metabolites

MDPI

Article

Differentiation of Geographical Origin of White and Brown Rice Samples Using NMR Spectroscopy Coupled with Machine Learning Techniques

Maham Saeed [1], Jung-Seop Kim [1], Seok-Young Kim [1], Ji Eun Ryu [1], JuHee Ko [1], Syed Farhan Alam Zaidi [2], Jeong-Ah Seo [3], Young-Suk Kim [4], Do Yup Lee [5] and Hyung-Kyoon Choi [1,*]

[1] College of Pharmacy, Chung-Ang University, Seoul 06974, Korea
[2] Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Korea
[3] School of Systems Biomedical Science, Soongsil University, Seoul 06978, Korea
[4] Department of Food Science and Biotechnology, Ewha Womans University, Seoul 03760, Korea
[5] Center for Food and Bioconvergence, Department of Agricultural Biotechnology, Research Institute for Agricultural and Life Sciences, Seoul National University, Seoul 08826, Korea
* Correspondence: hykychoi@cau.ac.kr; Tel.: +82-2-8205605

Abstract: Rice (*Oryza sativa* L.) is a widely consumed food source, and its geographical origin has long been a subject of discussion. In our study, we collected 44 and 20 rice samples from different regions of the Republic of Korea and China, respectively, of which 35 and 29 samples were of white and brown rice, respectively. These samples were analyzed using nuclear magnetic resonance (NMR) spectroscopy, followed by analyses with various data normalization and scaling methods. Then, leave-one-out cross-validation (LOOCV) and external validation were employed to evaluate various machine learning algorithms. Total area normalization, with unit variance and Pareto scaling for white and brown rice samples, respectively, was determined as the best pre-processing method in orthogonal partial least squares–discriminant analysis. Among the various tested algorithms, support vector machine (SVM) was the best algorithm for predicting the geographical origin of white and brown rice, with an accuracy of 0.99 and 0.96, respectively. In external validation, the SVM-based prediction model for white and brown rice showed good performance, with an accuracy of 1.0. The results of this study suggest the potential application of machine learning techniques based on NMR data for the differentiation and prediction of diverse geographical origins of white and brown rice.

Keywords: rice; geographical origin; NMR spectroscopy; machine learning; prediction model

Citation: Saeed, M.; Kim, J.-S.; Kim, S.-Y.; Ryu, J.E.; Ko, J.; Zaidi, S.F.A.; Seo, J.-A.; Kim, Y.-S.; Lee, D.Y.; Choi, H.-K. Differentiation of Geographical Origin of White and Brown Rice Samples Using NMR Spectroscopy Coupled with Machine Learning Techniques. *Metabolites* **2022**, *12*, 1012. https://doi.org/10.3390/metabo12111012

Academic Editor: Anna Piasecka

Received: 23 August 2022
Accepted: 21 October 2022
Published: 24 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Rice (*Oryza sativa*) is a primary food source for almost 50% of the global population because of its high caloric content and various nutrients, such as minerals and vitamins [1]. Rice is an important crop in Asia and is widely consumed in various forms such as rice flour, cooked rice, and rice cookies [2]. There are two significant subspecies of *O. sativa*, indica and japonica, with an enormous number of varieties. Japonica is the most commonly cultivated crop in East Asia, particularly in Korea, Japan, and China. Genotype and environmental factors, such as rainfall intensity, soil, and temperature, heavily influence rice metabolite profiles [3]. Rice fraud has been a serious global problem. However, the assessment of botanical and geographical origin as well as cultivation methods of rice are very important [4].

Metabolomics has been employed in the field of crops and agriculture research to discriminate genetic and environmental differences, control crop quality, and determine geographical origin [5–7]. Various analytical platforms, such as gas chromatography–

mass spectrometry, liquid chromatography/mass spectrometry, nuclear magnetic resonance (NMR) spectroscopy, Fourier-transform infrared spectroscopy, and direct-infusion mass spectrometry, have been employed to discriminate the geographical origin of crops and plants [8].

Several studies have been conducted to determine the types or geographical origins of rice samples to prevent adulteration and substitution, and to standardize their safety and quality assurance. For example, various rice samples were distinguished by multi-element fingerprinting using high-resolution inductively coupled plasma mass spectrometry (ICP-MS) [1], elemental imaging using laser ablation ICP-MS [9], and ^{1}H-NMR spectroscopy coupled with principal component analysis (PCA) and discriminant analysis [10], and ^{1}H-NMR spectroscopy coupled with partial least square–discriminant analysis (PLS-DA) and independent component analysis [11]. Mass spectrometry coupled with the random forest (RF) classification algorithm has also been reported to differentiate white rice samples from China and Korea [12].

Recently, machine learning (ML) algorithms have been demonstrated to significantly enhance the predictability and validity of prediction models for the geographical origin of various crops and medicinal plants. ^{1}H-NMR and inductively coupled plasma atomic emission spectroscopy/ICP-MS techniques coupled with ML algorithms have been applied to discriminate the geographical origins of medicinal plants from Korea and China (*Astragalus membranaceus* and *Paeonia albiflora*) [13]. The geographical origin of 237 samples of asparagus (*Asparagus officinalis* L.) from six different countries (The Netherlands, Germany, Spain, Poland, Greece, and Peru) was successfully distinguished using ^{1}H-NMR spectroscopy coupled with ML algorithms [14]. *Ixeris denticulata* samples from eight different origins were differentiated using ultra-high performance liquid chromatography-quadrupole time-of-flight mass spectrometry followed by ML algorithms [15]. Differences in soybean seed vigor were evaluated using infrared spectroscopy and machine-learning techniques [16].

The total amount of rice imported into Korea in 2021 was 492,901 tons (accounting for 12.7% of domestic production), of which 40.8% was imported from China (196,322.2 tons of brown rice, 5001.8 tons of polished rice). Most of the imported rice from China into Korea was brown rice because of its advantages in storage and variety of utilization [17,18]. It is important to distinguish the geographical origin of white and brown rice. The problem of counterfeiting the origin of imported rice is leading to economic problems as well as confusion in the domestic rice market. Over the past four years, the number of cases of illegal distribution of imported rice in the Republic of Korea has been increasing every year with about 425 cases, which gives local farmers a sense of relative deprivation and disrupts the order of the healthy rice distribution market [19]. In addition, this has affected the increase in the inventory of rice in conjunction with the overproduction of domestic rice, resulting in a surge in the amount of sales loss for feed and inventory management costs [20]. However, there have been no reports on the differentiation of the geographical origins of rice samples (with two milling types) using ^{1}H-NMR analysis coupled with ML techniques.

In this study, we collected white and brown rice samples from different regions of the Republic of Korea (hereafter referred to as Korea) and China. We analyzed the data using NMR spectroscopy coupled with various ML algorithms such as PCA, orthogonal PLS-DA (OPLS-DA), random forest, decision tree, support vector machine (SVM), logistic regression, and k-nearest neighbors. The main objective of our study was to explore the utilization of NMR spectroscopy coupled with various ML algorithms to develop a convenient method for predicting the geographical origin of rice.

2. Materials and Methods

2.1. Rice Sample Collection

Rice samples, collected from Korea (44 samples) and China (20 samples) with two types of milling (white and brown rice) as shown in Supplementary Figure S1, were prepared for NMR spectroscopy analysis. Brown rice is obtained by dehusking of paddy rice, and white rice is obtained by removing the bran layer and the germ from the brown rice. Korean rice samples were harvested in 2018 and collected by the National Institute of Crop Science. The Korean rice samples were cultivated in Kangwon-do (Cheorwon, Chuncheon, Hoengseong, and Yangyang), Gyeonggi-do (Suwon, Hwaseong, Paju, and Yeoncheon), Chungcheongbuk-do (Chungju and Cheongju), Chungcheongnam-do (Asan and Seosan), Gyeongsangbuk-do (Sangju, Yecheon, and Gyeongju), Gyeongsangnam-do (Miryang and Haman), Jeollabuk-do (Jeonju, Kimje, and Jinan), and Jeollanam-do (Kangjin and Haenam). Chinese rice samples were harvested in 2018 and bought from online suppliers. These samples were obtained from Heilongjiang, Henan, Liaoning, Jilin, Jiangsu, Shandong, Shanxi, Hubei, and Sichuan provinces. The provinces, cities, and weather information for the rice samples are summarized in Supplementary Table S1 and Figure S2.

2.2. Chemicals and Reagents

Deuterium oxide (D_2O, 99.9% atom D) including 0.05% 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid sodium salt (TSP), deuterium oxide (D_2O, 99.9% atom D), and monopotassium phosphate (KH_2PO_4) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Sodium deuteroxide solution (NaOD, 99.5% atom D; 40% in D_2O) was purchased from Cambridge Isotope Laboratories, Inc. (Andover, MA, USA).

2.3. Pre-Preparation and Extraction of Rice

Pooled samples from each location were promptly frozen in liquid nitrogen, pulverized using a blender, and stored in a deep freezer until NMR analysis. Then, 100 mg of rice powder and 1.5 mL of 100% D_2O (0.1 mM TSP) were transferred into a 2 mL centrifuge tube (Eppendorf tube, Hamburg, Germany), vortexed for 1 min, and sonicated for 15 min. Subsequently, the suspension was centrifuged at 17,000× *g*, 4 °C for 10 min. A buffer solution of 90 mM KH_2PO_4 was prepared from D_2O, and NaOD was used to adjust the pH to 6.0. The clear supernatant was filtered using a 0.45 μm PVDF filter (Chemco Scientific, Osaka, Japan), and 600 μL of the sample was transferred into a 5 mm NMR tube (Norell, Landisville, NJ, USA).

2.4. Peak NMR Spectra Assignment

A 600-MHz Bruker Avance spectrometer (Bruker, Germany) was employed to analyze rice samples at 25 °C to record all NMR spectra. For the ^{1}H-NMR spectra, 64K data points were obtained with a relaxation delay of 2.0 s and a spectral width of 10,775.9 Hz. A total of 128 scans and an acquisition time of 3.0 s were used. Water suppression was conducted to exclude the region between δ = 4.7 and 5.0 using a pre-saturation pulse sequence (Bruker 1D noesygppr1d). For two-dimensional NMR spectra, ^{1}H–^{1}H correlation spectroscopy (COSY) spectra were acquired under the following conditions: 32 scans, relaxation delay of 2.0 s, and 7812.5 Hz (for white rice) and 6465.5 Hz (for brown rice) spectral widths. ^{1}H–^{13}C heteronuclear single quantum correlation (HSQC) spectra were obtained with 32 scans, 2.0 s relaxation delay, and spectral widths of 5122.9 Hz and 36,235.5 Hz in the F1 and F2 dimensions, respectively (for white rice), and 6465.5 Hz and 36,150.3 Hz in the F1 and F2 dimensions, respectively (for brown rice). Baseline correction and assignments of all ^{1}H–NMR spectra were performed using Chenomx NMR suite software (version 8.2, Chenomx, Edmonton, AB, Canada). Metabolites were further identified based on the HMDB database (http://www.hmdb.ca/) (accessed on 02 March 2022). Non-overlapping peaks were used for peak assignment. MestReNova (version 6.0.4, Mestrelab

Research, Santiago de Compostela, Spain) was employed to measure the peak *J* values and identify the peaks of the ^{1}H–^{1}H COSY and ^{1}H–^{13}C HSQC spectra.

2.5. NMR Data Pre-Processing and Measurement

Binning and normalization of the ^{1}H–NMR spectral data were performed using the Chenomx NMR suite software. Baseline-corrected NMR spectral data ranging from 0.08 to 10.00 ppm were segmented into a series of small bins (total 245) with widths of 0.04 ppm, while excluding the water suppression region (4.70–4.86 ppm). The raw NMR spectral data were normalized using total area and standardized area normalization techniques. The total area normalization method was used to compute the relative intensities of the binned spectral data by dividing the spectral data by the total area of all bins. In contrast, in standardized area normalization, the relative intensities of the binned spectral data were calculated by dividing the spectral data by the area of the reference peak. Subsequently, the results of the binned datasets were converted to Microsoft Office Excel in a suitable format to quantify each compound by its loading value, and the binning values of compounds with numerous non-overlapping peaks were summed. All the pre-processed NMR spectral data with peak values were converted into comma-separated value (CSV) files for ML analysis.

2.6. Statistical Analysis

After normalization of the NMR data, SIMCA-P+ software (version 13.0, Umetrics, Umeå, Sweden) was used to perform multivariate statistical analysis. PCA and OPLS-DA were performed using the SIMCA software. PCA is a clustering approach that minimizes the dimensions of multivariate data, while retaining the majority of its variance without any prerequisite information about the dataset, whereas the OPLS-DA model is a supervised classification method [21]. The autofit function in the SIMCA program was used to choose the number of components such that a significant number of principal components were selected from the models.

Mean-centering was performed, unit variance (UV) and Pareto (Par) scaling were applied in both PCA and OPLS-DA, and the outcomes were compared to determine the best scaling method. The goodness-of-fit and predictability of the model were evaluated using R^2Y and Q^2Y parameters. The R^2Y and Q^2Y values are expected to be close to 1. The 10-fold cross-validation and permutation test were performed to prevent the overfitting of the model. Intercept values of R^2Y and Q^2Y below 0.4 and 0.05, respectively, were regarded as valid models.

2.7. Development of Differentiation Models and ML Algorithms

Python is a scripting language widely used in data science [22]. Differentiation models implicit in various ML algorithms were employed using the SciKit-Learn 0.24 software package. The SciKit-Learn library is a Python module that makes ML accessible to everyone and covers various supervised and unsupervised ML algorithms [23,24]. In metabolomics research, different linear and nonlinear supervised ML methods can be employed, such as OPLS-DA, logistic regression, SVM, k-nearest neighbors, decision tree, and RF. However, OPLS-DA is accepted as the gold standard among supervised algorithms because it supplies information related to contributing metabolites (variables) for group separation.

"GridSearchCV" is an algorithm in the SciKit-Learn library that selects optimal hyperparameters for each ML algorithm to identify the best differentiation model. A range of hyperparameter values can be assigned to the algorithm as inputs. The algorithm then builds models using each possible hyperparameter set from the ranges of the hyperparameters and shows the best hyperparameter settings for the selected ML algorithm. It also uses a CV method to find optimal hyperparameter values over k-fold CV [25].

Leave-one-out cross-validation (LOOCV) was used to evaluate the performance of machine learning algorithms. Most often used cross-validation techniques are k-fold and

LOOCV. For larger datasets, k-fold is preferable to LOOCV. Data are split into K sets for k-fold cross-validation, with one set serving as the validation set for each iteration. In comparison, LOOCV is a special case of k-fold that employs test and training data from each sample in the dataset. LOOCV chooses one sample from the data as a validation set so that each sample can reflect the test data. However, utilizing several trained and testing models by LOOCV estimates more reliable outcomes, thus it is suitable for small datasets [26,27].

In ML algorithms, true positives (TP) are the positive classes that the model correctly classifies, and true negatives (TN) are the negative classes that are classified correctly by the model. False positives (FP) are the classes that the model incorrectly classifies as positive, and false negatives (FN) are the classes that are incorrectly classified as negative by the model [24].

To evaluate the ML algorithms, six evaluators compared the performance of the established models, including accuracy, receiver operating characteristic (ROC)–area under the curve (AUC), specificity, precision, recall, and F1_score. Accuracy measures the ratio of correctly predicted samples to the total number of samples evaluated ((TP + TN)/(TP + FP + TN + FN)) [28]. Specificity measures the fraction of negative patterns that are correctly classified (TN/(TN + FP)) [24]. Precision measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class (TP/(TP + FP)) [28]. Recall measures the fraction of positive patterns that are correctly classified (TP/(TP + FN)) [24]. The F1_score is the harmonic mean of precision and recall ((2 × precision × recall)/(precision + recall)) [29]. The AUC is widely used to determine the predictability of an established model; a high AUC value represents the best performance of the model [24].

3. Results and Discussion

3.1. Identification of Metabolites in Rice

The putatively assigned peaks for white and brown rice are presented in Table 1, and the representative NMR spectrum for the metabolite extract is shown in Figure 1. We obtained 105 and 87 ^{1}H-NMR spectra using three experimental replicates from white rice (Korea, 30 and China, 5) and brown rice (Korea, 14 and China, 15) samples, respectively.

Table 1. Putative peak assignment of nuclear magnetic resonance (NMR) spectra in rice.

No.	Compound	InChI Key	Chemical Shift (Multiplicity, J Value)	Assignment Method	
				White Rice	Brown Rice
	Amino Acids				
1	4-Aminobutyrate	BTCSSZJGUNDROE-UHFFFAOYSA-N	1.84–1.92 (m), 2.29 (t, J = 7.4), 3.00 (t, J = 7.2)	1D	1D, HSQC
2	Alanine	QNAYBMKLOCPYGJ-REOHCLBHSA-N	1.47 (d, J = 7.2)	1D	1D
3	Asparagine	DCXYFEDJOCDNAF-REOHCLBHSA-N	2.80–2.92 (m), 2.88–3.00 (m)	1D	1D
4	Aspartate	CKLJMWTZIZZHCS-REOHCLBHSA-N	2.80 (dd, J = 17.4, 3.9)	1D, COSY	1D
5	Glutamate	WHUUTDBJXJRKMK-UHFFFAOYSA-N	2.00–2.10 (m)	1D, COSY	1D, COSY
6	Glutamine	ZDXPYRJPNDTMRX-VKHMYHEASA-N	2.06–2.20 (m), 2.38–2.50 (m)	1D	1D
7	Glycine	DHMQDGOQFOQNFH-UHFFFAOYSA-N	3.56 (s)	1D	1D
8	Isoleucine	AGPKZVBTJJNPAG-WHFBIAKZSA-N	0.92 (t, J = 7.2), 1.00 (d, J = 7.2)	1D, HSQC	1D, COSY, HSQC

9	Leucine	ROHFNLRQFUQHCH-YFKPBYRVSA-N	0.95 (t, J = 6.2)	1D	1D, COSY
10	Methionine	FFEARJCKVFRZRR-BYPYZUCNSA-N	2.66 (t, J = 7.8)	1D, COSY	1D, COSY
11	Threonine	AYFVYJQAPQTCCC-GBXIJSLDSA-N	1.31 (d, J = 6.6)	1D, COSY	1D, COSY
12	Valine	KZSNJWFQEVHDMF-BY-PYZUCNSA-N	0.98 (d, J = 6.9), 1.03 (d, J = 6.6)	1D	1D, COSY
	Organic acids				
13	Malate	BJEPYKJPYRNKOW-UHFFFAOYSA-N	4.33 (d, J = 7.8)	1D, COSY	1D, COSY
14	Fumarate	VZCYOOQTPOCHFL-OWOJBT-EDSA-N	6.51 (s)	1D	1D
15	Succinate	KDYFGRWQOYBRFD-UHFFFAOYSA-N	2.42 (s)	1D	1D
16	Acetate	QTBSBXVTEAMEQO-UHFFFAOYSA-M	1.91 (s)	1D	1D
17	Glycolate	AEMRFAOFKBGASW-UHFFFAOYSA-N	3.95 (s)	1D	1D
	Sugars				
18	Glucose	WQZGKKKJIJFFOK-GASJEMHNSA-N	4.63 (d, J = 7.8), 5.22 (d, J = 3.6)	1D, HSQC	1D, HSQC
19	Maltose	GUBGYTABKSRVRQ-PICCSMPSSA-N	5.41 (d, J = 3.6)	1D, COSY, HSQC	1D, COSY, HSQC
20	Sucrose	CZMRCDWAGMRECN-UG-DNZRGBSA-N	3.46 (t, J = 9.6), 3.67 (s), 3.75 (t, J = 9.6), 4.04 (t, J = 8.6), 4.21 (d, J = 8.7), 5.39 (d, J = 3.6)	1D, COSY, HSQC	1D, COSY, HSQC
	Alcohol				
21	Ethanol	LFQSCWFLJHTTHZ-UHFFFAOYSA-N	1.17 (t, J = 7.2)	1D	1D
	Others				
22	Pyruvate	LCTONWCANYUPML-UHFFFAOYSA-N	2.36 (s)	1D	1D
23	Threonate	JPIJQSOTBSSVTP-STHAYSLISA-N	4.00 (d, J = 2.4)	1D, HSQC	1D, HSQC
24	Choline	OEYIOHPDSNJKLS-UHFFFAOYSA-N	3.19 (s)	1D, HSQC	1D, HSQC

s, singlet; d, doublet; dd, doublet of doublets; t, triplet; q, quartet; m, multiplet; 1D, 1-dimensional; COSY, correlation spectroscopy; HSQC, heteronuclear single quantum correlation.

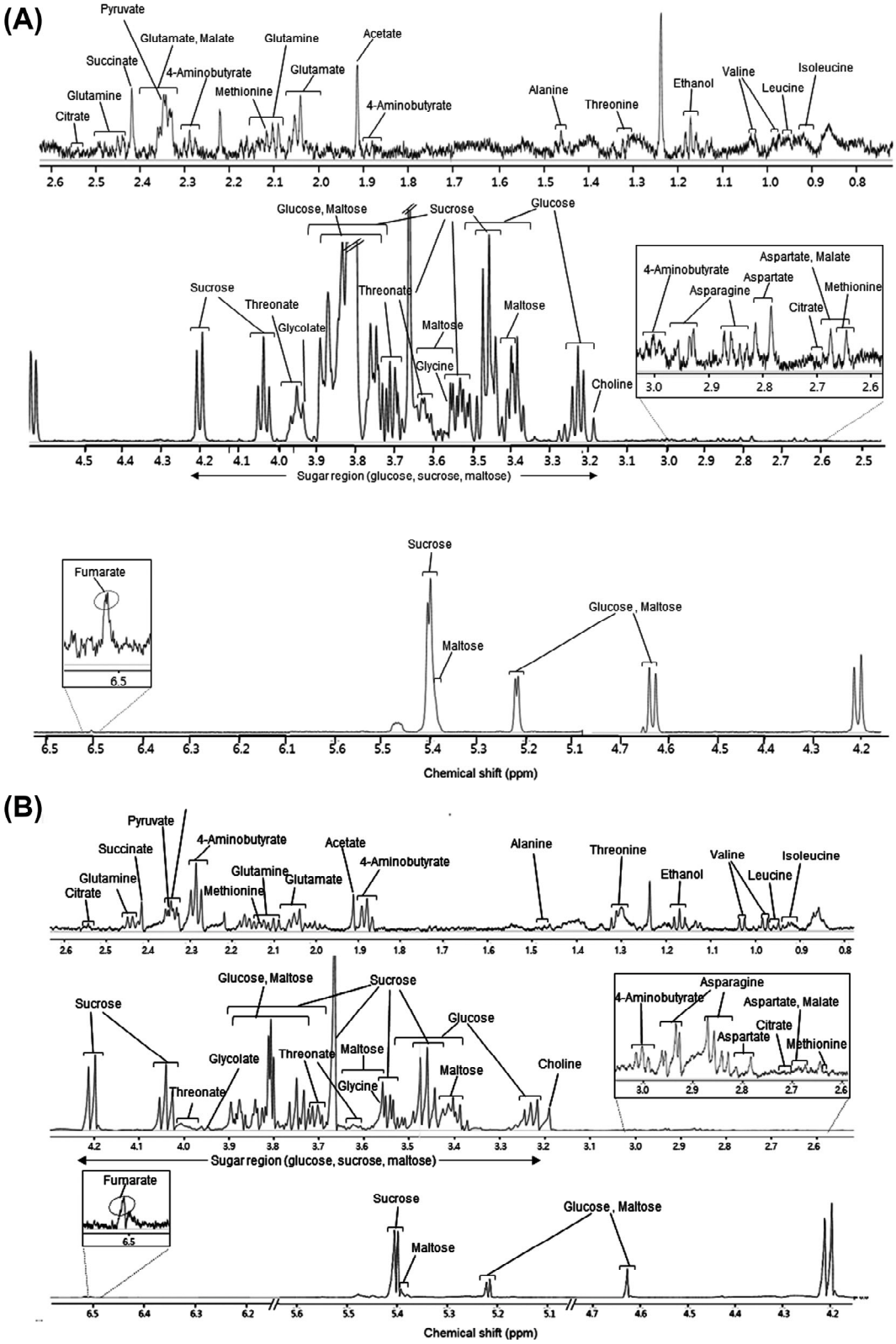

Figure 1. Representative one-dimensional ^{1}H-nuclear magnetic resonance (NMR) spectra of white rice (**A**) and brown rice (**B**).

Twenty-four metabolites, including 12 amino acids, 5 organic acids, 3 sugars, 1 alcohol, and 3 others, were identified in the rice samples using a one-dimensional NMR technique. Among the 12 amino acids, isoleucine, leucine, methionine, threonine, and valine were identified as essential amino acids. Acetate, malate, fumarate, glycolate, and succinate were found to be the organic acids. Sugars found in the rice samples included glucose, maltose, and sucrose. Two-dimensional NMR spectroscopy (COSY, HSQC) was performed to support the identification of various metabolites by one-dimensional NMR (Supplementary Figures S3 and S4).

Like other agricultural crops, metabolic profiles of rice are influenced by genotype and various environmental factors, such as rainfall, temperature, and soil [3,30]. This study considered the average rainfall and temperature in the regions from which the white and brown rice samples from Korea and China were collected. Supplementary Table S1 shows the average rainfall and temperatures of all regions. Supplementary Figure S2A,B show the average rainfall and temperature for white and brown rice collection regions in Korea and China, respectively. No significant differences in temperature were observed between Korea and China. However, a significant difference in the rainfall was observed between Korea and China. Rainfall has significant consequences for specific geographical locations and seasons. Most notably, during critical phases of crop growth, protracted durations of rainfall may drain considerable amounts of essential substances from plants, such as amino acids, organic acids, and polysaccharides [31]. Rainfall and solar radiation have been reported to significantly affect tea production and quality [32]. In addition, in most plants, a change in one factor can affect the metabolite content, even when other factors remain constant [33]. Therefore, we speculate that rainfall was the main environmental factor responsible for the differences in the metabolites of rice samples.

3.2. PCA Model Establishment for Predicting the Geographical Origin of Rice

In this study, PCA was used to objectively analyze the ^{1}H-NMR data. In the PCA score plot (Figure 2), the white and brown rice samples from Korea and China were distinctly separated by partial conjoining. The UV and Par scaling methods for white and brown rice, respectively, showed better clustering of 10 quality control (QC) samples, demonstrating the instrumental stability and reliability of NMR spectroscopy. For white rice, the principal components (PC1 and PC2) collectively accounted for 55.8% of the total variation, with $R^2X = 0.975$ and $Q^2 = 0.607$. For brown rice, the principal components (PC1 and PC2) collectively accounted for 68.7% of the total variation, with values of $R^2X = 0.984$, and $Q^2 = 0.692$.

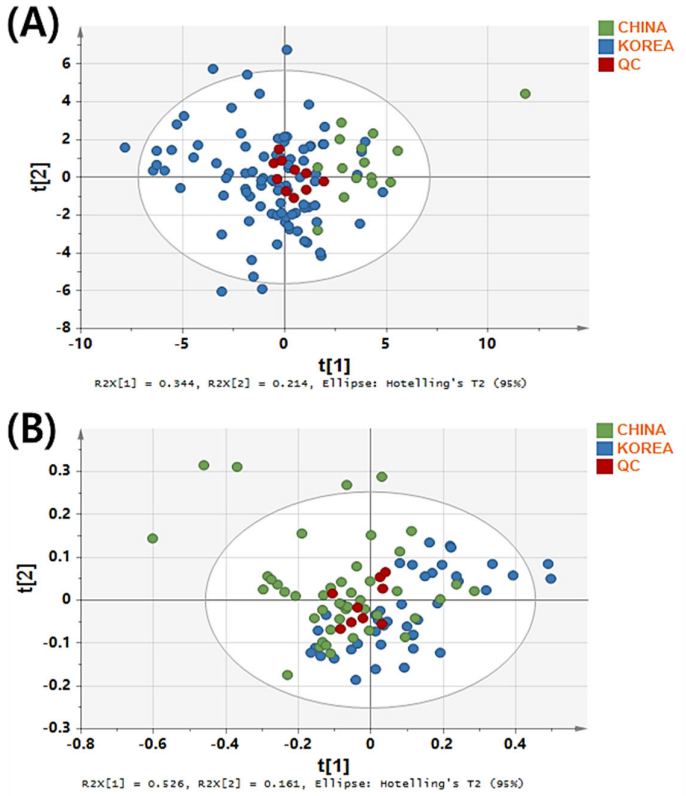

Figure 2. Principal component analysis (PCA) score plots for discriminating the geographical origin of white rice (**A**) and brown rice (**B**) samples from Korea and China.

3.3. Comparing ML Models for Predicting the Geographical Origin of Rice

Table 2 lists the model performance (R^2Y and Q^2Y) and parameters (intercept values of R^2Y and Q^2Y) of permutation for predicting Korean and Chinese white and brown rice samples with different normalization and scaling methods. For white rice samples, the model established with total area normalization and UV scaling showed the highest R^2Y and Q^2Y values, 0.673 and 0.566, respectively. Therefore, we selected this model as the optimal model. The permutation test was also satisfied with R^2Y and Q^2Y intercept values of 0.0731 and −0.196, respectively. OPLS-DA-derived score plots showed an explicit separation between the Korean and Chinese rice samples (Figure 3A).

Table 2. Orthogonal partial least square–discriminant analysis (OPLS-DA) model parameters based on various normalization and scaling methods for discriminating the geographical origin of rice samples.

Group No.	Normalization Method	Scaling Method	Component Number	R^2Y	Q^2Y	R^2Y Intercept	Q^2Y Intercept
			White Rice				
1	Total area	UV	1 + 3 + 0	0.673	0.566	0.0731	−0.196
2		Par	1 + 3 + 0	0.623	0.538	0.0941	−0.244
3	Standardized area	UV	1 + 1 + 0	0.396	0.233	0.0276	−0.292
4		Par	-	-	-	-	-

			Brown rice				
1	Total area	UV	1 + 7 + 0	0.844	0.736	0.172	−0.403
2		Par	1 + 4 + 0	0.702	0.597	0.119	−0.275
3	Standardized area	UV	1 + 6 + 0	0.827	0.723	0.144	−0.386
4		Par	1 + 7 + 0	0.82	0.702	0.152	−0.399

UV, unit variance; Par, pareto.

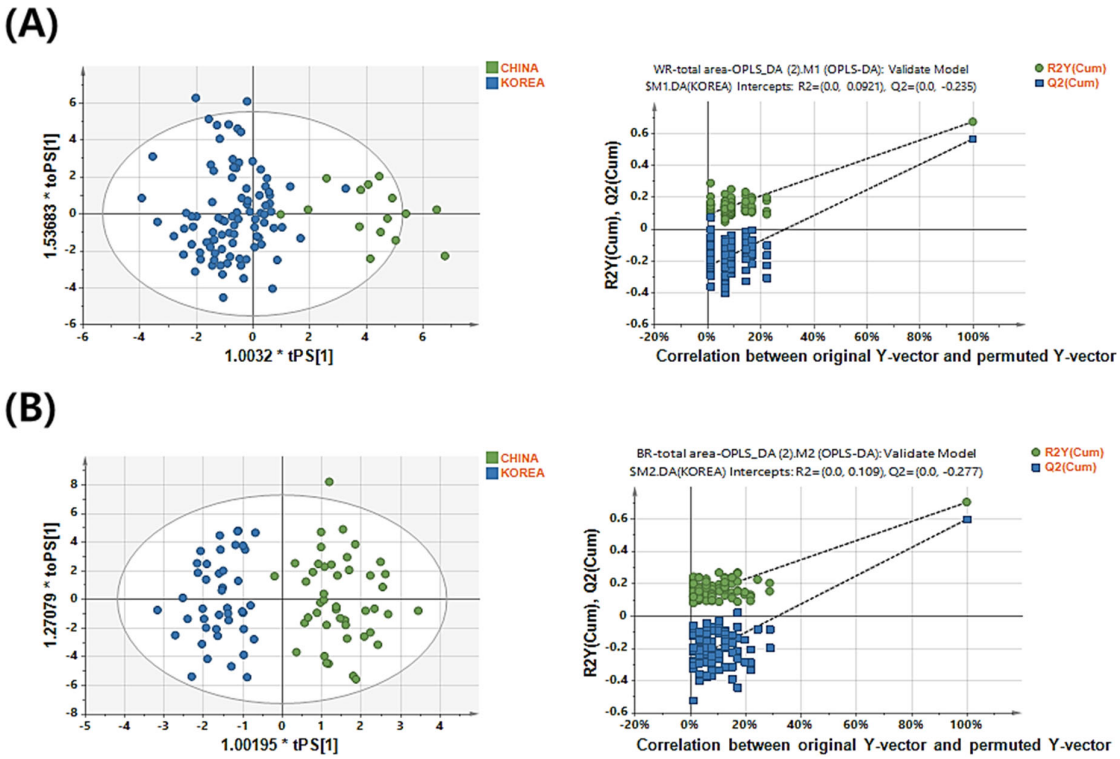

Figure 3. Orthogonal partial least square–discriminant analysis (OPLS-DA) score plots and permutation test plots for discriminating the geographical origin of white rice (**A**) and brown rice (**B**) samples from Korea and China.

For brown rice samples, the total area normalization and Par scaling methods were optimal with satisfactory R^2Y and Q^2Y values of 0.702 and 0.597, respectively (Table 2). The score plots showed an explicit separation between the Korean and Chinese rice samples (Figure 3B). The permutation test also yielded R^2Y and Q^2Y intercept values of 0.119 and −0.275, respectively (Figure 3B, Table 2). The OPLS-DA-derived score plots also showed a clear separation between the Korean and Chinese rice samples (Figure 3B). The model with total area normalization and UV scaling showed the highest value closest to 1 for brown rice; however, it showed scattered plots of QC samples in the PCA-derived score plots. The clustering of QC samples represents the robustness and reproducibility of the analysis [34,35]. Therefore, total area normalization and Par scaling were selected for discrimination of the Korean and Chinese brown rice samples because they showed better clustering of the QC samples than that with the other methods.

To distinguish between Chinese and Korean rice, the total area normalization method, which divided each metabolite peak area by the total peak area, was used, giving each sample the same total peak area of 1. The total area normalization method is one of the most widely used normalization techniques for NMR data in metabolomics research.

Consequently, each peak intensity can be reported as a percentage of the total peak intensity, making it possible to compare metabolite levels across samples in the same unit [36,37]. By assigning equal values to each metabolite and setting the standard deviation to one for all metabolites, the UV scaling strategy is one of the simplest ways to normalize metabolic variability [37–39]. Appropriate normalization and scaling techniques are crucial for improving the biological information in metabolomics data. These techniques decrease unwanted biases induced by biological and technical variance and compensate for different ranges between samples or variables for comparison [37,40,41]. Normalization considerably decreases metabolite intensity variance between samples (sample-to-sample variance), allowing all samples to be compared. Scaling balances intensity variation between metabolites (metabolite-to-metabolite comparison), allowing all metabolites to be compared [36,37].

Table 3 presents the comparison of the performance of the differentiation models established by various ML algorithms after LOOCV. The SVM-based differentiation model outperformed the RF-, decision tree-, k-nearest-neighbors-, and OPLS-DA-based differentiation models. The mentioned parameters were selected using "GridSearchCV". SVM is prone to overfitting; however, the correct type of kernel selection makes it robust to noise and overfitting [14,42]. The SVM can handle outliers efficiently because it uses a maximum margin solution. The maximum margin solution uses a maximum margin separating hyperplane for optimization [43]. SVM has a significant advantage over PLS and OPLS because the SVM model can be built using both linear and nonlinear kernels [42].

Table 3. Comparison of leave-one-out cross-validation (LOOCV) performance of various machine learning algorithms for discriminating the geographical origin of white and brown rice from Korea and China.

White Rice	**Parameters**	**Accuracy**		**ROC-AUC**		**Specificity**		**Precision**		**Recall**		**F1_Score**	
Methods		**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Random forest	·criterion = 'gini' max_depth = 4, min_samples_leaf = 2, min_samples_split = 20, random state = 0 n_estimators = 10	0.94	0.92	0.83	0.78	0.99	0.98	0.94	0.92	0.94	0.92	0.94	0.92
Decision tree	·criterion = 'gini' max_depth = 2, random state=0	0.99	0.91	0.95	0.81	0.99	0.96	0.99	0.91	0.99	0.91	0.99	0.91
SVM	·C = 3, gamma = 0.01, kernel = 'linear'	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99
Logistic regression	·C = 2, max_iter = 100,' random_state = 0, solver = 'lbfgs	1.00	0.96	1.00	0.89	1.00	0.99	1.00	0.96	1.00	0.96	1.00	0.96
KNN	·n_neighbors = 2, weights = 'distance'	1.00	0.97	1.00	0.96	1.00	0.98	1.00	0.97	1.00	0.97	1.00	1.97
OPLS-DA	·components = 1 + 3 + 0	0.96	0.98	0.98	0.98	0.99	0.98	0.95	0.96	0.89	0.89	0.92	0.91
Brown rice	**Parameters**	**Accuracy**		**ROC-AUC**		**Specificity**		**Precision**		**Recall**		**F1_score**	
Methods		**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Random forest	·criterion = 'entropy' max_depth = 3, min_samples_leaf = 2, min_samples_split = 10, random state = 0	0.99	0.92	0.98	0.92	1.00	0.95	0.99	0.92	0.99	0.92	0.99	0.92

Decision tree	n_estimators = 20 ·criterion = 'gini' max_depth = 2, random state = 0	0.98	0.94	0.99	0.94	0.98	0.95	0.98	0.94	0.98	0.94	0.98	0.94
SVM	·C = 250, kernel = 'linear'	1.00	0.96	1.00	0.96	1.00	0.96	1.00	0.95	1.00	0.95	1.00	0.95
Logistic regression	·C = 2, max_iter = 10,' random_state = 0, solver = 'liblinear'	0.83	0.78	0.82	0.78	0.82	0.78	0.83	0.78	0.83	0.78	0.82	0.78
KNN	·n_neighbors = 6, weights = 'distance'	1.00	0.91	1.00	0.92	1.00	0.93	1.00	0.91	1.00	0.91	1.00	0.91
OPLS-DA	·components = 1 + 4 + 0	0.98	0.95	1.00	0.96	0.99	0.98	0.99	0.98	0.97	0.94	0.98	0.96

SVM, LR, RF, KNN, and DT were performed using Scikit-Learn software, and the parameters were selected by the "GridSearchCV" function in SciKit-Learn. OPLS-DA was performed using SIMCA software, and the parameters were selected by the "Autofit" function in SIMCA software. DT, decision tree; KNN, k-nearest neighbors; LR, logistic regression; OPLS-DA, orthogonal partial least squares–discriminant analysis; RF, random forest; SVM, support vector machine.

We employed datasets obtained from the ^{1}H-NMR analysis, which are not high-dimensional because the number of features is less than the number of samples. High-dimensional data correspond to data with more features than the number of samples [44]. As discussed previously, the choice of kernel is essential. Therefore, "GridSearchCV" was used with different kernels and parameter sets to find the best kernel type to distinguish the rice samples. In our case, the SVM model with a linear kernel showed better train and test accuracy than the radial basis function and polynomial and sigmoid kernels. Kernel type may differ for different ^{1}H-NMR data for other crops.

The linear SVM classified rice samples using a single line or hyperplane. A line or hyperplane is adjusted by updating weights or intercept values during training. The SVM finds the best weights and intercept values that create a hyperplane or decision boundary [45,46]. The closest samples to the decision boundary were identified as support vectors, which draw lines parallel to the decision boundary to provide an optimized solution, called the maximum margin solution [47,48].

Furthermore, SVM can perform better than other standard classification algorithms. Logistic regression is also a linear classifier that can classify data by a hyperplane; however, the activation function makes it different from SVM because logistic regression uses a sigmoid function instead of a maximum margin solution [49], which does not give an optimal solution. However, the logistic regression algorithm requires a large sample size for better and stable model training and shows poor performance with irrelevant and highly correlated data. The decision tree algorithm uses Gini or information gain for building the tree. The bias of the decision tree is to find the smallest tree that can classify the data. If data change slightly or are noisy, the outcomes can vary considerably.

Moreover, decision tree algorithm cannot deal with high-dimensional data and can easily be overfitted to training data [50]. Random forest algorithms ensemble many trees during training [51], slowing down the algorithms by increasing the number of trees. Moreover, the predictions of the trees need to be uncorrelated [50]. KNN is time-consuming for large datasets and requires data scaling because it uses distance for finding neighbors, is sensitive to outliers [49], cannot handle missing values, and does not work well for imbalanced datasets [49,50].

In comparison, SVM can perform better than other standard classification algorithms for imbalanced datasets [52]. The imbalance dataset has significantly more samples than other classes, which is a cause of model overfitting. However, for our rice dataset, SVM's optimal nature showed better performance than other standard classifiers because it can handle both balanced (brown rice dataset; Korea, 14 and China, 15) and imbalanced datasets (white rice dataset; Korea, 30 and China, 5).

UV-scaled data were used for white rice to establish the model, and the best differentiation model was developed by applying SVM (accuracy and ROC-AUC of 0.99 in the test set). Par-scaled data were used for brown rice to establish the model, and the best differentiation model was developed by applying an SVM (accuracy and ROC-AUC of 0.96 in the test set). AUC values of 0.99 (for white rice) and 0.96 (for brown rice) were determined through the ROC curve analysis for predicting the geographical origin of rice, which suggested that the discovered 24 metabolites might be utilized to distinguish between rice samples from Korea and China. Thus, the SVM model developed in this study can be used to distinguish and predict Korean and Chinese rice samples.

In total, 105 and 87 rice sample spectra were analyzed using NMR. SVM showed better performance than other ML algorithms, as it offers great generalization ability for a small sample size [53,54]. Generalization is a ML term, which means that the model should be able to make appropriate decisions for unseen data based on previously observed data [55]. Hou et al. [53] identified 11 types of edible oil (from 52 samples) by employing an SVM based on a low-field nuclear magnetic resonance dataset comprising five extracted features.

Table 4 presents the performance of the SVM model in discriminating between white and brown rice using external validation. When establishing a differentiation model, internal validation is essential; however, external validation is also suggested to acquire important information regarding the existing or previously developed performance of the model [56]. External validation was performed using the previously selected SVM method by importing validation samples. For white rice, the entire dataset (n = 105) was randomly divided into development (n = 90) and validation (n = 15) samples. For brown rice, the entire dataset (n = 87) was randomly divided into development (n = 74) and validation (n = 13) samples. The performance scores (white and brown rice) were 0.96 or higher for the development set, whereas they were 1.0 for the validation set.

Table 4. Prediction performance of the support vector machine (SVM)-based machine learning model to discriminate the geographical origin of rice from the development and external validation datasets.

Evaluators	White Rice (SVM)			Brown Rice (SVM)		
	Developmental Model (n = 90)		Validation Model (n = 15)	Developmental Model (n = 74)		Validation Model (n = 13)
	Train	Test		Train	Test	
Accuracy	1.00	0.97	1.00	1.00	0.96	1.00
ROC-AUC	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	1.00	0.96	1.00	1.00	1.00	1.00
Precision	1.00	0.97	1.00	1.00	0.96	1.00
Recall	1.00	0.97	1.00	1.00	0.96	1.00
F1_score	1.00	0.97	1.00	1.00	0.96	1.00

Compared with mass spectrometry (MS)-based metabolic profiling, NMR-based metabolic profiling has advantages in rapid sample preparation and higher reproducibility [57]. However, it has lower sensitivity than MS-based metabolic profiling. Thus, it is suggested that both MS- and NMR-based metabolic profiling be employed as complementary methods for the discrimination of various crops including rice. In a previous report, MS-based metabolic profiling coupled with PLS-DA and random forest models discriminated the geographical origin of white rice samples [12]. In our study, we established optimal normalization and scaling methods for NMR datasets to differentiate white and brown rice samples from Korea and China. We also found that the SVM applied to NMR-based metabolic profiles outperforms the PLS-DA and random forest in predicting the

geographical origins of white and brown rice from Korea and China. In particular, differentiation of geographical origins of brown rice was conducted for the first time in our study using NMR-based metabolic profiling.

The limitation of machine learning techniques compared with the widely used multivariate statistical analyses, such as PLS-DA or OPLS-DA, is the lack of information about the contributing factors (metabolites) for the differentiation of each group. The machine learning techniques should be employed when the main aim is the practical differentiation or prediction, rather than revelation of contributing factors.

For practical use of established methods in this study, the expensive cost of the NMR equipment and its maintenance should be considered, especially in developing countries. Establishment and effective management of a nationwide centralized laboratory system can be a promising approach for the high-cost problem. In future studies, an extensive sample collection and analysis could be performed to establish a robust differentiation model for discriminating rice samples from various countries worldwide.

4. Conclusions

This is the first study to discriminate the geographical origin of rice from Korea and China with two milling types (white and brown) using NMR spectroscopy coupled with the most widely used ML algorithms. The SVM-based classification showed the best results in the LOOCV and external validation of the white and brown rice samples. This study can be employed as a complementary and alternative approach to previously reported analytical techniques for the geographical discrimination of rice samples. The concept and results of this study could be used for establishing a robust model for differentiation of rice samples from various countries worldwide in future studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo12111012/s1, Figure S1. Map showing the origin of Korean (A) and Chinese (B) rice samples used in the experiment; Figure S2. Climate data for white rice (A) and brown rice (B) samples from Korea and China; Figure S3. Two-dimensional NMR spectra of white rice samples. (A) ^{1}H-^{1}H COSY spectrum and (B) ^{1}H-^{13}C HSQC spectrum; Figure S4. Two-dimensional NMR spectra of brown rice samples. (A) ^{1}H-^{1}H COSY spectrum and (B) ^{1}H-^{13}C HSQC spectrum; Table S1. The provinces, cities, and weather information (year: 2018) of rice samples from the Republic of Korea and China.

Author Contributions: Conceptualization: H.-K.C., S.-Y.K., J.-A.S., Y.-S.K. and D.Y.L.; methodology, M.S. and S.-Y.K.; software, S.F.A.Z. and M.S.; formal analysis, M.S. and S.F.A.Z.; resources, H.-K.C.; data curation, M.S., S.-Y.K., J.E.R. and J.K.; writing—original draft preparation, M.S., S.-Y.K. and J.-S.K.; writing—review and editing, H.-K.C., M.S., J.-S.K., S.F.A.Z., J.E.R. and J.K.; supervision, H.-K.C.; project administration, H.-K.C., J.-A.S., Y.-S.K. and D.Y.L.; funding acquisition, H.-K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by SRC project (grant number 2022R1A5A6000760) and Chung-ang University Young Scientist Scholarship (CAYSS) in 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AUC	area under the curve
COSY	correlation spectroscopy
HMDB	Human Metabolome Database
HSQC	heteronuclear single quantum correlation
NMR	nuclear magnetic resonance

OPLS-DA	orthogonal partial least square–discriminant analysis
Par	Pareto
PCA	principal component analysis
PLS-DA	partial least square–discriminant analysis
RF	random forest
ROC	receiver operating characteristic
SVM	support vector machine
UV	unit variance

References

1. Cheajesadagul, P.; Arnaudguilhem, C.; Shiowatana, J.; Siripinyanond, A.; Szpunar, J. Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. *Food Chem.* **2013.** *141*, 3504–3509. https://doi.org/10.1016/j.foodchem.2013.06.060.
2. Song, E.-H.; Kim, H.-J.; Jeong, J.; Chung, H.-J.; Kim, H.-Y.; Bang, E.; Hong, Y.-S. A ^{1}H HR-MAS NMR-based metabolomic study for metabolic characterization of rice grain from various *Oryza Sativa* L. cultivars. *J. Agric. Food Chem.* **2016**, *64*, 3009–3016. https://doi.org/10.1021/acs.jafc.5b05667.
3. Kang, Y.; Lee, B.M.; Lee, E.M.; Kim, C.-H.; Seo, J.-A.; Choi, H.-K.; Kim, Y.-S.; Lee, D.Y. Unique metabolic profiles of Korean rice according to polishing degree, variety, and geo-environmental factors. *Foods* **2021**, *10*, 711. https://doi.org/10.3390/FOODS10040711.
4. Śliwińska-Bartel, M.; Burns, D.T.; Elliott, C. Rice fraud a global problem: A review of analytical tools to detect species, country of origin and adulterations. *Trends Food Sci. Technol.* **2021**, *116*, 36–46. https://doi.org/10.1016/J.TIFS.2021.06.042.
5. Yang, S.O.; Lee, S.W.; Kim, Y.O.; Lee, S.W.; Kim, N.H.; Choi, H.K.; Jung, J.Y.; Lee, D.H.; Shin, Y.S. Comparative analysis of metabolites in roots of *Panax Ginseng* obtained from different sowing methods. *Korean J. Med. Crop Sci.* **2014**, *22*, 17–22. https://doi.org/10.7783/KJMCS.2014.22.1.17.
6. Lee, B.-J.; Zhou, Y.; Lee, J.S.; Shin, B.K.; Seo, J.A.; Lee, D.; Kim, Y.S.; Choi, H.K. Discrimination and prediction of the origin of Chinese and Korean soybeans using fourier transform infrared spectrometry (FT-IR) with multivariate statistical analysis. *PLoS ONE* **2018**, *13*, e0196315. https://doi.org/10.1371/JOURNAL.PONE.0196315.
7. D'Urso, G.; Montoro, P.; Lai, C.; Piacente, S.; Sarais, G. LC-ESI/LTQOrbitrap/MS based metabolomics in analysis of *Myrtus Communis* leaves from Sardinia (Italy). *Ind. Crops Prod.* **2019**, *128*, 354–362. https://doi.org/10.1016/J.INDCROP.2018.11.022.
8. Dunn, W.B.; Ellis, D.I. Metabolomics: Current analytical platforms and methodologies. *TrAC Trends Anal. Chem.* **2005**, *24*, 285–294. https://doi.org/10.1016/j.trac.2004.11.021.
9. Promchan, J.; Günther, D.; Siripinyanond, A.; Shiowatana, J. Elemental imaging and classifying rice grains by using laser ablation inductively coupled plasma mass spectrometry and linear discriminant analysis. *J. Cereal Sci.* **2016**, *71*, 198–203. https://doi.org/10.1016/J.JCS.2016.08.017.
10. Huo, Y.; Kamal, G.M.; Wang, J.; Liu, H.; Zhang, G.; Hu, Z.; Anwar, F.; Du, H. ^{1}H NMR-based metabolomics for discrimination of rice from different geographical origins of China. *J. Cereal Sci.* **2017**, *76*, 243–252. https://doi.org/10.1016/J.JCS.2017.07.002.
11. Monakhova, Y.B.; Rutledge, D.N.; Roßmann, A.; Waiblinger, H.-U.; Mahler, M.; Ilse, M.; Kuballa, T.; Lachenmeier, D.W. Determination of rice type by ^{1}H NMR spectroscopy in combination with different chemometric tools. *J. Chemom.* **2013**, *28*, 83–92. https://doi.org/10.1002/cem.2576.
12. Lim, D.K.; Mo, C.; Lee, J.H.; Long, N.P.; Dong, Z.; Li, J.; Lim, J.; Kwon, S.W. The integration of multi-platform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of *Oryza Sativa* L. *J. Food Drug Anal.* **2018**, *26*, 769–777. https://doi.org/10.1016/J.JFDA.2017.09.004.
13. Kwon, Y.-K.; Bong, Y.-S.; Lee, K.-S.; Hwang, G.-S. An integrated analysis for determining the geographical origin of medicinal herbs using ICP-AES/ICP-MS and ^{1}H NMR analysis. *Food Chem.* **2014**, *161*, 168–175. https://doi.org/10.1016/J.FOODCHEM.2014.03.124.
14. Klare, J.; Rurik, M.; Rottmann, E.; Bollen, A.; Kohlbacher, O.; Fischer, M.; Hackl, T. Determination of the geographical origin of *Asparagus Officinalis* L. by ^{1}H NMR spectroscopy. *J. Agric. Food Chem.* **2020**, *68*, 14353–14363. https://doi.org/10.1021/ACS.JAFC.0C05642/SUPPL_FILE/JF0C05642_SI_001.PDF.
15. Li, Y.; Wang, X.; Li, C.; Huang, W.; Gu, K.; Wang, Y.; Yang, B.; Li, Y. Exploration of chemical markers using a metabolomics strategy and machine learning to study the different origins of *Ixeris Denticulata* (Houtt.) Stebb. *Food Chem.* **2020**, *330*, 127232. https://doi.org/10.1016/J.FOODCHEM.2020.127232.
16. Larios, G.; Nicolodelli, G.; Ribeiro, M.; Canassa, T.; Reis, A.R.; Oliveira, S.L.; Alves, C.Z.; Marangoni, B.S.; Cena, C. Soybean seed vigor discrimination by using infrared spectroscopy and machine learning algorithms. *Anal. Methods* **2020**, *12*, 4303–4309. https://doi.org/10.1039/D0AY01238F.
17. KOSTAT. Available online: https://kostat.go.kr/portal/korea/index.action (accessed on 4 April 2022).
18. KATI (Korean Association of Translators & Interpreters) in Republic of Korea. Available online: https://www.kati.net/statistics/monthlyPerformanceByProduct.do (accessed on 4 April 2022).

19. Park, J. Reports of the National Assembly and Members of the National Assembly in Republic of Korea. Available online: https://nanet.go.kr/lowcontent/assamblybodo/selectAssamblyBodoDetail.do?searchSeq=99307&searchNoSeq=2019101199307 (accessed on 12 August 2022).
20. Ministry of Agriculture, Food and Rural Affairs (MAFRA) in Republic of Korea. Available online: https://www.mafra.go.kr/mafra/294/subview.do?enc=Zm5jdDF8QEB8JTJGYmJzJTJGbWFmcmElMkY2OSUyRjMxODcxMyUyRmFydGNsVmlldy5kbyUzRg%3D%3D (accessed on 12 August 2022).
21. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi-and Megavariate Data Analysis, Part 1*; Umetrics Academy: Umeå, Sweden, 2006; pp. 63–101. Available online: https://www.worldcat.org/title/multi-and-megavariate-data-analysis-part-i-basic-principles-and-applications/oclc/900729892?referer=di&ht=edition (accessed on 10 April 2022).
22. Mendez, K.M.; Pritchard, L.; Reinke, S.N.; Broadhurst, D.I. Toward collaborative open data science in metabolomics using jupyter notebooks and cloud computing. *Metabolomics* **2019**, *15*, 125. https://doi.org/10.1007/S11306-019-1588-0.
23. Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
24. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281. https://doi.org/10.1186/S12911-019-1004-8.
25. Paper, D. Scikit-learn classifier tuning from complex training sets. In *Hands-on Scikit-Learn for Machine Learning Applications*; Apress: Berkeley, CA, USA, 2020; pp. 165–188. https://doi.org/10.1007/978-1-4842-5373-1_6.
26. Wong, T.T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846. https://doi.org/10.1016/J.PATCOG.2015.03.009.
27. Cha, G.W.; Moon, H.J.; Kim, Y.C. Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8530. https://doi.org/10.3390/IJERPH18168530.
28. Sahli, H. An introduction to machine learning. In *TORUS 1-Toward an Open Resource Using Services: Cloud Computing for Environmental Data*; Wiley: Hoboken, NJ, USA, 2020; pp. 61–74. https://doi.org/10.1002/9781119720492.ch7.
29. Chicco, D.; Jurman, G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. https://doi.org/10.1186/S12864-019-6413-7.
30. Fiehn, O. Metabolomics—the link between genotypes and phenotypes. In *Functional Genomics*; Springer: Dordrecht, The Netherlands, 2002; pp. 155–171. https://doi.org/10.1007/978-94-010-0448-0_11.
31. Tukey, H.B. Implications of allelopathy in agricultural plant science. *Bot. Rev.* **1969**, *35*, 1–16. https://doi.org/10.1007/BF02859885.
32. Marx, W.; Haunschild, R.; Bornmann, L. Global warming and tea production—The bibliometric view on a newly emerging research topic. *Climate* **2017**, *5*, 46. https://doi.org/10.3390/CLI5030046.
33. Yang, L.; Wen, K.S.; Ruan, X.; Zhao, Y.X.; Wei, F.; Wang, Q. Response of plant secondary metabolites to environmental factors. *Molecules* **2018**, *23*, 762. https://doi.org/10.3390/MOLECULES23040762.
34. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249–2264. https://doi.org/10.4155/BIO.12.204.
35. Gika, H.G.; Theodoridis, G.A.; Earll, M.; Wilson, I.D. A QC approach to the determination of day-to-day reproducibility and robustness of LC–MS methods for global metabolite profiling in Metabonomics/Metabolomics. *Bioanalysis* **2012**, *4*, 2239–2247. https://doi.org/10.4155/BIO.12.212.
36. Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J.K.; Lindon, J.C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* **2006**, *78*, 2262–2267. https://doi.org/10.1021/AC0519312.
37. Zhou, Y.; Kim, S.-Y.; Lee, J.-S.; Shin, B.-K.; Seo, J.-A.; Kim, Y.-S.; Lee, D.-Y.; Choi, H.-K.; Zhou, Y.; Kim, S.-Y.; et al. Discrimination of the geographical origin of soybeans using NMR-based metabolomics. *Foods* **2021**, *10*, 435. https://doi.org/10.1016/J.JCS.2015.08.001.
38. Li, B.; Tang, J.; Yang, Q.; Cui, X.; Li, S.; Chen, S.; Cao, Q.; Xue, W.; Chen, N.; Zhu, F. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci. Rep.* **2016**, *6*, 38881. https://doi.org/10.1038/srep38881.
39. Weljie, A.M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C.M. Targeted pofiling: Quantitative analysis of ^{1}H NMR metabolomics data. *Anal. Chem.* **2006**, *78*, 4430–4442. https://doi.org/10.1021/ac060209g.
40. Kohl, S.M.; Klein, M.S.; Hochrein, J.; Oefner, P.J.; Spang, R.; Gronwald, W. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* **2012**, *8*, 146–160. https://doi.org/10.1007/S11306-011-0350-Z.
41. van den Berg, R.A.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genom.* **2006**, *7*, 142. https://doi.org/10.1186/1471-2164-7-142.
42. Vu, T.; Siemek, P.; Bhinderwala, F.; Xu, Y.; Powers, R. Evaluation of multivariate classification models for analyzing NMR metabolomics data. *J. Proteome Res.* **2019**, *18*, 3282–3294. https://doi.org/10.1021/acs.jproteome.9b00227.
43. Garcés, M.A.; Orosco, L.L. EEG signal processing in brain–computer interface. In *Smart Wheelchairs and Brain-Computer Interfaces Mobile Assistive Technologies*; Academic Press: Cambridge, MA, USA, 2008; pp. 95–110. https://doi.org/10.1016/B978-0-12-812892-3.00005-4.
44. Narisetty, N.N. Bayesian model selection for high-dimensional data. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2020; Volume 43, pp. 207–248. https://doi.org/10.1016/BS.HOST.2019.08.001.

45. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000; ISBN 0-387-98780-0.
46. Chang, Y.-W.; Lin, C.-J.; Guyon, I.; Aliferis, C.; Cooper, G.; Elisseeff, A.; Pellet, J.-P.; Spirtes, P.; Statnikov, A. Feature ranking using linear SVM. *JMLR Work. Conf. Proc.* **2008**, *3*, 53–64.
47. Temko, A.; Thomas, E.; Marnane, W.; Lightbody, G.; Boylan, G. EEG-based neonatal seizure detection with support vector machines. *Clin. Neurophysiol.* **2011**, *122*, 464–473. https://doi.org/10.1016/J.CLINPH.2010.06.034.
48. Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2001, 2**, 45–66.
49. Kotsiantis, S.B.; Zaharakis, I.D.; Pintelas, P.E. Machine Learning: A review of classification and combining techniques. *Artif. Intell. Rev.* **2006**, *26*, 159–190. https://doi.org/10.1007/s10462-007-9052-3.
50. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; IEEE: Piscataway, NJ, USA, 2016. Available online: https://ieeexplore.ieee.org/abstract/document/7724478 (accessed on 20 August 2022).
51. Qi, Y. Random forest for bioinformatics. In *Ensemble Machine Learning*; Springer: Boston, MA, USA, 2012; pp. 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11.
52. Tang, Y.; Zhang, Y.Q.; Chawla, N.V. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2009**, *39*, 281–288. https://doi.org/10.1109/TSMCB.2008.2002909.
53. Hou, X.; Wang, G.; Su, G.; Wang, X.; Nie, S. Rapid identification of edible oil species using supervised support vector machine based on low-field nuclear magnetic resonance relaxation features. *Food Chem.* **2019**, *280*, 139–145. https://doi.org/10.1016/J.FOODCHEM.2018.12.031.
54. Liu, X.; Gao, C.; Li, P. A comparative analysis of support vector machines and extreme learning machines. *Neural Netw.* **2012**, *33*, 58–66. https://doi.org/10.1016/J.NEUNET.2012.04.002.
55. Heinemann, J. Machine learning in untargeted metabolomics experiments. In *Methods in Molecular Biology*; Humana Press: New York, NY, USA, 2019; Volume 1859, pp. 287–299.
56. Moons, K.G.M.; Altman, D.G.; Reitsma, J.B.; Ioannidis, J.P.A.; Macaskill, P.; Steyerberg, E.W.; Vickers, A.J.; Ransohoff, D.F.; Collins, G.S. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73. https://doi.org/10.7326/M14-0698.
57. Garcia-Perez, I.; Posma, J.M.; Serrano-Contreras, J.I.; Boulangé, C.L.; Chan, Q.; Frost, G.; Stamler, J.; Elliott, P.; Lindon, J.C.; Holmes, E.; et al. Identifying unknown metabolites using NMR-based metabolic profiling techniques. *Nat. Protoc.* **2020**, *15*, 2538–2567. https://doi.org/10.1038/s41596-020-0343-3.