Geographical discrimination of Asian red pepper powders using ^{1}H NMR spectroscopy and deep learning-based convolution neural networks

Byung Hoon Yun [a,1], Hyo-Yeon Yu [a,1], Hyeongmin Kim [a], Sangki Myoung [a], Neulhwi Yeo [a], Jongwon Choi [b], Hyang Sook Chun [c,*], Hyeonjin Kim [d,e,*], Sangdoo Ahn [a,*]

[a] Department of Chemistry, Chung-Ang University, Seoul 06974, South Korea
[b] Department of Advanced Imaging, Chung-Ang University, Seoul 06974, South Korea
[c] Department of Food Science & Technology, Chung-Ang University, Anseong 17546, South Korea
[d] Department of Medical Sciences, Seoul National University, Seoul 03080, South Korea
[e] Department of Radiology, Seoul National University Hospital, Seoul 03080, South Korea

ARTICLE INFO

ABSTRACT

This study investigated an innovative approach to discriminate the geographical origins of Asian red pepper powders by analyzing one-dimensional ^{1}H NMR spectra through a deep learning-based convolution neural network (CNN). ^{1}H NMR spectra were collected from 300 samples originating from China, Korea, and Vietnam and used as input data. Principal component analysis − linear discriminant analysis and support vector machine models were employed for comparison. Bayesian optimization was used for hyperparameter optimization, and cross-validation was performed to prevent overfitting. As a result, all three models discriminated the origins of the test samples with over 95 % accuracy. Specifically, the CNN models achieved a 100 % accuracy rate. Gradient-weighted class activation mapping analysis verified that the CNN models recognized the origins of the samples based on variations in metabolite distributions. This research demonstrated the potential of deep learning-based classification of ^{1}H NMR spectra as an accurate and reliable approach for determining the geographical origins of various foods.

1. Introduction

The geographical origin of food refers to the country, region, or sea area where agricultural or marine products are produced, harvested, or captured. It is an issue closely related to concerns about food authenticity. One common example of geographical origin fraud is the misrepresentation of the country of origin on the label, a trend exacerbated by the globalization of the food market due to the liberalization of food trade between countries (Katerinopoulou et al., 2020). Such food fraud leads to increased costs for consumers who end up paying more for lower-quality products or unknowingly purchasing items they did not intend to buy. Moreover, it can sometimes have serious health implications, making it crucial to control and prevent such fraudulent practices. Various factors, such as climatic conditions, soil, cultivation practices, and processing methods, contribute to differences in the chemical composition of food products based on their geographic origin (Medina et al., 2019). Therefore, the chemical composition of food serves as a unique fingerprint, enabling the verification of its authenticity through appropriate analysis techniques (J. Zhang et al., 2019).

Chemical analysis techniques, including chromatography, elemental

analysis, mass spectrometry, and spectroscopy, have been used to determine the chemical composition of food for the purpose of discriminating its geographical origin (Tahir et al., 2022). While chromatography offers good sensitivity and applicability to a wide range of samples, it requires skilled personnel and involves time-consuming experimental processes. Spectroscopic techniques such as Fourier transform infrared (FT-IR) and nuclear magnetic resonance (NMR) offer advantages in determining the authenticity of food due to their non-destructive nature and minimal time requirements for data acquisition. Among these techniques, NMR spectroscopy holds particular importance in the field of food authenticity due to its ability to handle samples in various phases and provide molecular information about the system. By combining NMR spectroscopy with multivariate data analysis methods, like principal component analysis–linear discriminant analysis (PCA-LDA) and pattern recognition technology, it becomes a powerful tool for determining geographical origin by investigating complex food matrices' profiles or quantitatively analyzing specific components present in the matrices (Luykx & van Ruth, 2008). In particular, PCA-LDA is an effective method for determining the geographic origin due to its characteristic features, such as dimensionality reduction, data visualization, feature selection, and interpretability (Tahir et al., 2022).

As the volume and complexity of data obtained from analytical devices increase, there is a possibility of errors in extracting representative features from the sample. Recently, the introduction of artificial intelligence (AI) technology has enabled the extraction of useful information from larger and more complex datasets, leading to active research in the field of food geographical origin (Lia et al., 2020; Maione et al., 2019). Machine learning algorithms can identify patterns in datasets through a process of self-learning (Kotsiantis et al., 2006). In particular, support vector machine (SVM) learning performs classification by finding a hyperplane that best differentiates between classes (Noble, 2006). SVM is also a commonly used technique for determining geographical origin because of its robustness to overfitting, wide adoption, and that it can be applied to linear and nonlinear classification (Jin et al., 2023). Deep learning, a subset of machine learning, utilizes artificial neural networks (ANNs) inspired by the human brain to achieve more complex learning (Schmidhuber, 2015), and it has found extensive applications in various fields. Convolutional neural networks (CNNs) are a class of ANNs that specialize in image classification tasks, an ability attributed to their feature-learning process based on convolution and pooling (Shin et al., 2016; Chen et al., 2020). Recently, CNN models have been successfully applied to spectroscopic data in order to discriminate the variety or geographical origin of various foods (Cai et al., 2023; Feng et al., 2021; Hou et al., 2020).

Red pepper (*Capsicum annuum L.*) powder is a widely used seasoning in Asian cuisine, and fraudulent misrepresentation of its geographical origin is a prevalent issue. While the demand and price for domestic products are much higher than imported ones, the supply chain often faces challenges. Therefore, there is a need for an analysis method that can accurately and efficiently determine the geographical origin of red pepper powders. In a previous study, discriminant variables were identified through multivariate statistical analysis of ^{1}H NMR signals related to the metabolite content of Asian red pepper powders (Lee et al., 2020). A linear discriminant function that can distinguish the geographical origins of red pepper powders from Asia was derived from the discriminant variable. However, that analysis was limited by the number of samples used and had certain accuracy limitations.

The aim of this study is to develop an innovative method for accurately discriminating the geographical origin of Asian red pepper powders by analyzing their ^{1}H NMR spectral profiles using deep learning-based CNN models. To train CNN models, the one-dimensional (1D) ^{1}H NMR spectra were converted into two-dimensional (2D) images and were used as input data. The spectral range and resolution of the input image were set differently to examine their influence on discrimination. Various methods to optimize hyperparameters for effective model development were incorporated. To assess the discrimination

capabilities of the CNN models, PCA-LDA and SVM models were also introduced in the analyses. The feasibility and accuracy of the developed CNN models were validated using test samples, and their performance was compared with that of the other models.

2. Materials and methods

2.1. Materials

A total of 300 red pepper powders (100 each from China, Korea, and Vietnam) distributed in Korea were collected as samples. Korean red pepper powders were purchased from local producers and/or reliable suppliers, such as agricultural cooperatives. Chinese and Vietnamese red pepper powders imported to Korea through the Korea Agro-Fisheries Trade Corporation were purchased from local markets. All samples were stored in a refrigerator at 4 ℃.

Deuterium oxide (D$_2$O, 99.9 atom% D, containing 1.0 %, w/w sodium 4,4-dimethyl-4-silapentane-1-sulfonate (DSS)) and methanol-d_4 (≥99.8 atom% D) for extraction and ^{1}H NMR measurement were purchased from Merck (Seoul, Korea). All solvents were used without further purification.

2.2. Sample preparation

The dried-red pepper powder samples were further ground into a fine powder with particle diameters of ≤ 200 μm in a food grinder to facilitate effective metabolite extraction. The extraction procedures of metabolites from red pepper powder samples were carried out in the same manner as that described by Lee et al. (2020). For each red pepper powder, sonication of 500 mg was performed for 30 min using 1,600 μL of methanol − water mixed solvent (3:1, v/v). At this time, deuterated methanol and water were used as extraction solvents to enable direct measurement of ^{1}H NMR spectra without the need for additional treatment. The samples were then centrifuged at 13,500 rpm for 1 min, and all supernatants were filtered through a 0.45 μm pore size filter to remove particulate impurities that may have arisen during the extraction process.

2.3. ^{1}H NMR experiments

1D high-resolution ^{1}H NMR spectra were acquired on a Varian 600 MHz NMR spectrometer (Palo Alto, CA, USA) operating at an NMR frequency of 600.14 MHz. The NMR spectrometer was equipped with a 5-mm PFG dual broad-band 600NB probe controlled by the VnmrJ v3.2A software. The NMR conditions were as follows: temperature of 25 ℃, pulse angle of 45°, acquisition time of 3.408 s, and 16 scan averages with a relaxation delay of 10 s. The data of 64 k-points were recorded with a spectral width of 9,615.4 Hz. Additionally, the presaturation pulse sequence was used to suppress the water signal from any moisture present in the red pepper powder sample. The chemical shifts of metabolites were reported relative to the DSS signal at 0 ppm. The preprocessing tasks, such as phase correction and baseline correction of all NMR spectra, were carried out using Mnova NMR software from MestreNova (Mestrelab Research, Santiago de Compostela, Spain).

To assess the impact of spectral region on classification accuracy, the study conducted analyses by dividing the ^{1}H NMR spectrum of red pepper powder sample extract into three distinct regions: 0.7–8.5 ppm (*full region*), 3.1–4.1 ppm (*carbohydrate region*), and 6.5–8.5 ppm (*aromatic region*). All three spectral regions were applied to PCA-LDA, SVM, and CNN models. For the PCA-LDA and SVM approaches, spectral data were binned with a binning size of 0.01 ppm. Consequently, the full, carbohydrate, and aromatic regions had 780, 100, and 200 variables, respectively. Each variable integral was normalized to the overall integral of its respective region to minimize the effects of concentration differences among samples (Mahadevan et al., 2008). For the CNN models, each of the three NMR spectral regions was transformed into a

2D image and stored as a PNG file. Subsequently, using the OpenCV library, black-and-white inversion as well as size were adjusted. At that time, cv2.INTER_AREA was used for interpolation. Additionally, in order to evaluate the effect of resolution on discrimination, 2D images were generated with varying resolutions: 256 × 64, 256 × 256, and 1,024 × 64.

2.4. Development of discriminant methods for the geographic origin of Asian red pepper powders

2.4.1. PCA-LDA

The data analysis was performed using Python 3.8 with the scikit-learn 1.1.2 library. LDA is a supervised pattern recognition method used for classification purposes and is widely applied in the field of spectroscopy. It aims to find one or more linear functions that maximize the ratio of between-class variance and minimize the ratio of within-class variance (Varmuza & Filzmoser, 2016). However, LDA cannot be directly employed if the data exhibit collinearity or if the number of variables is larger than the number of observations (Gromski et al., 2015). To circumvent this limitation, PCA was employed to reduce data collinearity and the number of variables (Kumar et al., 2014).

In this study, the number of variables (i.e., chemical shifts) exceeded the total number of red pepper powder samples ($n = 300$), which made it challenging to apply LDA directly. Therefore, the principal components (PCs) were used as the input variables for LDA, and the number of PCs was optimized using 5-fold cross-validation. The red pepper powder dataset for each country of origin ($n = 100$) was divided into training and test sets at a ratio of 8:2 ($n = 80$ and 20, respectively), and 20 % of the training set was used for validation during the training process. This division was equally applied to the SVM and CNN models. To identify the variables that contributed significantly to the classification, the

shapley additive explanations (SHAP) approach was used with the scikit-learn 1.1.2 library module in Python 3.8 (Lundberg & Lee, 2017).

2.4.2. SVM

SVM performs data classification by setting a hyperplane that maximizes the margin between different classes. This is an efficient approach because only support vectors near each class boundary are used for training (Pisner & Schnyer, 2020). In this study, SVM was developed using Python 3.8 with the scikit-learn 1.1.2 library. As with LDA, the best SVM hyperparameters, such as kernel, gamma, and cost, were optimized using 5-fold cross-validation. SHAP was also employed to identify important variables in the classification process.

2.4.3. CNN

The CNN models were developed using Python 3.8 with the Tensorflow 2.7.0 and Keras 2.2 libraries on a computer equipped with an Intel Core-i7 10,700 CPU. As shown in Fig. 1, the CNN consists of an input layer, hidden layers, and an output layer (Albawi et al., 2017; L. Zhang et al., 2020). In this study, as mentioned in section 2.3, ^{1}H NMR spectra from three distinct regions were transformed into 2D images with varying resolutions, which served as the input data for training the CNN. The hidden layers consist of a combination of convolutional (CONV) layers, pooling layers (Max-Pool), and fully connected layers (Dense 1 and Dense 2). For the CONV layers, the number of filters was determined by Bayesian optimization using the validation set. The filter size was set to 3 × 3. A rectified linear unit (ReLU) was used as the activation function (Agarap, 2018). The max pooling was used to reduce the feature size while maintaining the characteristics of the data, and the pool size was set to 3 × 3.

The features learned elaborately through the CONV and Max-pool are flattened into a 1D array by the flatten layer, which is then
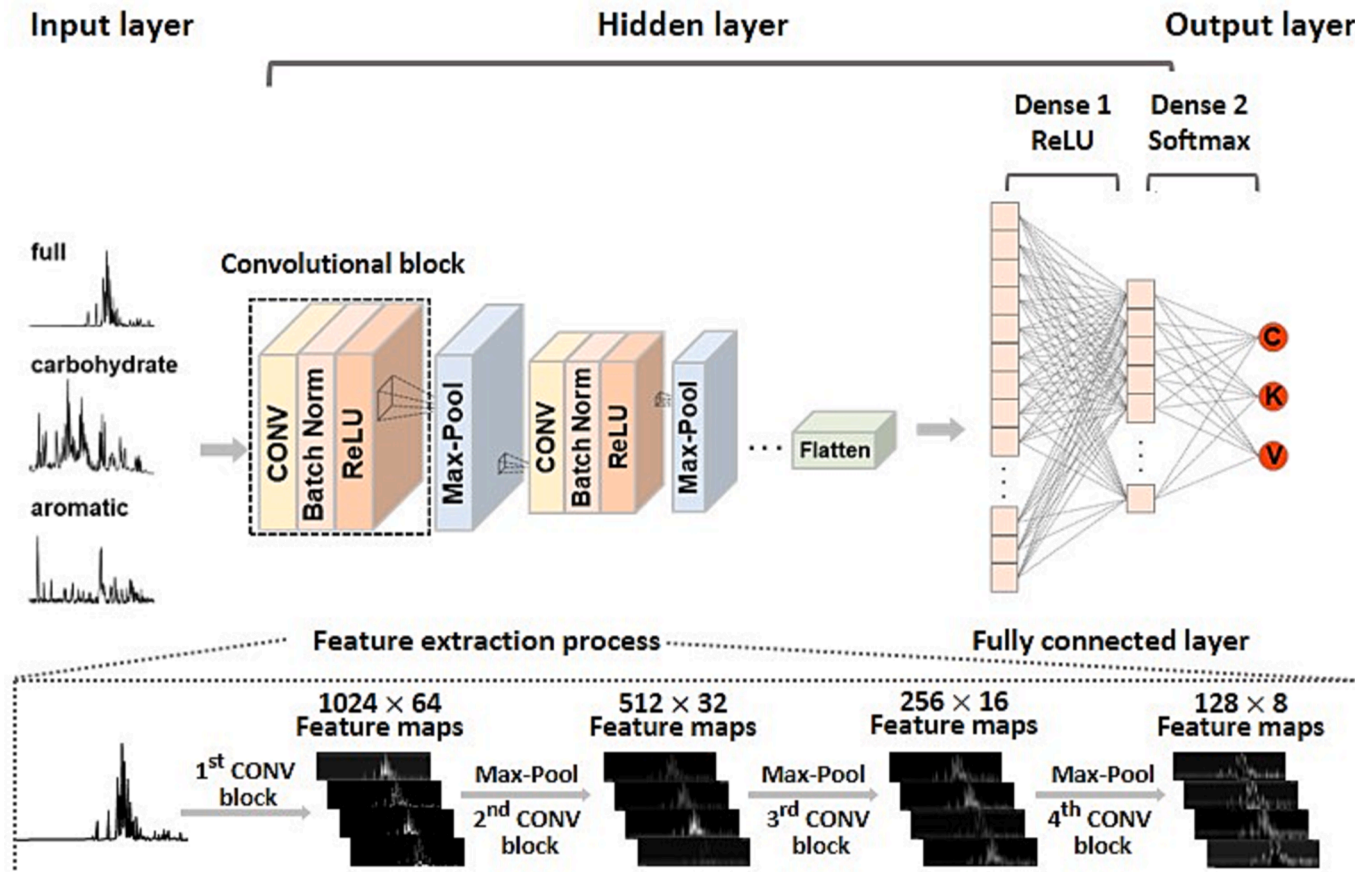

Fig. 1. The process of CNN deep learning with ^{1}H NMR spectra. (C: Chinese, K: Korean, V: Vietnamese). CNN, convolutional neural network.

received as input data to perform learning in the fully connected layers. Two fully connected layers (Dense 1 and Dense 2) were included in this CNN model, and the neurons for the first fully connected layer (Dense 1) were activated using the ReLU activation function (Yamashita et al., 2018). To discriminate the geographical origins of the red pepper powders from China, Korea, and Vietnam, the output layer was composed of three units. The final output of the CNN model was obtained following the second fully connected layer (Dense 2) and a Softmax activation function (Wang et al., 2018). The cross-entropy loss and Adam optimizer were used in the training (Kingma & Ba, 2014). The numbers of the CONV-Batch Norm-ReLU blocks and learning rate were also optimized by Bayesian optimization simultaneously with the number of filters. The detailed architecture of the CNN model for the full region is shown in Table S1.

For the sake of the transparency of the trained neural networks, gradient-weighted class activation mapping (Grad-CAM) was used to identify the signal areas with a significant influence on discriminating the geographical origin (Selvaraju et al., 2020). Grad-CAM is a technique for providing a "visual explanation" for the decisions made by deep learning models. This technique involves computing the gradients of the target class with respect to each individual input feature, which represents the influence of each pixel in the input image on the model's decision. A higher gradient value indicates a greater level of influence. These gradients are then used as weights in the generation of a final heatmap, hence the term "gradient-weighted." The application of Grad-CAM is demonstrated by highlighting specific regions of the input spectrum in reddish color. This coloration indicates that these spectral regions have a significant positive impact on determining the geographical origin of the red peppers. Conversely, the bluish regions are interpreted as being independent of the determination process (Selvaraju et al., 2020; Shin et al., 2021).

In general, preprocessing is important when using chemometrics methods because it helps remove the artifacts and noise signals caused by imperfect instruments and sample preparation. Therefore, preprocessing techniques, such as baseline correction, phase correction, and binning, are frequently employed in chemometrics. However, in the case of CNNs, the same effect as preprocessing can be achieved during the extraction of feature maps, providing the advantage of using raw data directly (Brereton et al., 2017; X. Zhang et al., 2020).

3. Results and discussion

3.1. ^{1}H NMR spectra of red pepper powders

A representative ^{1}H NMR spectrum for discrimination of geographical origin of red pepper powder sample extracted with methanol-d_4 and D_2O is shown in Fig. 2. The spectrum reveals numerous high-intensity and overlapping signals within the 3.0–4.5 ppm range, which can be attributed to the presence of various metabolites containing multiple carbohydrate hydrogens in the red pepper powder. Due to the similarity in chemical environments of most carbohydrate hydrogens, these signals appear as overlapping, which poses a challenge in distinguishing individual metabolite signals. Consequently, in the study by Lee et al. (2020), this region was excluded when determining the geographical origin of Asian red pepper powder through the statistical analysis of integral values of a limited number of well-separated specific metabolite signals in the NMR spectrum. However, in this present study, three different fingerprint ranges of the ^{1}H NMR spectrum, including this region, were evaluated to compare the effect of spectral regions on classification accuracy without assigning specific metabolite signals.

The first region, 0.7–8.5 ppm (*full*), encompasses the entire metabolite signals. By testing the method on the entire spectrum region, it is

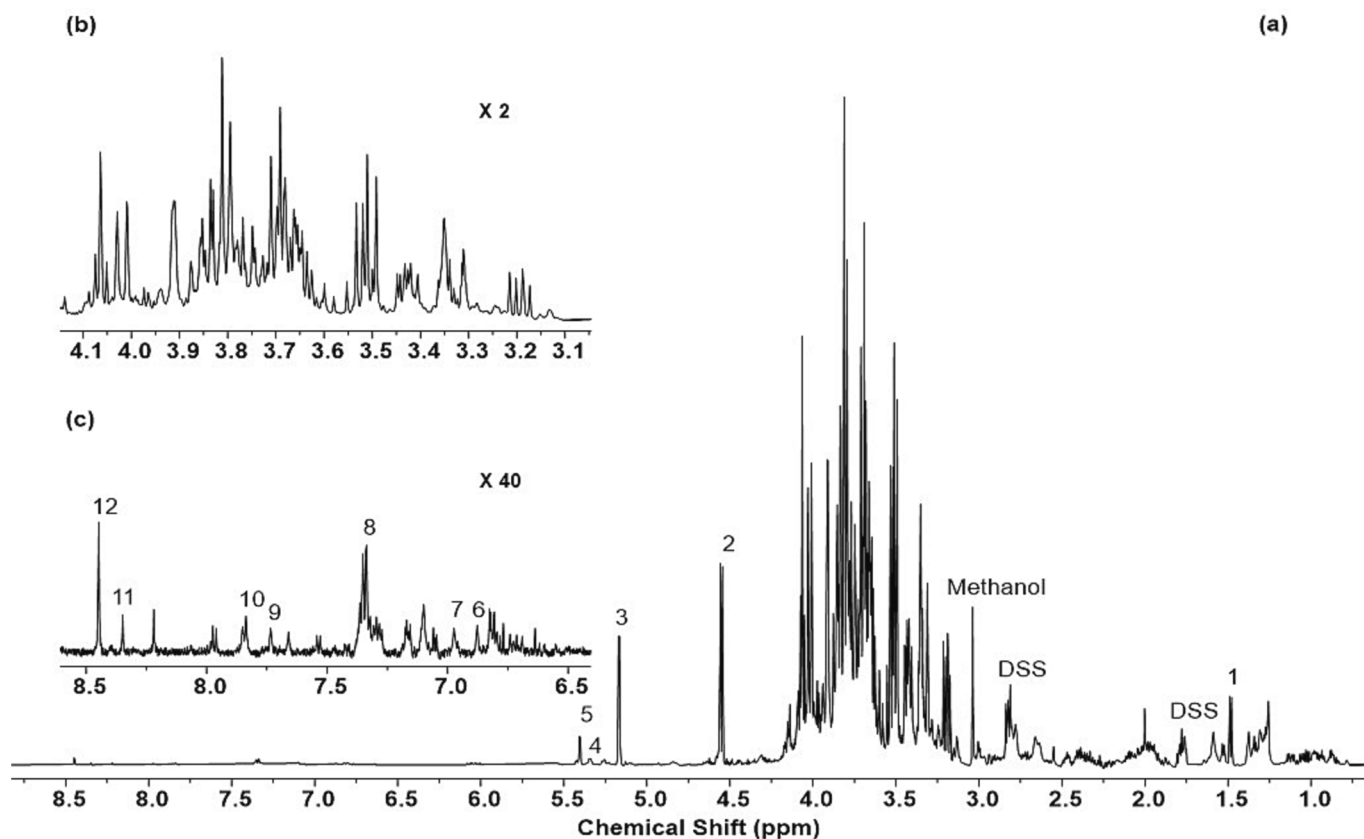**Fig. 2.** Representative ^{1}H NMR spectra of (a) full, (b) carbohydrate, and (c) aromatic regions for Korean red pepper powder. The labeled peaks are as follows (Becerra-Martínez et al., 2017); DSS: sodium 4,4-dimethyl-4-silapentane-1-sulfonate; 1: alanine; 2: β-glucose; 3: α-glucose; 4: unsaturated fatty acid; 5: sucrose; 6: tyrosine; 7: kaempferol; 8: phenylalanine; 9: tryptophan; 10: Uridine; 11: adenosine; 12: histidine.

anticipated that data preprocessing can be performed with relative ease, and results can also be obtained without the potential distortion arising from the selection of specific signals. The second region, 3.1–4.1 ppm (*carbohydrates*), contains relatively prominent metabolite signals associated with carbohydrates, such as glucose, fructose, and sucrose (Fig. 2b). As mentioned earlier, the signals are significant, but it is challenging to attribute them to a specific metabolite signal due to overlap; hence, the reason Lee et al. (2020) excluded this region from their study.

However, in the CNN model, which extracts various features from spectral images, it was expected that useful information for discrimination could be obtained from overlapping signals. Therefore, this region will be useful for comparing the performance of CNN models with the previous study (Lee et al., 2020) and other AI models. The third region, 6.5–8.5 ppm (*aromatic*), exhibits relatively low-intensity signals but contains well-resolved peaks from metabolites such as tyrosine, phenylalanine, tryptophan, uridine, adenosine, and histidine (Fig. 2c). Many of the signals used to determine geographical origin are distributed in this region (Lee et al., 2020). This region is expected to exhibit a significantly different spectral pattern based on the geographical origin despite the low-intensity signals compared to other regions, and thus, it is suitable for comparing similarities between AI and statistical model analyses.

Generally, using the full spectrum offers the advantage of incorporating a broader range of signals. However, in the case of a red pepper powder sample that may contain moisture, it is necessary to suppress the water signal using a presaturation pulse sequence, as in this study. If water suppression is impossible or incomplete, an alternative region unaffected by the water signal should be considered for analysis instead.

3.2. Discrimination of the geographical origin of Asian red pepper powders by PCA-LDA models

PCA and LDA are two commonly used techniques for reducing the dimensionality of data. The main purpose of PCA is to enable visualization or to remove noise through dimensionality reduction, which reduces variables in the data while maintaining as much variance as possible (Jolliffe, 2005). However, because each axis of maximum variance does not select an appropriate feature that distinguishes between classes well, the accuracy of classification can be improved through combination with LDA. Contrary to PCA, LDA reduces the dimension while maximizing the ratio of between-class variance to the within-class variance in any data set, thereby guaranteeing maximal separability (Balakrishnama & Ganapathiraju, 1998). In this study, PCs obtained through dimensionality reduction using PCA were used as input data for LDA. If the number of PCs is too small, some information disappears, and if the number of PCs is too large, dimensionality reduction is hardly achieved (Jolliffe, 2005). Therefore, 5-fold cross-validation and random search were used to find the appropriate number of PCs. As a result, the PCs for the full, carbohydrate, and aromatic regions were 17, 15, and 16, respectively.

As summarized in Table 1, the accuracies of the training set of the

Table 1
Accuracies of training and testing of PCA-LDA, SVM, and CNN models.

Data set		Accuracy (%)				
		PCA-LDA	SVM	CNN		
				1,024 × 64	256 × 256	256 × 64
Training set	Full	98.3	100	100	100	100
	Carbohydrate	99.1	100	100	100	100
	Aromatic	98.8	100	100	100	100
Test set	Full	96.6	98.3	100	100	100
	Carbohydrate	93.3	95.0	100	100	100
	Aromatic	95.0	98.3	100	100	100

PCA-LDA models on the full, carbohydrate, and aromatic datasets were 98.3 %, 99.1 %, and 98.7 %, respectively. The accuracies of the test were 96.6 %, 93.3 %, and 95.0 %, respectively (the corresponding confusion matrices are depicted in Fig. S1). With PCA-LDA, it becomes difficult to determine which variables have a major influence on classification because LDA is applied after performing dimensionality reduction through PCA. Therefore, we used the SHAP algorithm to identify variables that influence classification. Fig. 3(a − c) shows the mean absolute value of the SHAP value of each variable used in the PCA-LDA model. It can be said that the larger the SHAP value for each variable, the greater the effect on the classification (Lundberg & Lee, 2017). In the full region, signals at 1.59–1.60, 3.70, 3.80–3.84, 3.92, and 4.07 ppm were confirmed as main variables. Signals at 3.70–4.07 ppm originate from the metabolites related to carbohydrates, such as glucose, fructose, and sucrose, and 1.59–1.60 ppm is the region of fatty acid resonances. In the carbohydrate region, signals at 3.70–3.92 and 4.07 ppm were used as main variables, as with the full region, and additionally, a signal at 3.22 ppm, which is the region of phosphocholine resonances, was confirmed as a main variable. In the aromatic region, signals at 8.44–8.45 and 6.74–6.89 ppm were identified as main variables, which were peaks corresponding to histidine and tyrosine. In the previous study (Lee et al., 2020), 14 distinguishable metabolites ^{1}H NMR signals, excluding the carbohydrate region, were used for statistical analysis. Among these, seven signals, such as amino acids (e.g., histidine) and glucose, which showed significant variations in integral values for red pepper powders from each geographical origin, were selected as discriminant indicators. However, the present study, employing AI techniques, has demonstrated that signals in this region can also be effectively used in determining the geographical origin of Asian red pepper powders.

3.3. Discrimination of the geographical origin of Asian red pepper powders by SVM models

SVM models are affected by parameters such as the cost function, gamma function, and kernel function. Cost means how much data is allowed to be included in different classes. A large cost value creates a hyperplane with a small margin so that all data belong to its own class, which risks overfitting. In contrast, when the cost value is small, there is a possibility of underfitting because a hyperplane with a large margin is created. Gamma refers to the space that one data point affects. A higher gamma value indicates that a data point has a limited impact within a narrow range, whereas a lower gamma value suggests that a data point has a broader influence over a wider range. Therefore, if the gamma value is small, a smooth hyperplane can be found because many data sets affect the creation of a hyperplane. In addition, a nonlinear SVM is employed when the data cannot be separated by a straight line (Noble, 2006). Nonlinear SVMs use kernel functions, such as the radial basis function (RBF), and polynomial functions in order to transform the original nonlinear data pattern into a linearly separable form in a high-dimensional feature space. For the SVM model to show good performance, appropriate hyperparameters must be set according to each data set. Therefore, 5-fold cross-validation and random search were used to find the appropriate hyperparameter. As a result, hyperparameters for the full, carbohydrate, and aromatic regions were obtained individually and are shown in Table S2. For the carbohydrate region, the cost value is higher than that of other regions because intensities by country of origin are similar.

The SVM model, constructed with the specified hyperparameters, achieved training accuracies of 100 % for the full, carbohydrate, and aromatic regions. However, as shown in Table 1, the test set accuracies for these regions were 98.3 %, 95.0 %, and 98.3 %, respectively (the corresponding confusion matrices are depicted in Fig. S2). In the case of SVM models, it is difficult to determine which variable has a significant influence on the classification because the data dimensionality is increased through a kernel function. Therefore, as applied in the PCA-LDA model, the SHAP algorithm was used to determine which
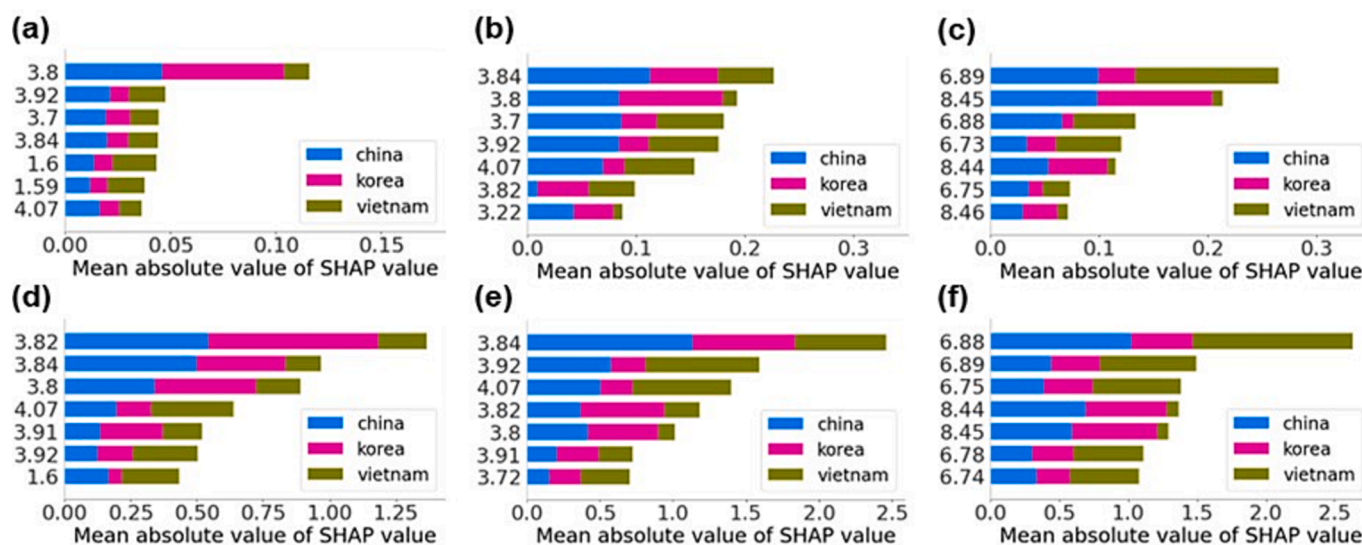
Fig. 3. The mean absolute value of SHAP values for each variable of (**a** and **d**) full, (**b** and **e**) carbohydrate, and (**c** and **f**) aromatic regions for PCA-LDA (top) and SVM (bottom) models.

variables affected the classification. The mean absolute value of the SHAP value for each variable in the SVM model is depicted in Fig. 3(d – f). Variables with a large mean absolute value of the SHAP value are considered to have a significant influence on the classification process (Amari & Wu, 1999; Rodríguez-Pérez & Bajorath, 2019). In the full region, signals at 3.80–3.84, 3.91–3.92, 4.07, 3.69, and 1.60 ppm had a significant influence on the classification. As with PCA-LDA, it was found that the origin of the sample was determined by the difference in the content of fatty acids and carbohydrates. As with the PCA-LDA model, signals at 3.72–3.92 and 4.07 ppm in the carbohydrate region were used as main variables. Additionally, in the aromatic region, 8.44–8.45 and 6.74–6.89 ppm peaks, which correspond to histidine and tyrosine, respectively, were identified as important variables. These findings align with the results obtained from the PCA-LDA model.

### 3.4. Discrimination of the geographical origin of Asian red pepper powders by CNN models

The performance of CNN models can be influenced by various hyperparameters, such as the number of filters, the number of layers, and the learning rate. Filters with weights are responsible for scanning input data and extracting specific features. The number of filters determines the number of feature maps generated. If the number of filters is excessively large, it can lead to overfitting and result in a high generalization error. Conversely, if the number of filters is too small, the training process may not be executed effectively (Johnson et al., 2020). The number of layers in a CNN is associated with the complexity of the network structure. Increasing the number of layers generally leads to improved results. However, this also increases the computational complexity, resulting in longer learning times (Yu & Zhu, 2020). The learning rate is related to the training process. In order to find the optimal weights and biases, the loss function should have a minimum value. The learning rate refers to the amount of learning to be learned in one learning process, and the weight parameter is updated after each learning process in a direction toward minimizing the loss function. If the learning rate is too large, the model may pass the global minimum, and if the learning rate is too small, it may fall into the local minimum (Behera et al., 2006). Therefore, it is important to fine-tune these hyperparameters.

Bayesian optimization is a statistical method that uses Gaussian processes to model the performance of an algorithm. It employs an iterative approach to identify the optimal parameter sets within the

parameter space of the given algorithm. CNNs can be regarded as intricate mathematical functions with underlying statistical models that are not fully known. Bayesian optimization addresses this challenge by creating a surrogate model that estimates the objective function based on previous outcomes, thereby determining the hyperparameter values that result in the highest learning accuracy. This optimization technique offers a significant improvement in the efficiency of hyperparameter tuning compared to a simple grid search (Gurbani et al., 2018; Snoek et al., 2012). Furthermore, Bayesian optimization has been coupled with batch normalization and early stopping techniques to mitigate overfitting in CNN models (Gurbani et al., 2018; Li et al., 2022; Shahriari et al., 2015; Snoek et al., 2012). Consequently, in this study, hyperparameter tuning was conducted using Bayesian optimization (Shahriari et al., 2015). The optimal hyperparameter sets for CNNs in the full, carbohydrate, and aromatic spectral regions are presented in Table S3, and the corresponding loss curves on the training and validation datasets are depicted in Fig. S3.

Table 1 shows the results of determining the geographical origins of red pepper powders using the CNN models with the optimal hyperparameters. The accuracies of training and testing from all types of the input data (3 spectral regions × 3 image resolutions) were unanimously 100 % (the corresponding confusion matrices are depicted in Fig. S4, S5, and S6). PCA-LDA and SVM showed relatively low accuracies in the carbohydrate region compared to other regions, possibly due to the significant overlap of carbohydrate signals. However, in the CNN models, the carbohydrate region also showed 100 % accuracy. This may be because even a slight difference could be captured by CNN feature extraction layers. In addition, it is a superior result compared to the accuracy of 93.75 % reported by Lee et al. (2020), who used the NMR spectra of a total of 62 red pepper powder samples for data analysis.

In the field of AI, CNNs are often referred to as ''black boxes'' due to their complex structure involving multiple hidden layers. However, it is crucial to understand the specific regions of the input that CNNs utilize for classification, especially considering the potential for misinterpretation. To address this concern, Grad-CAM, an explainable artificial intelligence (XAI) technique, was employed to identify the regions of substantial influence in determining the geographic origin. The Grad-CAM results (for 1,024 × 64 pixel images), as depicted in Fig. 4, highlight the red region as having a significant impact on the classification of the data. Specifically, in the full region (Fig. 4a), it is observed that the signals originating from the carbohydrate region have a substantial influence on the classification of Korean red pepper. More specifically, the
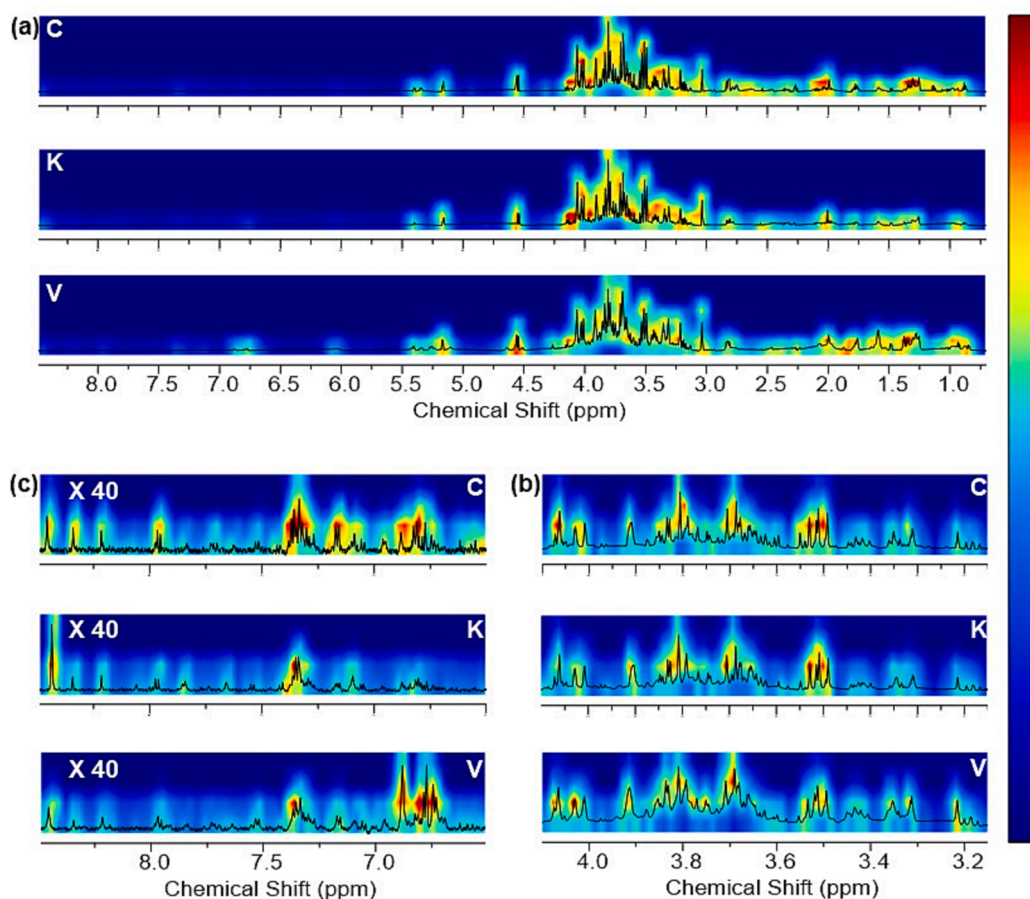
Fig. 4. Grad-CAM results of (**a**) full, (**b**) carbohydrate, and (**c**) aromatic regions in the CNN model with 1,024 × 64 pixel images. (C: Chinese, K: Korean, V: Vietnamese).

signal at 4.3 ppm, corresponding to the sucrose peak, has a significant effect on the classification. Similarly, in the case of Chinese red pepper, the signals from the carbohydrate region also have a notable impact on the classification. Additionally, the signals from fatty acids at 2.0 and 1.4 ppm have a considerable influence on the classification. For Vietnamese red pepper, the β-glucose peak near 4.5 ppm and the fatty acid signal at 1.4 ppm have a significant effect on the classification.

In the carbohydrate region (Fig. 4b), Korean red peppers might be classified based on signals around 3.5 ppm, originating from β-glucose. Chinese red peppers, however, might be classified through a signal at 4.06 ppm, originating from fructose. In the Vietnamese red peppers, the fructose peak near 3.7 ppm had a strong effect on the classification.

In the aromatic region (Fig. 4c), Korean red peppers might be classified through the histidine peak near 8.4 ppm and the phenylalanine peak near 7.2 ppm. Chinese red peppers might be classified through the phenylalanine peak near 7.2 ppm and the tyrosine peak near 6.8 ppm. Vietnamese red peppers, similarly, might be classified based on the tyrosine peak near 6.8 ppm.

The Grad-CAM results for the full region indicated that the CNN-based classification placed significant emphasis on the relatively strong carbohydrate signals, which aligns with the findings obtained from the SHARP values of the PCA-LDA and SVM models (Fig. 3a and d). Additionally, the CNNs also incorporated the weaker peaks from metabolite peaks, such as β-glucose and fatty acids, which were previously primarily used for determining geographic origin through ^{1}H NMR analysis of metabolite contents in conjunction with canonical discriminant analysis (Lee et al., 2020). Consequently, it appears that the CNN model exhibited superior performance in comparison to the PCA-LDA and SVM models.

Fig. 5(a) shows the input spectra of the samples that were misclassified by the PCA-LDA and SVM models but correctly classified by the CNN models. The misclassification occurred when a Chinese sample was classified as Korean by the PCA-LDA model, and a Korean sample was classified as Chinese by the SVM model. The spectral regions at 3.84 and 3.92 ppm were identified as significant in the classification using the PCA-LDA and SVM models. However, these regions appeared quite similar in the misclassified samples, which likely contributed to the misclassification. In contrast to those models, the CNN model was able to accurately classify the samples by capturing even the subtle differences in these spectral regions (Fig. 5b). This outcome highlights the potential advantage of CNN models over conventional approaches in achieving accurate classification of red pepper powders.

Furthermore, an investigation was conducted to determine the impact of image resolution on learning accuracy. Specifically, images with resolutions of 256 × 64, 256 × 256, and 1,024 × 64 pixels were utilized and compared. The results showed that all image resolutions yielded a training and testing accuracy of 100 % (Table 1). Because the CNN model identifies differences in the 1D NMR spectra through the feature extraction process, the same accuracy can be attained even when using low-resolution image data within the level representing regional characteristics. All resolutions applied in this study appeared to satisfy these conditions and were able to achieve 100 % accuracy (see the loss curves of CNN models shown in Fig. S3 and the confusion matrices in Fig. S5 and S6). However, upon examining the Grad-CAM outcomes for the 256 × 64 or 256 × 256 images (Fig. S7 and S8), it was observed that the resulting heatmaps were excessively blurred, making it challenging to discern the crucial spectral regions involved in the classification process. Consequently, although the CNNs achieved a classification
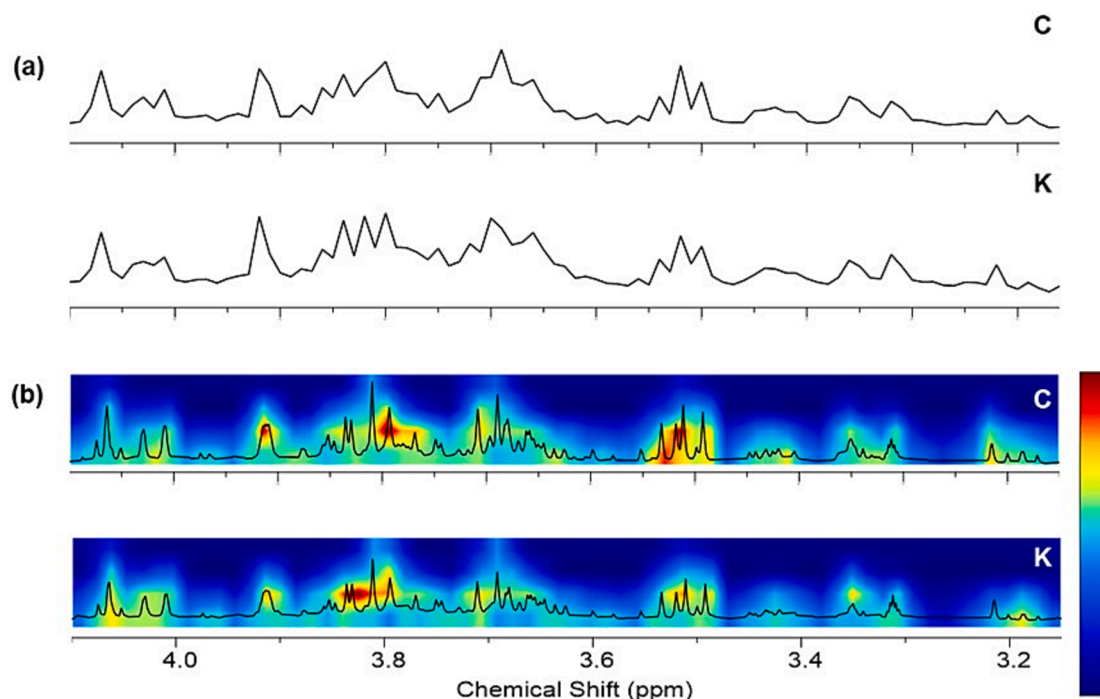
Fig. 5. (**a**) The input spectra of the misclassified samples by PCA-LDA (top) and SVM (bottom) models. (**b**) The CNN Grad-CAM results of misclassified samples in PCA-LDA and SVM models. (C: Chinese, K: Korean).

accuracy of 100 % across all image resolutions, only those CNNs trained with a resolution of 1,024 × 64 were easily interpretable. Notably, the resolution along the *x*-axis, which directly influences the spectral dispersion, holds greater significance than the resolution along the *y*-axis. Therefore, considering the transparency of the trained neural networks, an image resolution of 1,024 × 64 is deemed the most suitable for effectively discriminating the geographic origins of Asian red pepper powders in our study.

One advantage of using 2D CNN models for analyzing spectroscopic data is that they can be directly applied, even if only an image of the spectrum is available, without the need for raw data. To assess this applicability of the developed CNN model, the 1D ^{1}H NMR spectrum of Korean red pepper powder published by Lee et al. (2020) was captured as an image, and then its geographical origin was classified. Due to the lack of available literature on the ^{1}H NMR spectra of Asian red pepper powders, various verifications could not be conducted. However, the spectrum was accurately classified as that of Korean red pepper powder (Fig. S9).

4. Conclusion

In this study, deep learning-based CNN models were examined to discriminate the geographical origin of Asian red pepper powders by analyzing their ^{1}H NMR spectra. To assess the impact of the spectral domain on discrimination, the 1D spectra of three different regions were transformed into 2D images and used as input data. Furthermore, the discrimination performance was compared by employing PCA-LDA and SVM models. As a result of discrimination, the PCA-LDA and SVM models showed training and testing accuracies of over 95 %, although there were some differences depending on the input region. However, the CNN model showed 100 % accuracy regardless of the input area region and/or resolution. Grad-CAM analysis clearly showed that the developed CNN models recognized the origin of samples based on changes in the distribution of various metabolites. A larger dataset, including external validations, would greatly facilitate the development of a model with higher generalizability.

The results of this study demonstrated that the CNN method has the potential to accurately determine the geographical origin of various foods by analyzing their 1D ^{1}H NMR spectra converted into 2D images. This effectiveness could be attributed to the CNN model's ability to discern subtle variations in the data, thereby recognizing them as significant features. Compared to previous studies, this approach may be considered unique and innovative because it uses 2D images of 1D NMR spectra as input for the CNN models, eliminating the need for raw data in the classification process.

CRediT authorship contribution statement

Byung Hoon Yun: Conceptualization, Formal analysis, Writing – original draft. **Hyo-Yeon Yu:** Conceptualization, Investigation, Writing – original draft. **Hyeongmin Kim:** Formal analysis, Investigation, Visualization. **Sangki Myoung:** Conceptualization, Methodology, Visualization. **Neulhwi Yeo:** Formal analysis, Funding acquisition, Visualization. **Jongwon Choi:** Validation, Writing – review & editing. **Hyang Sook Chun:** Validation, Writing – review & editing. **Hyeonjin Kim:** Formal analysis, Investigation, Visualization. **Sangdoo Ahn:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2023.138082.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (ReLu). *arXiv preprint arXiv: 1803.08375.* 1–7. 10.48550/arXiv.1803.08375.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1–6). IEEE. 10.1109/ICEngTechnol.2017.8308186.

Amari, S., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks, 12*(6), 783–789. https://doi.org/10.1016/S0893-6080(99)00032-5

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *The Institute for Signal and Information Processing, 18*, 1–8.

Becerra-Martínez, E., Florentino-Ramos, E., Pérez-Hernández, N., Zepeda-Vallejo, L. G., Villa-Ruano, N., Velázquez-Ponce, M., … Bañuelos-Hernández, A. E. (2017). 1H NMR-based metabolomic fingerprinting to determine metabolite levels in serrano peppers (Capsicum annum L.) grown in two different regions. *Food research international, 102*, 163–170. https://doi.org/10.1016/j.foodres.2017.10.005

Behera, L., Kumar, S., & Patnaik, A. (2006). On adaptive learning rate that guarantees convergence in feedforward networks. *IEEE transactions on neural networks, 17*(5), 1116–1125. https://doi.org/10.1109/TNN.2006.878121

Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., … Tauler, R. (2017). Chemometrics in analytical chemistry—part I: History, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry, 409*(25), 5891–5899. https://doi.org/10.1007/s00216-017-0517-1

Cai, Z., Huang, Z., He, M., Li, C., Qi, H., Peng, J., … Zhang, C. (2023). Identification of geographical origins of Radix Paeoniae Alba using hyperspectral imaging with deep learning-based fusion approaches. *Food Chemistry., 422*, 1–10. https://doi.org/10.1016/j.foodchem.2023.136169

Chen, D., Wang, Z., Guo, D., Orekhov, V., & Qu, X. (2020). Review and prospect: Deep learning in nuclear magnetic resonance spectroscopy. *Chemistry–A. European Journal, 26*(46), 10391–10401. https://doi.org/10.1002/chem.202000246

Feng, X., Chen, X., Zheng, X., Zhu, H., Qi, Q., Liu, S., … Che, J. (2021). Latest trend of milk derived exosomes: Cargos, functions, and applications. *Frontiers in nutrition, 8*, 1–12. https://doi.org/10.3389/fnut.2021.747294

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879*, 10–23. https://doi.org/10.1016/j.aca.2015.02.012

Gurbani, S. S., Schreibmann, E., Maudsley, A. A., Cordova, J. S., Soher, B. J., Poptani, H., … Cooper, L. A. (2018). A convolutional neural network to filter artifacts in spectroscopic MRI. *Magnetic Resonance in Medicine, 80*(5), 1765–1775. https://doi.org/10.1002/mrm.27166

Hou, X., Wang, G., Wang, X., Ge, X., Fan, Y., & Nie, S. (2020). Convolutional neural network based approach for classification of edible oils using low-field nuclear magnetic resonance. *Journal of Food Composition and Analysis, 92*, 1–8. https://doi.org/10.1016/j.jfca.2020.103566

Jin, G., Zhu, Y., Cui, C., Yang, C., Hu, S., Cai, H., … Hou, R. (2023). Tracing the origin of Taiping Houkui green tea using ¹H NMR and HS-SPME-GC–MS chemical fingerprints, data fusion and chemometrics. *Food Chemistry, 425*, 1–9. https://doi.org/10.1016/j.foodchem.2023.136538

Johnson, F., Valderrama, A., Valle, C., Crawford, B., Soto, R., & Ñanculef, R. (2020). Automating configuration of convolutional neural network hyperparameters using genetic algorithm. *IEEE Access, 8*, 156139–156152. https://doi.org/10.1109/ACCESS.2020.3019245

Jolliffe, I. (2005). *Principal Component Analysis* (2nd ed.). Springer. 10.1002/0470013192.bsa501.

Katerinopoulou, K., Kontogeorgos, A., Salmas, C. E., Patakas, A., & Ladavos, A. (2020). Geographical origin authentication of agri-food products: A review. *Foods, 9*(4), 1–16. https://doi.org/10.3390/foods9040489

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*, 1–15, 10.48550/arXiv.1412.6980.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Kumar, N., Bansal, A., Sarma, G. S., & Rawal, R. K. (2014). Chemometrics tools used in analytical chemistry: An overview. *Talanta, 123*, 186–199. https://doi.org/10.1016/j.talanta.2014.02.003

Lee, D., Kim, M., Kim, B. H., & Ahn, S. (2020). Identification of the geographical origin of Asian red pepper (Capsicum annuum L.) powders using 1H NMR spectroscopy. *Bulletin of the Korean Chemical Society, 41*(3), 317–322. https://doi.org/10.1002/bkcs.11974

Li, G., Jian, X., Wen, Z., & AlSultan, J. (2022). Algorithm of overfitting avoidance in CNN based on maximum pooled and weight decay. *Applied Mathematics and Nonlinear Sciences, 7*(2), 965–974. https://doi.org/10.2478/amns.2022.1.00011

Lia, F., Vella, B., Mangion, M. Z., & Farrugia, C. (2020). Application of ¹H and ¹³C NMR fingerprinting as a tool for the authentication of Maltese extra virgin olive oil. *Foods, 9*(6), 1–14. https://doi.org/10.3390/foods9060689

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*, 1–10. https://doi.org/10.5555/3295222.3295230

Luykx, D. M. A. M., & van Ruth, S. M. (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry, 107*(2), 897–911. https://doi.org/10.1016/j.foodchem.2007.09.038

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Analytical Chemistry, 80*(19), 7562–7570. https://doi.org/10.1021/ac800954c

Maione, C., Barbosa, F., Jr, & Barbosa, R. M. (2019). Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review. *Computers and Electronics in Agriculture, 157*, 436–446. https://doi.org/10.1016/j.compag.2019.01.020

Medina, S., Perestrelo, R., Silva, P., Pereira, J. A. M., & Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology, 85*, 163–176. https://doi.org/10.1016/j.tifs.2019.01.017

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli, & V. Sandra (Eds.), *Machine learning: methods and applications to brain disorders* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

Rodríguez-Pérez, R., & Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medical Chemistry, 63*(16), 8761–8777. https://doi.org/10.1021/acs.jmedchem.9b01101

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE, 104*(1), 148–175. 10.1109/JPROC.2015.2494218.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., … Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging, 35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Shin, S., Lee, Y., Kim, S., Choi, S., Kim, J. G., & Lee, K. (2021). Rapid and non-destructive spectroscopic method for classifying beef freshness using a deep spectral network fused with myoglobin information. *Food Chemistry, 352*, 1–10. https://doi.org/10.1016/j.foodchem.2021.129329

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems, 25*, 1–12. 10.48550/arXiv.1206.2944.

Tahir, H. E., Arslan, M., Mahunu, G. K., Mariod, A. A., Hashim, S. B., Xiaobo, Z., … Musa, T. H. (2022). The use of analytical techniques coupled with chemometrics for tracing the geographical origin of oils: A systematic review (2013–2020). *Food Chemistry, 366*, 1–14. https://doi.org/10.1016/j.foodchem.2021.130633

Varmuza, K., & Filzmoser, P. (2016). Introduction to multivariate statistical analysis in chemometrics. *CRC Press.* https://doi.org/10.1201/9781420059496

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters, 25*(7), 926–930. https://doi.org/10.1109/LSP.2018.2822810

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights Imaging, 9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yu, T., & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*, 1–56. 10.48550/arXiv.2003.05689.

Zhang, J., Yang, R., Chen, R., Li, Y. C., Peng, Y., & Wen, X. (2019). Geographical origin discrimination of pepper (Capsicum annuum L.) based on multi-elemental concentrations combined with chemometrics. *Food Science and Biotechnology, 28*, 1627–1635. https://doi.org/10.1007/s10068-019-00619-3

Zhang, L., Ding, X., & Hou, R. (2020). Classification modeling method for near-infrared spectroscopy of tobacco based on multimodal convolution neural networks. *Journal of Analytical Methods in Chemistry, 2020*, 1–13. https://doi.org/10.1155/2020/9652470

Zhang, X., Xu, J., Yang, J., Chen, L., Zhou, H., Liu, X., … Ying, Y. (2020). Understanding the learning mechanism of convolutional neural networks in spectral analysis. *Analytica Chimica Acta, 1119*, 41–51. https://doi.org/10.1016/j.aca.2020.03.055