



OPEN

## Potential of spectroscopic analyses for non-destructive estimation of tea quality-related metabolites in fresh new leaves

Hiroto Yamashita<sup>1,2</sup>, Rei Sonobe<sup>1,3</sup>✉, Yuhei Hirono<sup>3,4</sup>, Akio Morita<sup>1,3</sup> & Takashi Ikka<sup>1,3</sup>✉

Spectroscopic sensing provides physical and chemical information in a non-destructive and rapid manner. To develop non-destructive estimation methods of tea quality-related metabolites in fresh leaves, we estimated the contents of free amino acids, catechins, and caffeine in fresh tea leaves using visible to short-wave infrared hyperspectral reflectance data and machine learning algorithms. We acquired these data from approximately 200 new leaves with various status and then constructed the regression model in the combination of six spectral patterns with pre-processing and five algorithms. In most phenotypes, the combination of de-trending pre-processing and Cubist algorithms was robustly selected as the best combination in each round over 100 repetitions that were evaluated based on the ratio of performance to deviation (RPD) values. The mean RPD values were ranged from 1.1 to 2.7 and most of them were above the acceptable or accurate threshold (RPD = 1.4 or 2.0, respectively). Data-based sensitivity analysis identified the important hyperspectral regions around 1500 and 2000 nm. Present spectroscopic approaches indicate that most tea quality-related metabolites can be estimated non-destructively, and pre-processing techniques help to improve its accuracy.

Plants collectively produce many metabolites with estimates ranging from 100,000 to 1 million, and many metabolites are thought to play essential roles in resistance to biotic stresses and tolerance of abiotic stresses<sup>1–5</sup>. In addition, natural products synthesized in plants provide indispensable resources for human health and survival<sup>5</sup>. Given the importance of plant metabolites to plant development and adaptation, and for human health, various quantitative and qualitative analyses have been developed. The main examples are based on chromatography techniques such as gas chromatography or high-performance liquid chromatography (HPLC) with improved mass resolution and sensitivity<sup>6,7</sup>. However, these analytical methods require the destructive collection and pre-treatment of plant samples, which makes them slow in acquiring analytical data and unsuitable for real-time diagnosis of metabolite level.

Hyperspectral reflectance sensing is an established spectroscopic method that can provide rapid analysis without the need for sample pre-treatment. It is commonly applied to visible (VIS; 400–700 nm), near-infrared (NIR; 700–1000 nm), and short-wave infrared (SWIR; 1000–2500 nm) spectral ranges and has been used to estimate leaf pigments and water contents<sup>8,9</sup>. The VIS is dominated by absorption of the photosynthetic pigments such as chlorophylls, carotenoids, and anthocyanins<sup>8</sup>. On the other hand, NIR spectroscopy is directly relevant to the overtones and combinations of the fundamental C–H, O–H, and N–H bonds in organic molecules<sup>10,11</sup>. Thus, NIR spectroscopy provides physical and chemical information and has shown good potential in estimating different parameters in biotic samples, including metabolites in plants, agricultural products, and food<sup>12–14</sup>. In addition, machine learning techniques provide powerful tools for constructing regression or classification models in agricultural indices from hyperspectral reflectance data<sup>15</sup>. The methodology of machine learning algorithms provides a flexible model not only for data-driven decision-making but also for capturing expertise into the algorithms<sup>16</sup>. The technique shows good potential for analyzing hyperspectral reflectance data with all

<sup>1</sup>Faculty of Agriculture, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka 422-8529, Japan. <sup>2</sup>United Graduate School of Agricultural Science, Gifu University, 1-1 Yanagito, Gifu 501-1193, Japan. <sup>3</sup>Institute for Tea Science, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka 422-8529, Japan. <sup>4</sup>Division of Tea Research, Institute of Fruit Tree and Tea Science, National Agriculture and Food Research Organization (NARO), 2769 Shishidoi, Kanaya, Shimada, Shizuoka 428-8501, Japan. ✉email: sonobe.rei@shizuoka.ac.jp; ikka.takashi@shizuoka.ac.jp

spectral information based on a large number of bands<sup>17</sup>. Machine learning techniques also enable the assessment of hyperspectral features that are informative for high accuracy predictive modelling<sup>16,18</sup>.

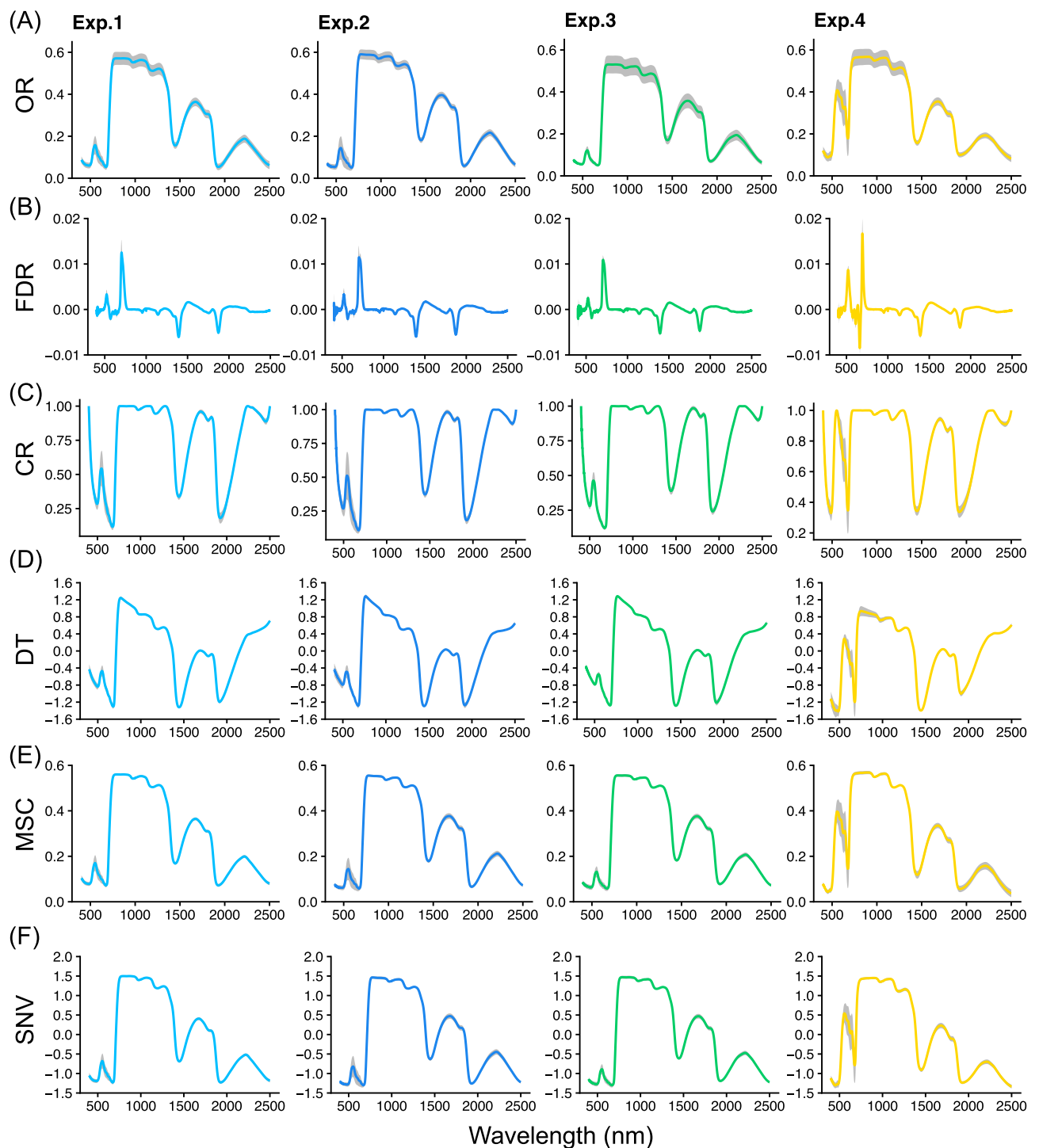
Tea plants (*Camellia sinensis* L.) are mainly distributed and cultivated in Asia to produce several tea types, such as green tea, oolong tea, and black tea, which are popular non-alcoholic beverages consumed all over the world. Tea-drinking reportedly has numerous and diverse health benefits<sup>19</sup>. Generally, tea quality and function are defined by the profile of various chemical components, such as catechins, caffeine, and theanine, which are characteristics to tea leaves. Tea catechins, which comprise a major class of polyphenols, contribute to the taste of astringency and bitterness of tea and have been studied for their health functions such as antibacterial activities<sup>20</sup> and free radical scavenging activities<sup>21</sup>. Free amino acids, especially glutamate (Glu) and theanine, contribute to the *umami* taste of green tea<sup>22,23</sup>. In particular, theanine, a unique amino acid in tea plants, has the activities of promoting relaxation<sup>24</sup> and reducing blood pressure<sup>25</sup>. Caffeine (1,3,7-trimethylxanthine) is a kind of purine alkaloid and its consumption may be associated with a reduced risk for type 2 diabetes<sup>26</sup>, but excessive intake of caffeine may cause inflammation of the digestive organs, insomnia, and arrhythmia<sup>27</sup>. Thus, unique tea quality-related metabolites are the most important agronomic traits targeted by modern and future tea cultivation and breeding. To evaluate the levels of these metabolites, many analytical tools have been employed to quantify tea quality-related metabolites including free amino acids, catechins, and caffeine contents in tea samples. Many analytical methods have been based on HPLC<sup>28,29</sup> and capillary electrophoresis<sup>30,31</sup>, but these methods destructively use plant tissues and are time-consuming and expensive to perform. Therefore, a rapid and accurate method for the evaluation of quantitative traits in tea leaves is in high demand for tea cultivation management and breeding programs. The NIR-based estimation of some chemical components in ground tea leaves has been established by previous studies<sup>32–34</sup>. Few studies have been reported in a non-destructive method for fresh leaves<sup>35,36</sup>. Huang et al.<sup>35</sup> have reported non-destructive estimation methods for four main catechins and caffeine in fresh green leaves based on VIS–NIR spectra (400–2498 nm) and partial least squares (PLS) model. However, the outcomes of this study were limited by fewer tea quality-related metabolites and the sample status from leaf positions and fewer tea quality-related metabolites, which cannot achieve robust results in actual agricultural management.

We have achieved the non-destructive estimation of chlorophyll and nitrogen contents in tea leaves by combining the VIS–NIR–SWIR (400–2500 nm) hyperspectral reflectance data and machine learning algorithms<sup>37</sup>. In the current study, we acquired the reflectance and 15 tea quality-related metabolites traits from the various nitrogen conditions, the leaf-stage, shading conditions, and albino tea leaves to construct the robust models. Pre-processing techniques and machine learning algorithms for hyperspectral data were used to perform regression modelling to non-destructively estimate the contents of free amino acids, catechins, and caffeine as tea quality-related metabolites in new fresh leaves. Our modelling indicated that most tea quality-related metabolites can be estimated by VIS–NIR–SWIR hyperspectral reflectance data and machine learning algorithms and that pre-processing techniques help to improve its accuracy. In particular, the combination of de-trending (DT) pre-processing methods and Cubist algorithms showed the highest model performance for most tea quality-related metabolites.

## Results

**Data distribution of reflectance data and tea quality-related metabolite contents.** Original reflectance (OR) data were obtained at 1-nm steps across the 400 to 2500 nm wavelength from approximately 200 leaves in four experiment conditions. Five pre-processing methods, namely first derivative reflectance (FDR), continuum-removed (CR), standard normal variate (SNV), multiplicative scatter correction (MSC), and DT, were applied to the OR data. Several spectral patterns were observed in OR and pre-processed reflectance (Fig. 1). In the same leaves that were measured by reflectance, we analyzed catechins, caffeine, and FAAs as tea quality-related metabolites by HPLC and acquired 15 phenotypic traits. For catechins, the contents of (+)-gallocatechin (GC), (+)-catechin (C), (–)-epicatechin (EC), (–)-epigallocatechin (EGC), (–)-catechin gallate (CG), (–)-epicatechin gallate (ECG), (–)-epigallocatechin gallate (EGCG), (–)-epigallocatechin-3-O-(3-O-methyl)-gallate (EGCG-3''Me), and total catechins were in the ranges of 3.4–64.6, 0.5–19.2, 1.1–25.3, 8.4–339.4, 21.4–459.4, 46.8–1003.1, 91.0–619.8, 1.3–43.3, and 206.2–2528.7  $\mu\text{g cm}^{-2}$ , respectively (Fig. 2). For FAAs, the contents of aspartate (Asp), glutamate (Glu), arginine (Arg), theanine (Thea), and total FAAs were in the ranges of 1.6–59.3, 3.1–49.1, 0.9–346.4, 0.2–264.5, and 12.3–746.0  $\mu\text{g cm}^{-2}$ , respectively (Fig. 2). Caffeine content was in the range of 1.8–393.1  $\mu\text{g cm}^{-2}$  (Fig. 2). The coefficient of variation (CV) in 15 phenotypes was in the range of 33.7%–138.6% (Fig. 2).

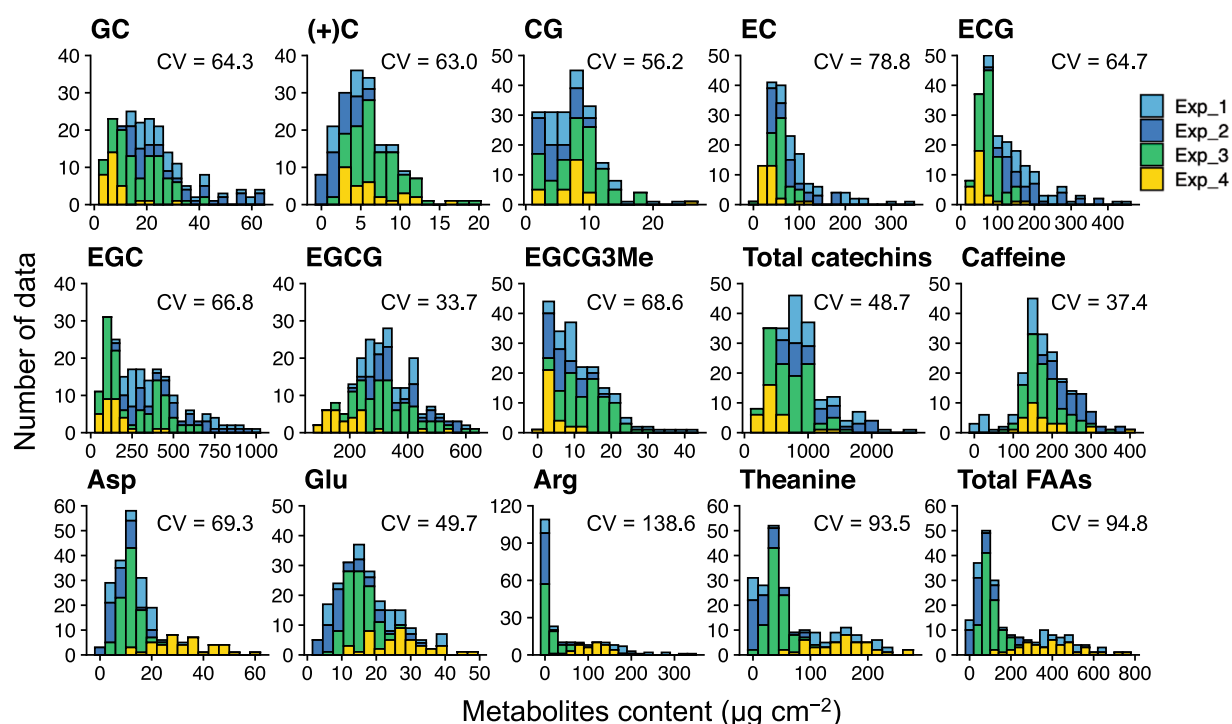
**Best combination of pre-processing and machine learning algorithms in regression model performance.** Using six spectral patterns (OR, FDR, CR, SNV, MSC, and DT) and five machine learning algorithms, Random Forest (RF), Support Vector Machine (SVM), Cubist, Stochastic Gradient Boosting (SGB), and Kernel-based Extreme Learning Machine (KELM), we performed regression modelling for 15 phenotypes of tea quality-related metabolites (Supplementary Fig. S1). Model performances in the combination of pre-processing and machine learning algorithms were evaluated based on the ratio of performance to deviation (RPD) values and robustness over 100 repetitions (Supplementary Table S2). In most phenotypes, the combination of DT and Cubist (DT-Cubist) was selected most often as the best performing combination in each round among the 100 repetitions (Table 1, Supplementary Table S2). The model performance based on DT-Cubist was different between the 15 phenotypes (Fig. 3A; two-way ANOVA,  $P < 0.001$ ). Except for CG and EGCG-3''Me, the mean RPD values in most of them were above the acceptable threshold (RPD = 1.4)<sup>38</sup>. In GC, EC, ECG, EGC, total catechins, Asp, and total FAAs, the mean RPD values were above the accurate threshold (RPD = 2.0)<sup>38</sup>. The modelling based on DT-Cubist significantly increased model performance over that based on OR-Cubist (Fig. 3A;



**Figure 1.** Pre-processing spectral patterns of original reflectance (OR) in tea leaves. Five pre-processing techniques were applied to the OR (A) base: first derivative reflectance (FDR, B), continuum-removed (CR, C), de-trending (DT, D), multiplicative scatter correction (MSC, E), and standard normal variate transformation (SNV, F). Colors in spectra (Exp. 1, light blue; Exp. 2, blue; Exp. 3, green; Exp. 4, yellow) and gray indicate mean and standard deviation, respectively. Figures were visualized by the R package “ggplot2” ver. 3.3.2.

two-way ANOVA,  $P < 0.001$ ). These results were also supported by the root-mean-square error (RMSE) values and the coefficient of determination ( $R^2$ ) values as a model performance index (Fig. 3B, Table 2).

**Detection of important hyperspectral regions by DSA.** Data-based sensitivity analysis (DSA) was performed to detect important hyperspectral regions in models to estimate tea quality-related metabolites, and their results based on OR-Cubist and DT-Cubist were visualized at 50-nm intervals (Fig. 4). Different shapes of



**Figure 2.** Data distribution of 15 phenotypes for tea quality-related metabolites. Number of samples: 201, 201, and 215 for catechins, caffeine and free amino acids (FAA), respectively. Coefficient of variation (CV) value for each metabolite is included on the Figure. Figures were visualized by the R package “ggplot2” ver. 3.3.2.

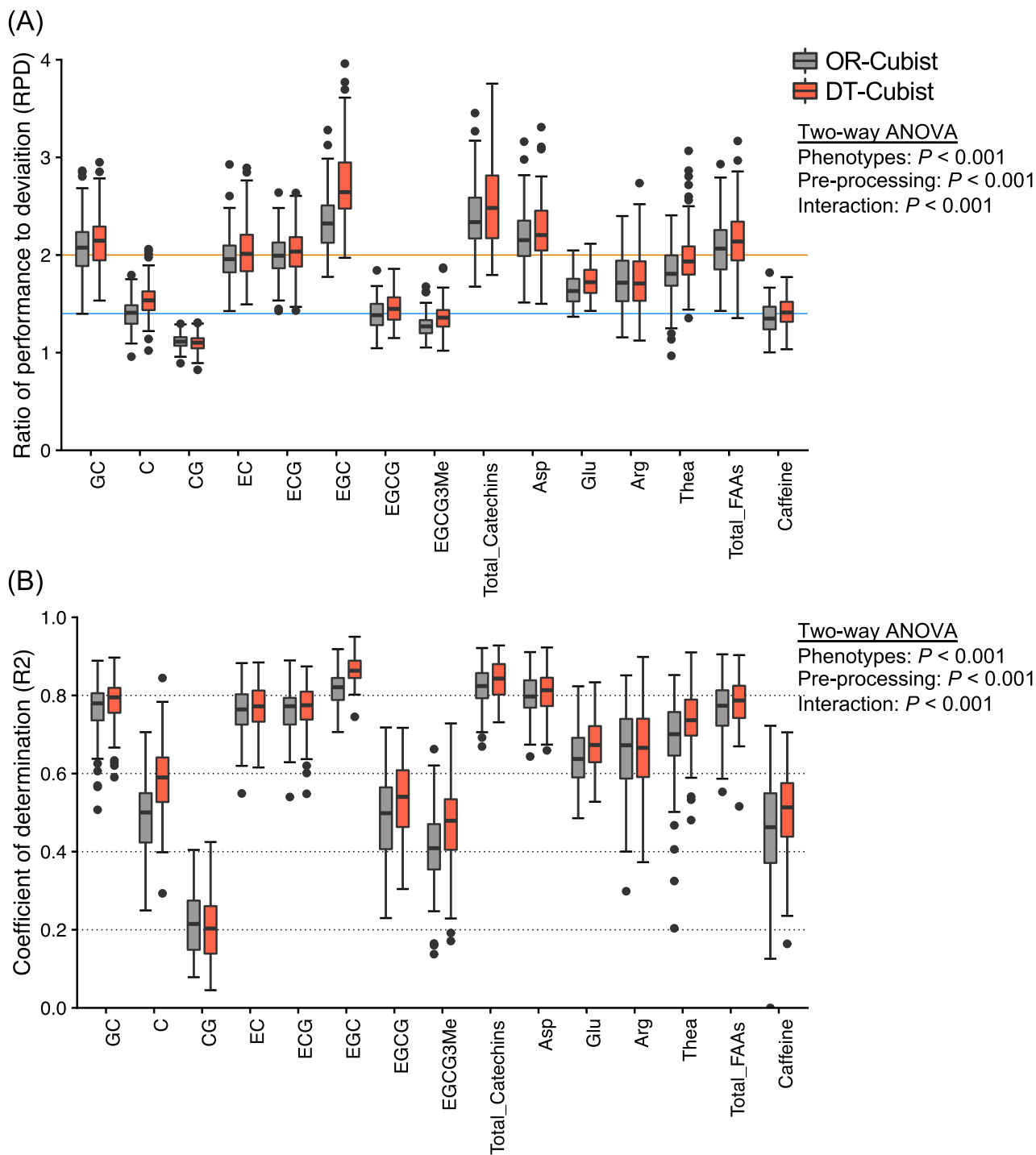
Phenotypes		Most selected pre-processing and models	Frequency (100 repeat <sup>-1</sup> )
Catechins	GC	DT-cubist	32
	C	SNV-cubist	17
	CG	DT-RF, OR-cubist	8
	EC	DT-cubist	18
	ECG	DT-cubist	23
	EGC	DT-cubist	65
	EGCG	DT-cubist	37
	EGCG3Me	DT-cubist	23
	Total	DT-cubist	42
FAAs	Asp	DT-cubist	23
	Glu	DT-cubist	50
	Arg	DT-cubist	25
	Thea	DT-cubist	27
	Total	DT-cubist	34
Caffeine		DT-cubist	20

**Table 1.** Best combination of pre-processing and machine learning algorithms after 100 repetitions. Combination of pre-processing and machine learning algorithms were evaluated based on RPD values.

DSA plots were observed for caffeine and individual catechins and amino acids (Fig. 4). For catechins without CG and EGCG-3''Me that showed poor prediction performance, the peak region consisting of high importance values was observed around 2000 nm (Fig. 4). For amino acids, the peak region of high importance values was around 1500 nm and 2000 nm (Fig. 4), and that for caffeine was around 750 nm and 1350 nm (Fig. 4).

## Discussion

To enable the non-destructive estimation of FAAs, catechins, and caffeine as tea quality-related metabolites, we performed regression modelling by combining the VIS–NIR–SWIR (400–2500 nm) hyperspectral reflectance data and machine learning algorithms. Datasets of hyperspectral data and tea quality-related metabolite contents were obtained from approximately 200 new leaves grown under different N conditions in hydroponics or from shading



**Figure 3.** Model performance and robustness based on OR-Cubist and DT-Cubist for tea quality-related metabolites. The ratio of performance to deviation (RPD, **A**) and coefficient of determination ( $R^2$ , **B**) were applied to evaluate the accuracy of each model. A stratified sampling approach for modelling was repeated 100 times to obtain robust results. Figure are plots of the RPD and  $R^2$  values in each repeat. Orange and blue lines indicate RPD values of 1.4 and 2.0, respectively, as accuracy thresholds. Statistical tests for significant differences by two-way ANOVA are shown on the right side of the Figure. Figures were visualized by the R package “ggplot2” ver. 3.3.2.

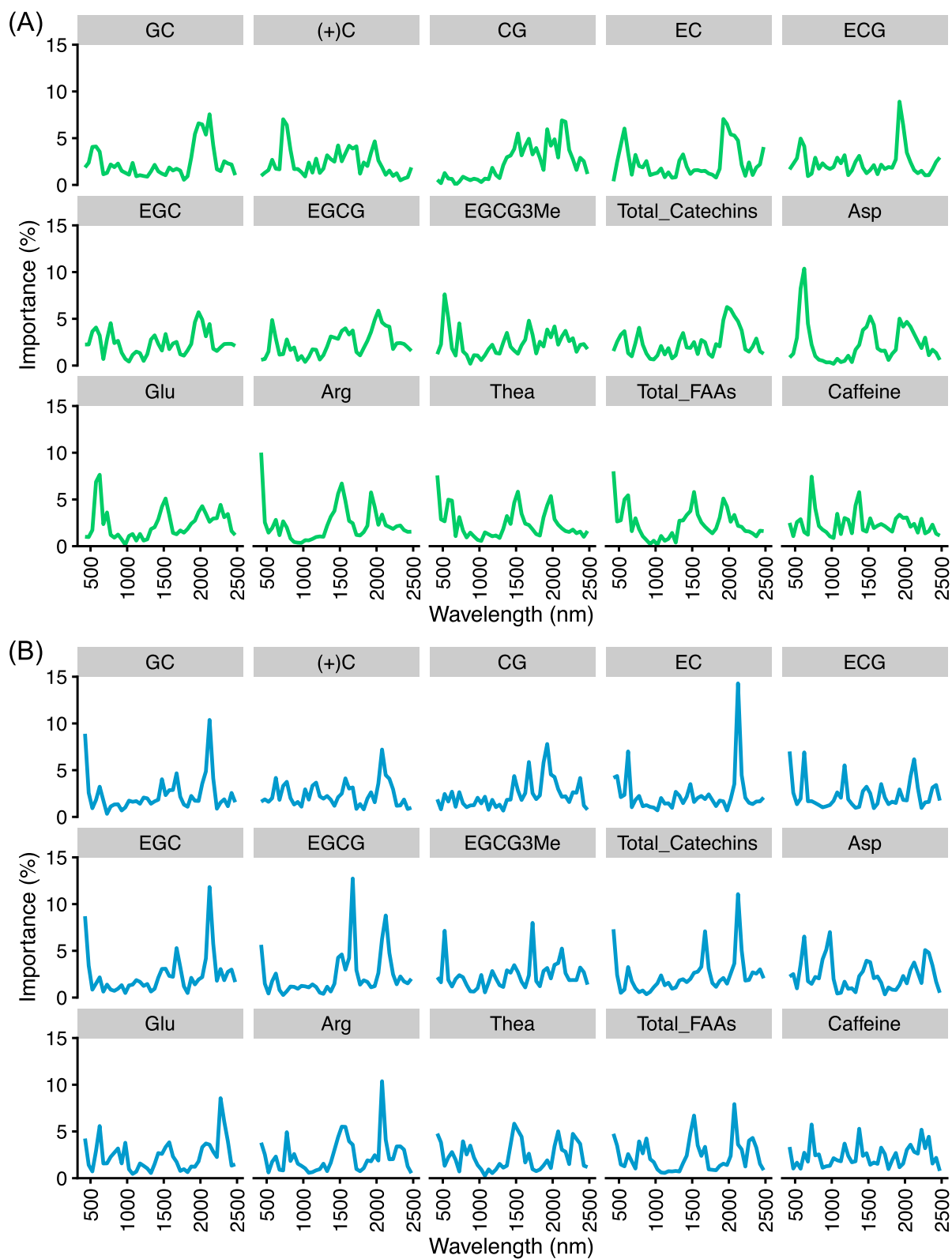
cultivations. The data showed wide variation that the CV in 15 phenotypes was in the range of 33.7%–138.6% (Fig. 2). The CV of EGCG (33.7%), ECG (64.7%), EGC (66.8%), EC (78.8%), caffeine (37.4%) in this study were higher than these (EGCG, 24.2%; ECG, 24.3%; EGC, 34.7%; EC, 14.0%; caffeine, 16.7%) in the previous study<sup>35</sup>. These results indicate that present datasets are suitable for robust regression modelling.

Phenotypes	Observed values (mg cm <sup>-2</sup> )			DT-cubist																Model performance <sup>a</sup>			
				RPD <sub>v</sub>			R <sup>2</sup> <sub>v</sub>			RMSE <sub>v</sub> (mg cm <sup>-2</sup> )			RPD <sub>p</sub>			R <sup>2</sup> <sub>p</sub>			RMSE <sub>p</sub> (mg cm <sup>-2</sup> )				
Catechins	GC	21.2	±	13.7	2.22	±	0.30	0.79	±	0.06	5.96	±	1.01	2.10	±	0.30	0.78	±	0.06	6.70	±	0.90	Accurate
	C	5.3	±	3.3	1.56	±	0.20	0.59	±	0.11	2.20	±	0.43	1.50	±	0.20	0.58	±	0.10	2.10	±	0.30	Acceptable
	CG	7.5	±	4.2	1.18	±	0.11	0.29	±	0.12	3.56	±	0.56	1.10	±	0.10	0.21	±	0.09	3.80	±	0.50	Poor
	EC	67.4	±	53.2	2.23	±	0.38	0.80	±	0.07	23.86	±	4.80	2.00	±	0.30	0.77	±	0.06	27.70	±	4.70	Accurate
	ECG	119.3	±	77.2	2.17	±	0.29	0.79	±	0.06	33.66	±	5.82	2.00	±	0.20	0.77	±	0.06	40.00	±	5.30	Accurate
	EGC	317.6	±	212.1	2.80	±	0.37	0.87	±	0.04	75.34	±	10.80	2.70	±	0.40	0.86	±	0.03	80.70	±	10.00	Accurate
	EGCG	320.6	±	107.9	1.50	±	0.19	0.55	±	0.11	71.87	±	10.26	1.50	±	0.20	0.54	±	0.09	72.80	±	8.70	Acceptable
	EGCG3Me	10.8	±	7.4	1.44	±	0.17	0.54	±	0.11	5.08	±	0.78	1.40	±	0.10	0.47	±	0.11	5.70	±	0.70	Poor
Total	869.7	±	423.2	2.46	±	0.46	0.83	±	0.06	171.14	±	30.96	2.50	±	0.40	0.84	±	0.05	175.70	±	28.30	Accurate	
FAAs	Asp	15.4	±	10.7	2.18	±	0.31	0.80	±	0.06	5.02	±	0.62	2.20	±	0.30	0.81	±	0.05	4.80	±	0.50	Accurate
	Glu	17.6	±	8.8	1.70	±	0.18	0.66	±	0.07	5.10	±	0.54	1.70	±	0.20	0.68	±	0.07	5.20	±	0.50	Acceptable
	Arg	51.1	±	70.9	1.88	±	0.32	0.71	±	0.10	39.16	±	7.60	1.80	±	0.30	0.67	±	0.11	41.00	±	7.30	Acceptable
	Thea	69.9	±	65.3	1.98	±	0.26	0.74	±	0.07	33.52	±	4.89	2.00	±	0.30	0.74	±	0.08	34.10	±	4.70	Acceptable
	Total	181.2	±	171.7	2.25	±	0.34	0.80	±	0.06	78.23	±	11.12	2.20	±	0.30	0.78	±	0.06	81.00	±	11.00	Accurate
Caffeine	183.4	±	68.7	1.39	±	0.19	0.48	±	0.13	49.58	±	8.35	1.40	±	0.20	0.50	±	0.11	51.80	±	6.90	Acceptable	

**Table 2.** Summary of validation and prediction performance based on DT-Cubist in 15 phenotypes for tea quality-related metabolites. “v” and “p” in each index means validation and prediction, respectively. <sup>a</sup>Prediction performance is represented as described by Chang et al., (2001); RPD<sub>p</sub> > 2.0, accurate prediction; 1.4 < RPD<sub>p</sub> < 2.0, acceptable prediction; RPD<sub>p</sub> < 1.4, poor prediction.

We applied five pre-processing techniques (Fig. 1; FDR, CR, DT, MSC, and SNV) to the OR data to enhance the more chemically associated peaks by reducing noise from spectral data and the effects of baseline shifts and overall curvature over the OR. Then we compared the model performance in the combination of six spectral patterns (OR, FDR, CR, SNV, MSC, and DT) and five machine learning algorithms (RF, SVM, Cubist, SGB, and KELM) based on the RPD values and robustness over 100 repetitions (Supplementary Table S2). In most phenotypes, the combination of DT and Cubist (DT-Cubist) was selected most often as the best performing combination in each round among the 100 repetitions (Table 1, Supplementary Table S2). DT has been used to correct wavelength-dependent scattering effects and to account for the variation in baseline shift and curvilinearity by fitting a second-degree polynomial through each spectrum<sup>39</sup>. Therefore, these results suggest that pre-processing based on DT was effective in improving accuracies when VIS-NIR-SWIR (400–500 nm) hyperspectral reflectance data from plant leaves were applied to the regression modelling. Cubist algorithms can generate so-called committee models that consist of a set of consecutive rule-based models to correct the predictions of previous member models<sup>40</sup>; this approach is computationally efficient and well suited to big data analytics<sup>40</sup>. Cubist is better equipped to handle extrapolations out of range of the training target data by relying on a rule-based multivariate linear regression model rather than an ensemble of decision trees with interconnected leaves associated with rigid target predictions<sup>41</sup>. Furthermore, Cubist algorithms achieved the best performance in a comparison of a large collection composed of 77 popular regression models<sup>42</sup>. Previous studies also showed that the Cubist algorithm had the potential of an efficient model algorithm for various plant traits using reflectance data such as leaf area index<sup>43</sup>. Our previous study also showed that the Cubist algorithm had the best regression performance with VIS-NIR-SWIR (400–2500 nm) hyperspectral reflectance data and the contents of N and chlorophyll in tea leaves<sup>37</sup>. These results and previous studies strongly show that the combination of the pre-processing technique based on the DT-Cubist algorithm was suitable for regression modelling of the VIS-NIR-SWIR reflectance data in plants.

These regression models based on DT-Cubist archived that the mean RPD values in most of the 15 phenotypes were above the acceptable threshold (RPD = 1.4)<sup>38</sup> except for CG and EGCG-3''Me (Fig. 3A). For catechins and caffeine, the mean RPD values of GC, EC, ECG, EGC, and total catechins were above the accurate threshold (RPD = 2.0)<sup>38</sup>, but those of EGCG and caffeine were not (Fig. 3A). A previous study based on NIR analysis of ground tea leaves indicated that the calibration models for caffeine, EGC, C, EGCG, EC, ECG, and total catechins, except for GC and EGCG-3''Me, had high performance with high R<sup>2</sup> (more than 0.90)<sup>34</sup>. The model's performance for EGCG and caffeine differs from that of other catechins in this study, and these may not be caused by chemical properties. In the dataset for our modelling, the CV values of EGCG and caffeine were drastically lower than those for other catechins (Fig. 2). These low variations in the reference dataset of EGCG and caffeine could have affected the regression modelling performance. Our model performance (R<sup>2</sup> = 0.50–0.86) was inferior to that (R<sup>2</sup> = 0.89–0.94) of the report of Huang et al.<sup>35</sup> that also performed the regression modelling based on 400–2498 nm reflectance for some catechins and caffeine content in fresh tea new leaves. Although Huang et al.<sup>35</sup> acquired the reflectance data using a near-infrared spectrometer under a dark environment in the room, we non-destructively did use a leaf clipping unit on the site under a field condition that could also cause some effect of spectral noise. These differences in measurement methods may affect the prediction performance. However, our measurement method was more designed to be applied in actual agricultural fields. In the previous work of Lee et al.<sup>34</sup> and in this study, the estimation of EGCG-3''Me was low (Fig. 2). The EGCG-3''Me content in the cultivars, Benifuuki, Benifuji, and Benihomare was drastically higher than the other tea cultivars<sup>44</sup>, including Yabukita, which was used in this study. Adding these data for high-EGCG-3''Me-content cultivars to the reference data would expand the data variation and possibly improve model performance.



**Figure 4.** Detection of important hyperspectral regions by data-based sensitive analysis (DSA). Importance values, which were averaged over 100 replicates and accumulated at 50-nm intervals, were visualized as DSA results based on OR-Cubist (A) and DT-Cubist (B) treatment. Figures were visualized by the R package “ggplot2” ver. 3.3.2.

The contributions of hyperspectral regions to generate the regression models for tea quality-related metabolite contents were detected using DSA. The different shapes of DSA plots based on OR-Cubist and DT-Cubist were observed for caffeine and individual catechins and amino acids (Fig. 4). These results suggest that the machine learning algorithms separately determine the variable contributions of important spectral regions to estimate each metabolite. In most catechins, the peak region consisting of high importance values was observed around 2000 nm by DSA (Fig. 4). These results overlapped with spectral regions of known absorption features associated with phenolic compounds and the bending and stretching of C–H and O–H bonds<sup>45–47</sup>. In amino acids, the peak regions of high importance were observed around 1500 nm and 2000 nm by DSA (Fig. 4). These results were also consistent with previously reported spectral regions (e.g., 1520–1523 nm) for amino acid estimation<sup>45</sup>. DSA based on DT reflected the importance of these regions more than the other pre-processing patterns (Fig. 4, Supplementary Fig. S2). NIR and SWIR spectra in fresh leaf exhibit confounding factors in water absorption regions (approximately 1350–1450 and 1850–1975 nm) that may mask optical chemical features<sup>48–50</sup>. Our dataset also indicated that many catechins and FAAs contents were negatively and positively correlated with water content, respectively (Supplementary Figs. S3, S4). Although each metabolite in fresh tea leaves may be affected by the water content, the relationship between the model performance and the correlation of each metabolite and the water content was inconsistent (Fig. 3, Supplementary Figs. S3, S4), which indicates that the prediction model in this study has been constructed with an optimized model that takes into account the water content in fresh leaves.

The results of the present study suggest that spectroscopic analyses based on VIS–NIR–SWIR (400–2500 nm) hyperspectral reflectance data and machine learning algorithms have good potential to non-destructively estimate the contents of FAAs, catechins, and caffeine as tea quality-related metabolites in new fresh leaves (Table 2). Our modelling approaches also indicate that pre-processing techniques help to improve the accuracy of model performance. In particular, the combination of DT pre-processing methods and Cubist algorithms showed the highest model performance for most tea quality-related metabolites. These findings will contribute to the non-destructive real-time diagnosis of metabolite levels in tea cultivation management and breeding programs.

## Methods

**Plant materials.** To obtain the dataset of tea quality-related metabolites contents with variations, a series of four experiments (Exp. 1 to Exp. 4) were conducted as described by Yamashita and Sonobe et al.<sup>37</sup>. New leaves were plucked from each experiment, and its reflectances were measured in site under a field condition. The reflectance datasets of these experiments were also used in our previous study<sup>37</sup>.

Exps. 1 and 2 were conducted based on hydroponic nutrient tests. One-year-old rooted tea cuttings of cv. Yabukita, a popular and leading Japanese cultivar for green tea, were used in the hydroponic cultures that were conducted under ambient light conditions in an unheated greenhouse (120 m<sup>2</sup>) at Shizuoka University (Shizuoka, Japan). A minor modification of the culture method described by Konishi et al. (1985) was used. Exp. 1 was conducted based on different six nitrogen (N) nutrient amount conditions using three to five biological replicates: 0 × N, 0.01 × N, 0.1 × N, 1 × N (40 mg L<sup>-1</sup>), 2 × N, 4 × N. After approximately 6 months of treatment, one or two new leaves were plucked from one individual. Exp. 2 was conducted based on low-light conditions (85% shading) and different four N nutrient amount conditions using three biological replicates: 0 × N, 0.1 × N, 1 × N, 4 × N. After 23 days for treatment, one or two new leaves were plucked from one individual.

Exp. 3 was conducted using mature tea plants (ridges) of cv. Yabukita at Shizuoka University (Shizuoka, Japan) based on low-light conditions (85% shading). New leaves in each leaf-stage were plucked from approximately random 15 shoots in sunlight and shaded tea ridges, and a total 87 leaves in Exp. 4, new leaves in each leaf-stage were plucked from approximately 20 shoots in a 7-year-old rooted tea cutting of a Japanese albino cultivar cv. Koganemidori, which had been bred from the natural etiolated bud sport, in hydroponics.

Finally, 215, 201, and 201 leaves samples in each experiment were freeze-dried, grounded into a fine powder, and then analyzed for free amino acids (FAAs), catechins, and caffeine, respectively.

**Reflectance measurements and pre-processing.** Reflectance data in new leaves was measured by an ASD FieldSpec4 unit (Analytical Spectral Devices, Boulder, CO, USA) with a leaf clipping (diameter 20 mm) (Supplementary Fig. S5). The widest part in the center of the leaf was measured three times so that a leaf clipping could fit inside the leaf and the average value of that was taken as the representative for each leaf. This spectroscopy contained three detectors, visible (VIS) and near-infrared (NIR), short-wave infrared (SWIR), and SWIR 2. ViewSpec Pro Software (Analytical Spectral Devices) was used to correct differences in the spectral drifts at 1000 and 1800 nm caused by inherent variation in these detector sensitivities. Finally, OR data were recorded with a sampling resolution of 1 nm steps across the entire wavelength domain from 400 to 2500 nm. Five pre-processing methods were also tested based on their success in previous studies, namely first FDR, CR, SNV, MSC, and DT. FDR is effective in reducing baseline variation and increasing the resolution of spectral peak features<sup>51,52</sup>. CR is a brightness normalization technique that has been applied to enhance related changes<sup>53</sup>. MSC and SNV have also been used to eliminate the effect of noise, baseline drift, and light scattering of the spectrogram<sup>54–56</sup>. DT has been used to correct wavelength-dependent scattering effects and accounts for the variation in baseline shift and curvilinearity by fitting a second-degree polynomial through each spectrum<sup>39</sup>. All methods were performed using R version 3.6.3 and the R package “prospectr” ver. 0.2.0.

**Measurement of tea quality-related metabolites.** Catechins and caffeine contents were measured according to the methods described by Horie et al.<sup>57</sup> and Yamashita et al.<sup>58</sup>. Dry ground leaf tissue (25 mg) was added to 5 mL of 50% (v/v) acetonitrile and shook with 130 strokes min<sup>-1</sup> for 60 min at room temperature. The suspended samples were centrifuged at 2000 × g for 15 min at 4 °C, and then the supernatants were individually passed through 0.45-µm polytetrafluoroethylene filters (Advantec, Tokyo, Japan). The resulting solutions were



stored at  $-30\text{ }^{\circ}\text{C}$  until they were analyzed by HPLC as described by Yamashita and Uchida et al.<sup>58</sup>. The eight catechins, GC, C, CG, EC, ECG, EGC, EGCG, EGCG-3''Me, and caffeine were quantified. Their total value without caffeine was also expressed as total catechins.

The FAAs contents were measured according to the method described by Goto et al.<sup>59</sup> and Yamashita et al.<sup>58</sup>. Dry ground leaf tissue (10 mg) was added to 10 mg of polyvinylpyrrolidone and 5 mL of ultra-pure water and was shook with  $130\text{ strokes min}^{-1}$  for 60 min at room temperature. The suspended samples were centrifuged at  $2000\times g$  for 15 min at  $4\text{ }^{\circ}\text{C}$ , and then the supernatants were individually passed through  $0.45\text{-}\mu\text{m}$  cellulose acetate filters (Advantec). The resulting solution was stored at  $-30\text{ }^{\circ}\text{C}$  until analysis by HPLC as described by<sup>58</sup>. Nine amino acids [Asp, asparagine (Asn), Glu, glutamine (Gln), serine (Ser), Arg, alanine (Aln), Thea, and  $\gamma$ -aminobutyric acid (GABA)] were quantified. Their total value was also expressed as total FAAs.

**Regression models based on machine learning algorithms.** The regression modelling was conducted as described by Yamashita and Sonobe et al.<sup>37</sup> with minor modification and its flow chart was shown in Supplementary Fig. S1. For modelling, a stratified random sampling approach was applied, for which strata were formed based on experiments and treatments, and then all measurements were divided into three dataset groups as follows; a training set (50%), which was used to fit the models; a validation set (25%), which was used to estimate the prediction error for model selection; and a test set (25%), which was used for assessing the generalization error in the final selected model. To evaluate the robustness of models, this flow was repeated 100 times before pre-processing the OR and generating regression models.

When performing regression modelling based on machine learning algorithms, a genetic algorithm (GA)-based approach was applied to select wavelengths using the "ga\_pls" function (with the parameter "GA.threshold" and others set as 50 and default values, respectively) of the R package "plsVarSel" ver. 0.9.6. and R ver. 3.6.3. GA were effective for removing noninformative wavelengths to construct simpler and better prediction models. Regression models were then constructed from the selected wavelengths using the following representative five algorithms: RF, SVM, Cubist, SGB, and KELM. The overviews of these five algorithms were described in Supplementary Table S1.

RF was performed and optimized with the five hyperparameters by the R package "randomForestSRC" ver. 2.9.3. SVM was performed with the Gaussian radial basis function kernel and optimized with the two hyperparameters by the R package "e1071" ver. 1.5-8. Cubist was performed and optimized with the two hyperparameters by the R package "Cubist" ver. 0.2.3. SGB was performed and optimized with the four hyperparameters by the R package "gbm" ver. 2.1.5. KELM was performed and optimized with the two hyperparameters by the MATLAB and Statistics Toolbox Release 2016a (MathWorks, Natick, MA, USA; source code downloaded from <https://www.ntu.edu.sg/home/egbhuang/>). The optimizations in the hyperparameters of these machine learning algorithms were conducted based on the Bayesian optimization approach that was applied with the Gaussian process<sup>60,61</sup> using the R package "rBayesianOptimization" ver. 1.1.0. The hyperparameters information of these algorithms is shown in Supplementary Table S1.

The validation ( $v$ ) and prediction ( $p$ ) accuracy of constructed models was assessed based on the following three indexes: the ratio of performance to deviation (RPD), the coefficient of determination ( $R^2$ ), root-mean-square error (RMSE). The performance of the prediction model was assessed according to the following three classes of RPD<sup>38,62,63</sup>: RPD  $> 2$ , accurate prediction; RPD of  $1.4\text{--}2$ , acceptable prediction; RPD  $< 1.4$ , poor prediction.

**Data-based sensitivity analysis (DSA).** To extract human-understandable knowledge from supervised learning black box data mining models, we performed the DSA<sup>64,65</sup> by using the "Importance" function of the R package "rminer" ver. 1.4.5, as previously described by Yamashita and Sonobe et al.<sup>37</sup>. Although DSA is similar to a computationally efficient one-dimensional sensitivity analysis<sup>64</sup>, this method uses several training samples instead of a baseline vector<sup>65</sup> and it could be applied to black-box functions by querying the fitted models with sensitivity samples and recording their responses.

Received: 13 November 2020; Accepted: 9 February 2021

Published online: 18 February 2021

## References

- Dixon, R. A. & Strack, D. Phytochemistry meets genome analysis, and beyond. *Phytochemistry* **62**, 815–816 (2003).
- Afendi, F. M. et al. Data mining methods for omics and knowledge of crude medicinal plants toward big data biology. *Comput. Struct. Biotechnol. J.* **4**, e201301010 (2013).
- Weng, J.-K. The evolutionary paths towards complexity: A metabolic perspective. *New Phytol.* **201**, 1141–1149 (2014).
- Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* **90**, 764–787 (2017).
- Fang, C., Fernie, A. R. & Luo, J. Exploring the diversity of plant metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
- Fernie, A. R., Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5**, 763–769 (2004).
- Wolfender, J.-L., Nuzillard, J.-M., van der Hoof, J. J. J., Renault, J.-H. & Bertrand, S. Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. *Anal. Chem.* **91**, 704–742 (2019).
- Carter, G. A. & Knapp, A. K. Leaf optical properties in higher plants: Linking spectral characteristics to stress and chlorophyll concentration. *Am. J. Bot.* **88**, 677–684 (2001).

9. Slaton, M. R., Raymond Hunt, E. & Smith, W. K. Estimating near-infrared leaf reflectance from leaf structural characteristics. *Am. J. Bot.* **88**, 278–284 (2001).
10. Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M. & Hanpin, M. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* **667**, 14–32 (2010).
11. Türker-Kaya, S. & Huck, C. W. A review of mid-infrared and near-infrared imaging: Principles, concepts and applications in plant tissue analysis. *Molecules* **22**, 1 (2017).
12. Nicolai, B. M. *et al.* Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biol. Technol.* **46**, 99–118 (2007).
13. Liu, Y., Gao, R.-J. & Sun, X.-D. Review of portable NIR instruments for detecting fruit interior quality. *Spectrosc. Spectr. Anal.* **30**, 2874–2878 (2010).
14. Prevolnik, M. *et al.* Accuracy of near infrared spectroscopy for prediction of chemical composition, salt content and free amino acids in dry-cured ham. *Meat Sci.* **88**, 299–304 (2011).
15. 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).
16. 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).
17. Van Wittenbergh, 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).
18. Panda, S. S., Ames, D. P. & Panigrahi, S. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing* **2**, 673–696 (2010).
19. Zhang, L. *et al.* Chemistry and biological activities of processed camellia sinensis teas: A comprehensive review. *Compr. Rev. Food Sci. Food Saf.* **18**, 1474–1495 (2019).
20. Fukai, K., Ishigami, T. & Hara, Y. Antibacterial activity of tea polyphenols against phytopathogenic bacteria. *Agric. Biol. Chem.* **55**, 1895–1897 (1991).
21. Bors, W. & Saran, M. Radical scavenging by flavonoid antioxidants. *Free Radic. Res. Commun.* **2**, 289–294 (1987).
22. Ekborg-ott, K. H., Taylor, A. & Armstrong, D. W. Varietal differences in the total and enantiomeric composition of theanine in tea. *J. Agric. Food Chem.* **45**, 353–363 (1997).
23. Narukawa, M., Morita, K. & Hayashi, Y. L-Theanine elicits an umami taste with inosine 5'-monophosphate. *Biosci. Biotechnol. Biochem.* **72**, 3015–3017 (2008).
24. Lu, K. *et al.* The acute effects of L-theanine in comparison with alprazolam on anticipatory anxiety in humans. *Hum. Psychopharmacol.* **19**, 457–465 (2004).
25. Yokogoshi, H. *et al.* Reduction effect of theanine on blood pressure and brain 5-hydroxyindoles in spontaneously hypertensive rats. *Biosci. Biotechnol. Biochem.* **59**, 615–618 (1995).
26. Iso, H., Wakai, K., Fukui, M. & Tamakoshi, A. The relationship between green tea and total caffeine intake and risk for self-reported type 2 diabetes among Japanese adults. *Ann. Intern. Med.* **144**, 554–562 (2006).
27. Chou, T. M. & Benowitz, N. L. Caffeine and coffee: Effect on health and cardiovascular disease. *Comp. Biochem. Physiol. C.* **109**, 173–189 (1994).
28. Miyauchi, S. *et al.* High-quality green tea leaf production by artificial cultivation under growth chamber conditions considering amino acids profile. *J. Biosci. Bioeng.* **118**, 710–715 (2014).
29. Yang, X. R., Ye, C. X., Xu, J. K. & Jiang, Y. M. Simultaneous analysis of purine alkaloids and catechins in *Camellia sinensis*, *Camellia pilophylla* and *Camellia assamica* var. *kucha* by HPLC. *Food Chem.* **100**, 1132–1136 (2007).
30. Horie, H., Mukai, T. & Kohata, K. Simultaneous determination of qualitatively important components in green tea infusions using capillary electrophoresis. *J. Chromatogr. A* **758**, 332–335 (1997).
31. Kotani, A., Takahashi, K., Hakamata, H., Kojima, S. & Kusu, F. Attomole catechins determination by capillary liquid chromatography with electrochemical detection. *Anal. Sci.* **23**, 157–163 (2007).
32. Goto, T. Studies on NIR analyses of the chemical components in fresh tea leaf and crude tea and the evaluation of tea quality. *Tea Res. J.* **1992**, 51–61 (1992).
33. Schulz, H., Engelhardt, U. H., Wegent, A., Drews, H. & Lapczynski, S. Application of near-infrared reflectance spectroscopy to the simultaneous prediction of alkaloids and phenolic substances in green tea leaves. *J. Agric. Food Chem.* **47**, 5064–5067 (1999).
34. Lee, M.-S., Hwang, Y.-S., Lee, J. & Choung, M.-G. The characterization of caffeine and nine individual catechins in the leaves of green tea (*Camellia sinensis* L.) by near-infrared reflectance spectroscopy. *Food Chem.* **158**, 351–357 (2014).
35. Huang, Y. *et al.* Development of simple identification models for four main catechins and caffeine in fresh green tea leaf based on visible and near-infrared spectroscopy. *Comput. Electron. Agric.* **173**, 105388 (2020).
36. Wang, Y.-J. *et al.* Onsite nutritional diagnosis of tea plants using micro near-infrared spectrometer coupled with chemometrics. *Comput. Electron. Agric.* **175**, 105538 (2020).
37. Yamashita, H., Sonobe, R., Hirono, Y., Morita, A. & Ikka, T. Dissection of hyperspectral reflectance to estimate nitrogen and chlorophyll contents in tea leaves based on machine learning algorithms. *Sci. Rep.* **10**, 17360 (2020).
38. 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).
39. 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).
40. Walton, J. T. Subpixel urban land cover estimation. *Photogramm. Eng. Remote Sens.* **74**, 1213–1222 (2008).
41. 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).
42. Fernández-Delgado, M. *et al.* An extensive experimental survey of regression methods. *Neural Netw.* **111**, 11–34 (2019).
43. 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).
44. Sano, M. *et al.* Simultaneous determination of twelve tea catechins by high-performance liquid chromatography with electrochemical detection. *Analyst* **126**, 816–820 (2001).
45. Bian, M. *et al.* Predicting foliar biochemistry of tea (*Camellia sinensis*) using reflectance spectra measured at powder, leaf and canopy levels. *ISPRS J. Photogramm. Remote Sens.* **78**, 148–156 (2013).
46. Kokaly, R. F. & Skidmore, A. K. Plant phenolics and absorption features in vegetation reflectance spectra near 1.66  $\mu\text{m}$ . *Int. J. Appl. Earth Obs. Geoinf.* **43**, 55–83 (2015).
47. Couture, J. J. *et al.* Spectroscopic determination of ecologically relevant plant secondary metabolites. *Methods Ecol. Evol.* **7**, 1402–1412 (2016).
48. Curran, P. J., Dungan, J. L., Macler, B. A., Plummer, S. E. & Peterson, D. L. Reflectance spectroscopy of fresh whole leaves for the estimation of chemical concentration. *Remote Sens. Environ.* **39**, 153–166 (1992).
49. Gao, B.-C. & Goetz, A. F. H. Extraction of dry leaf spectral features from reflectance spectra of green vegetation. *Remote Sens. Environ.* **47**, 369–374 (1994).
50. Ramoelo, A., Skidmore, A. K., Schlerf, M., Mathieu, R. & Heitkönig, I. M. A. Water-removed spectra increase the retrieval accuracy when estimating savanna grass nitrogen and phosphorus concentrations. *ISPRS J. Photogramm. Remote Sens.* **66**, 408–417 (2011).

51. Tsai, F. & Philpot, W. Derivative analysis of hyperspectral data. *Remote Sens. Environ.* **66**, 41–51 (1998).
52. Sun, X., Subedi, P., Walker, R. & Walsh, K. B. NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment. *Postharvest. Biol. Technol.* **163**, 111140 (2020).
53. Clark, R. N. & Roush, T. L. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res.* **89**, 6329–6340 (1984).
54. 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).
55. 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).
56. Ren, G., Sun, Y., Li, M., Ning, J. & Zhang, Z. Cognitive spectroscopy for evaluating Chinese black tea grades (*Camellia sinensis*): Near-infrared spectroscopy and evolutionary algorithms. *J. Sci. Food Agric.* **100**, 3950–3959 (2020).
57. Horie, H., Maeda-Yamamoto, M., Ujihara, T. & Kohata, K. Extraction of tea catechins for chemical analysis. *Tea Res. J.* **2002**, 60–64 (2002).
58. Yamashita, H. *et al.* Genomic predictions and genome-wide association studies based on RAD-seq of quality-related metabolites for the genomics-assisted breeding of tea plants. *Sci. Rep.* **10**, 17480 (2020).
59. Goto, T., Horie, H. & Mukai, T. Analysis of major amino acids in green tea by high-performance liquid chromatography coupled with OPA precolumn derivatization. *Tea Res. J.* **1993**, 29–33 (1993).
60. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
61. Snoek, J. *et al.* Scalable Bayesian Optimization Using Deep Neural Networks. In: *International Conference on Machine Learning*, pp. 2171–2180 (jmlr.org, 2015).
62. Du, C. *et al.* Determination of soil properties using Fourier transform mid-infrared photoacoustic spectroscopy. *Vib. Spectrosc.* **49**, 32–37 (2009).
63. 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).
64. 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).
65. 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., No. 20J10182; H.Y.), and the ESPEC Foundation for Global Environment Research and Technology (Charitable Trust; H.Y.). We thank Austin Schultz, PhD, from Edanz Group (<https://en-author-services.edanzgroup.com/ac>) 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 metabolites 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** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-83847-0>.

**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](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021