForestSRC" ver. 2.9.3 using the following five hyperparameters: number of trees; number of variables used to split nodes; minimum number of unique cases in a terminal node; maximum depth to which a tree should be grown; and number of random splittings. SVM was performed with the Gaussian radial basis function kernel and optimized with R package "e1071" using the following two hyperparameters: Regularization parameter (C) and kernel bandwidth (s). These were considered as user-defined hyperparameters using the R package "e1071" ver. 1.5-8. Cubist was performed and optimized with R package "Cubist" ver. 0.2.3 using the following two hyperparameters: Number of committee models (boosting iterations); number of neighbors. SGB was performed and optimized with R package "gbm" ver. 2.1.5 using the following four hyperparameters: Number of iterations and number of basis functions in the additive expansion; maximum depth of each tree; learning rate; and minimum number of observations in the terminal

nodes of the trees. KELM was performed and optimized with MATLAB and Statistics Toolbox Release 2016a (The MathWorks, Inc., Natick, MA, USA; source code downloaded from https://www.ntu.edu.sg/home/egbhuang/) using the following two hyperparameters: Regulation coefficient (Cr) and kernel parameter (Kp).

The estimation accuracy of each method was evaluated based on the coefficient of determination (R^2), root-mean-square error (RMSE), and ratio of performance to deviation (RPD). The quality of the prediction model was interpreted according to three classes of RPD[77–79], as follows: RPD > 2 represents accurate prediction by the model, RPD of 1.4–2 represents fairly acceptable prediction, and RPD < 1.4 represent poor performance by the prediction model.

Sensitivity analysis. Data-based sensitivity analysis (DSA) is proposed as a visualization approach for extracting human-understandable knowledge from supervised learning black box data mining models[80,81]. A black box analysis of the fitted models was performed with their machine learning algorithms by querying the fitted models with sensitivity samples and recording their responses, as previously described by Sonobe et al. (2020)[54].

Received: 30 June 2020; Accepted: 21 September 2020
Published online: 15 October 2020

References

1. Hucklesby, D. P., Brown, C. M., Howell, S. & Hageman, R. H. Late spring applications of nitrogen for efficient utilization and enhanced production of grain and grain protein of wheat 1. *Agron. J.* **63**, 274–276 (1971).
2. Salvagiotti, F. *et al.* Nitrogen uptake, fixation and response to fertilizer N in soybeans: a review. *Field Crops Res.* **108**, 1–13 (2008).
3. Spiertz, J. H. J. Nitrogen, Sustainable Agriculture and Food Security: A Review. In *Sustainable Agriculture* (eds Lichtfouse, E. *et al.*) (Springer Netherlands, Amsterdam, 2009). https://doi.org/10.1007/978-90-481-2666-8_39.
4. Inoue, Y., Dabrowska-Zierinska, K. & Qi, J. Synoptic assessment of environmental impact of agricultural management: a case study on nitrogen fertiliser impact on groundwater quality, using a fine-scale geoinformation system. *Int. J. Environ. Stud.* **69**, 443–460 (2012).
5. Crutzen, P. J. The influence of nitrogen oxides on the atmospheric ozone content. *Q. J. R. Meteorol. Soc.* **96**, 320–325 (1970).
6. Ishijima, K. *et al.* Temporal variations of the atmospheric nitrous oxide concentration and its $\delta^{15}N$ and $\delta^{18}O$ for the latter half of the 20th century reconstructed from firn air analyses. *J. Geophys. Res.* **112**, 1031 (2007).
7. Takebe, M. & Yoneyama, T. Measurement of leaf color scores and its implication to nitrogen nutrition of rice plants. *JARQ* **23**, 86–93 (1989).
8. Mahlangu, R. I. S., Maboko, M. M., Sivakumar, D., Soundy, P. & Jifon, J. Lettuce (*Lactuca sativa* L.) growth, yield and quality response to nitrogen fertilization in a non-circulating hydroponic system. *J. Plant Nutr.* **39**, 1766–1775 (2016).
9. Dehnavard, S., Souri, M. K. & Mardanlu, S. Tomato growth responses to foliar application of ammonium sulfate in hydroponic culture. *J. Plant Nutr.* **40**, 315–323 (2017).
10. Hak, R., Rinderle-Zimmer, U., Lichtenthaler, H. K. & Natr, L. Chlorophyll a fluorescence signatures of nitrogen deficient barley leaves. *Photosynthetica* **28**, 151–159 (1993).
11. Kutík, J., Nátr, L., Demmers-Derks, H. H. & Lawlor, D. W. Chloroplast ultrastructure of sugar beet (*Beta vulgaris* L.) cultivated in normal and elevated CO_2 concentrations with two contrasted nitrogen supplies. *J. Exp. Bot.* **46**, 1797–1802 (1995).
12. Peng, S., García, F. V., Laza, R. C. & Cassman, K. G. Adjustment for specific leaf weight improves chlorophyll meter's estimate of rice leaf nitrogen concentration. *Agron. J.* **85**, 987–990 (1993).
13. Ntamatungiro, S., Norman, R. J., McNew, R. W. & Wells, B. R. Comparison of plant measurements for estimating nitrogen accumulation and grain yield by flooded rice. *Agron. J.* **91**, 676–685 (1999).
14. Reeves, D. W., Mask, P. L., Wood, C. W. & Delaney, D. P. Determination of wheat nitrogen status with a hand-held chlorophyll meter: influence of management practices. *J. Plant Nutr.* **16**, 781–796 (1993).
15. Bullock, D. G. & Anderson, D. S. Evaluation of the Minolta SPAD-502 chlorophyll meter for nitrogen management in corn. *J. Plant Nutr.* **21**, 741–755 (1998).
16. Feibo, W., Lianghuan, W. & Fuhua, X. Chlorophyll meter to predict nitrogen sidedress requirements for short-season cotton (*Gossypium hirsutum* L.). *Field Crops Res.* **56**, 309–314 (1998).
17. Nageswara Rao, R. C., Talwar, H. S. & Wright, G. C. Rapid assessment of specific leaf area and leaf nitrogen in peanut (*Arachis hypogaea* L.) using a chlorophyll meter. *J. Agron. Crop Sci.* **186**, 175–182 (2001).
18. Yoder, B. J. & Pettigrew-Crosby, R. E. Predicting nitrogen and chlorophyll content and concentrations from reflectance spectra (400–2500 nm) at leaf and canopy scales. *Remote Sens. Environ.* **53**, 199–211 (1995).
19. Takahashi, W., Vu, N.-C., Kawaguchi, S., Minamiyama, M. & Ninomiya, S. Statistical models for prediction of dry weight and nitrogen accumulation based on visible and near-infrared hyper-spectral reflectance of rice canopies. *Plant Prod. Sci.* **3**, 377–386 (2000).
20. Johnson, L. F. Nitrogen influence on fresh-leaf NIR spectra. *Remote Sens. Environ.* **78**, 314–320 (2001).
21. Kokaly, R. F. Investigating a physical basis for spectroscopic estimates of leaf nitrogen concentration. *Remote Sens. Environ.* **75**, 153–161 (2001).
22. Lamb, D. W. *et al.* Estimating leaf nitrogen concentration in ryegrass (*Lolium* spp.) pasture using the chlorophyll red-edge: theoretical modelling and experimental observations. *Int. J. Remote Sens.* **23**, 3619–3648 (2002).
23. Hansen, P. M. & Schjoerring, J. K. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sens. Environ.* **86**, 542–553 (2003).
24. Feng, W., Yao, X., Zhu, Y., Tian, Y. C. & Cao, W. X. Monitoring leaf nitrogen status with hyperspectral reflectance in wheat. *Eur. J. Agron.* **28**, 394–404 (2008).
25. Inoue, Y., Sakaiya, E., Zhu, Y. & Takahashi, W. Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements. *Remote Sens. Environ.* **126**, 210–221 (2012).
26. Berger, K. *et al.* Crop nitrogen monitoring: recent progress and principal developments in the context of imaging spectroscopy missions. *Remote Sens. Environ.* **242**, 111758 (2020).
27. Kimura, E., Bell, J., Trostle, C., Neely, C. & Drake, D. Potential causes of yellowing during the tillering stage of wheat in Texas. *Texas A&M AgriLife Extension Service* **4**, 1–5 (2016).
28. Behmann, J., Mahlein, A.-K., Rumpf, T., Römer, C. & Plümer, L. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precis. Agric.* **16**, 239–260 (2015).

29. Chlingaryan, A., Sukkarieh, S. & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput. Electron. Agric.* **151**, 61–69 (2018).
30. Van Wittenberghe, S. *et al.* Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset. *J. Photochem. Photobiol. B* **134**, 37–48 (2014).
31. Panda, S. S., Ames, D. P. & Panigrahi, S. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sens.* **2**, 673–696 (2010).
32. Tokuda, S. I. & Hayatsu, M. Nitrous oxide flux from a tea field amended with a large amount of nitrogen fertilizer and soil environmental factors controlling the flux. *Soil Sci. Plant Nutr.* **50**, 365–374 (2004).
33. Akiyama, H., Yan, X. & Yagi, K. Estimations of emission factors for fertilizer-induced direct N_2O emissions from agricultural soils in Japan: Summary of available data. *Soil Sci. Plant Nutr.* **52**, 774–787 (2006).
34. Hirono, Y., Watanabe, I. & Nonaka, K. Trends in water quality around an intensive tea-growing area in Shizuoka, Japan. *Soil Sci. Plant Nutr.* **55**, 783–792 (2009).
35. Hirono, Y. & Nonaka, K. Nitrous oxide emissions from green tea fields in Japan: contribution of emissions from soil between rows and soil under the canopy of tea plants. *Soil Sci. Plant Nutr.* **58**, 384–392 (2012).
36. Hirono, Y. & Nonaka, K. Effects of application of lime nitrogen and dicyandiamide on nitrous oxide emissions from green tea fields. *Soil Sci. Plant Nutr.* **60**, 276–285 (2014).
37. Jumadi, O., Hala, Y. & Inubushi, K. Production and emission of nitrous oxide and responsible microorganisms in upland acid soil in Indonesia. *Soil Sci. Plant Nutr.* **51**, 693–696 (2005).
38. Morita, A. *et al.* Chemical composition of new shoots in the first crop season of 'white leaf tea' cultivated in Japan. *Tea Res. J.* **111**, 63–72 (2011).
39. Du, Y. Y. *et al.* A study on the chemical composition of albino tea cultivars. *J. Hortic. Sci. Biotechnol.* **0316**, 9–13 (2017).
40. Lu, M. *et al.* Significantly increased amino acid accumulation in a novel albino branch of the tea plant (*Camellia sinensis*). *Planta* **249**, 363–376. https://doi.org/10.1007/s00425-018-3007-6 (2018).
41. Cheng, S. *et al.* Differential accumulation of specialized metabolite L-theanine in green and albino-induced yellow tea (*Camellia sinensis*) leaves. *Food Chem.* **276**, 93–100 (2018).
42. Ma, Q. *et al.* Transcriptomic analyses identify albino-associated genes of a novel albino tea germplasm 'Huabai 1'. *Hortic. Res.* **5**, 54 (2018).
43. Saito, T. *et al.* Anthocyanins from New Red Leaf Tea 'Sunrouge'. *J. Agric. Food Chem.* **59**, 4779–4782. https://doi.org/10.1021/jf200250g (2011).
44. Walton, J. T. Subpixel urban land cover estimation. *Photogram. Eng. Remote Sens.* **74**, 1213–1222 (2008).
45. Li, S. *et al.* Geospatial big data handling theory and methods: a review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **115**, 119–133 (2016).
46. Lacoste, M. *et al.* High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma* **213**, 296–311 (2014).
47. Brown, J. F., Wardlow, B. D., Tadesse, T., Hayes, M. J. & Reed, B. C. The Vegetation Drought Response Index (VegDRI): a new integrated approach for monitoring drought stress in vegetation. *GISci. Remote Sens.* **45**, 16–46 (2008).
48. Houborg, R. & McCabe, M. F. A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. *ISPRS J. Photogramm. Remote Sens.* **135**, 173–188 (2018).
49. Johnson, D. M. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sens. Environ.* **141**, 116–128 (2014).
50. Horvath, G. CMAC neural network as an SVM with B-spline kernel functions. In *Proceedings of the 20th IEEE Instrumentation Technology Conference* (*Cat. No. 03CH37412*) Vol. 2, 1108–1113 (2003).
51. Huang, G.-B., Ding, X. & Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **74**, 155–163 (2010).
52. Maliha, A., Yusof, R. & Shapiai, M. I. Extreme learning machine for structured output spaces. *Neural Comput. Appl.* **30**, 1251–1264 (2018).
53. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **43**, 772–777 (1989).
54. Sonobe, R., Hirono, Y. & Oi, A. Non-destructive detection of tea leaf chlorophyll content using hyperspectral reflectance and machine learning algorithms. *Plants* **9**, 368 (2020).
55. Nonaka, K. Nitrogenous environmental load in tea fields and fertilizer application technology for the reduction of the environmental load. *Tea Res. J.* **100**, 29–41 (2005).
56. Roy, P. S. Spectral reflectance characteristics of vegetation and their use in estimating productive potential. *Proc. Plant Sci.* **99**, 59–81 (1989).
57. Zarco-Tejada, P. J., Miller, J. R., Noland, T. L., Mohammed, G. H. & Sampson, P. H. Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **39**, 1491–1507 (2001).
58. le Maire, G., François, C. & Dufrêne, E. Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. *Remote Sens. Environ.* **89**, 1–28 (2004).
59. Blackburn, G. A. Hyperspectral remote sensing of plant pigments. *J. Exp. Bot.* **58**, 855–867 (2007).
60. Elvidge, C. D. & Chen, Z. Comparison of broad-band and narrow-band red and near-infrared vegetation indices. *Remote Sens. Environ.* **54**, 38–48 (1995).
61. Filella, I., Serrano, L., Serra, J. & Peñuelas, J. Evaluating wheat nitrogen status with canopy reflectance indices and discriminant analysis. *Crop Sci.* **35**, 1400–1405 (1995).
62. Danson, F. M., Steven, M. D., Malthus, T. J. & Clark, J. A. High-spectral resolution data for determining leaf water content. *Int. J. Remote Sens.* **13**, 461–470 (1992).
63. Inoue, Y., Morinaga, S. & Shibayama, M. Non-destructive estimation of water status of intact crop leaves based on spectral reflectance measurements. *Jpn. J. Crop Sci.* **62**, 462–469 (1993).
64. Aldakheel, Y. Y. & Danson, F. M. Spectral reflectance of dehydrating leaves: measurements and modelling. *Int. J. Remote Sens.* **18**, 3683–3690 (1997).
65. Ceccato, P., Flasse, S., Tarantola, S., Jacquemoud, S. & Grégoire, J.-M. Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sens. Environ.* **77**, 22–33 (2001).
66. Konishi, S., Miyamoto, S. & Taki, T. Stimulatory effects of aluminum on tea plants grown under low and high phosphorus supply. *Soil Sci. Plant Nutr.* **31**, 361–368 (1985).
67. Sano, T., Horie, H. & Hirono, Y. Effect of shading intensity on morphological and color traits and on chemical components of new tea (*Camellia sinensis* L.) shoots under direct covering cultivation. *J. Sci. Food Agric.* **98**, 5666–5676 (2018).
68. Zarco-Tajeda, P. J., Miller, J. R., Haboudane, D., Tremblay, N. & Apostol, S. Detection of chlorophyll fluorescence in vegetation from airborne hyperspectral CASI imagery in the red edge spectral region. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings* (*IEEE Cat. No. 03CH37477*) Vol. 1, 598–600 (2003).
69. Sanches, I. D., Souza Filho, C. R. & Kokaly, R. F. Spectroscopic remote sensing of plant stress at leaf and canopy levels using the chlorophyll 680 nm absorption feature with continuum removal. *ISPRS J. Photogramm. Remote Sens.* **97**, 111–122 (2014).

70. Genkawa, T. *et al.* Baseline correction of diffuse reflection near-infrared spectra using searching region standard normal variate (SRSNV). *Appl. Spectrosc.* **69**, 1432–1441 (2015).
71. Maleki, M. R., Mouazen, A. M., Ramon, H. & De Baerdemaeker, J. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosyst. Eng.* **96**, 427–433 (2007).
72. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochim. Biophys. Acta* **975**, 384–394 (1989).
73. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn. (Springer Science and Business Media, New York, 2009).
74. Villar, A. *et al.* Cider fermentation process monitoring by Vis-NIR sensor system and chemometrics. *Food Chem.* **221**, 100–106 (2017).
75. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
76. Snoek, J. *et al.* Scalable Bayesian Optimization Using Deep Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, 2171–2180 (2015).
77. Chang, C.-W., Laird, D. A., Mausbach, M. J. & Hurburgh, C. R. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J.* **65**, 480–490 (2001).
78. Du, C. *et al.* Determination of soil properties using Fourier transform mid-infrared photoacoustic spectroscopy. *Vib. Spectrosc.* **49**, 32–37 (2009).
79. Razakamanarivo, R. H., Grinand, C., Razafindrakoto, M. A., Bernoux, M. & Albrecht, A. Mapping organic carbon stocks in eucalyptus plantations of the central highlands of Madagascar: a multiple regression approach. *Geoderma* **162**, 335–346 (2011).
80. Kewley, R. H., Embrechts, M. J. & Breneman, C. Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Trans. Neural Netw.* **11**, 668–679 (2000).
81. Cortez, P. & Embrechts, M. J. Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf. Sci.* **225**, 1–17 (2013).

Acknowledgements

We thank Mr. Hiromitsu Sato for providing rooted tea cutting cv. 'Koganemidori'. This research was supported by the Agriculture, Forestry and Fisheries Research Council (No. 1919102; R.S., Y.H., A.M., and T.I.), the Japanese Society for the Promotion of Science Grant-in-Aid for Scientific Research (No. 19K06313; R.S. and Y.H.), and the ESPEC Foundation for Global Environment Research and Technology (Charitable Trust; H.Y.). We thank Simon Partridge, PhD, from Edanz Group (https://en-author-services.edanzgroup.com/) for editing a draft of this manuscript.

Author contributions

H.Y., R.S., and T.I. designed this study. H.Y., Y.H., A.M., and T.I. managed the tea plants for experiments. H.Y. analyzed the nitrogen and chlorophyll contents. H.Y. and R.S. measured reflectance and performed modelling. H.Y., R.S., and T.I. performed most data visualization and writing. H.Y., R.S., Y.H., A.M., and T.I acquired funding. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73745-2.

Correspondence and requests for materials should be addressed to R.S. or T.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.