JAAS

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: J. Anal. At. Spectrom., 2025, 40, 2222

# Fusion of elemental and molecular fingerprints for accurate classification of kimchi by country of origin†

Sandeep Kumar,[a] Yujin Oh,[b] Hyemin Jung,[b] Kyung-Sik Ham,[c] Hyun-Jin Kim,[d] Song-Hee Han,[e] Sang-Ho Nam*[abf] and Yonghoon Lee*[abf]

The geographical origin of commercial kimchi products is a key indicator of their quality, authenticity, and economic value. In this study, we propose a spectroscopic classification method combining laser-induced breakdown spectroscopy (LIBS) and infrared (IR) spectroscopy to differentiate kimchi samples from South Korea and China. LIBS was used to obtain elemental profiles based on the emission intensities of K, Mg, Na, Ca, C, H, and O, while IR spectroscopy captured molecular features. Principal component analysis of IR spectra in the carbohydrate absorption region (1254–1018 $cm^{-1}$) identified the third principal component (PC3) as the most discriminative. Classification models using *k*-nearest neighbors (*k*-NN) were evaluated with leave-one-out cross-validation. Two LIBS-only models—using variable sets (i) K I (766 nm), O I (777 nm), C I (248 nm), and (ii) K I, O I, Mg II (279 nm)—achieved 94.4% accuracy. The IR-only model reached 86.4%. Fusion of LIBS and IR features, with optimized weighting for the IR variable, enhanced model performance. The best result (96.8% accuracy) was achieved by combining LIBS variables K I, O I, and C I with IR PC3. We also introduce a statistical method to predict the optimal weighting factor for fusion, reducing computational complexity by minimizing the number of neighbors in *k*-NN. This LIBS-IR fusion strategy provides a robust tool for verifying kimchi origin.

Received 19th May 2025
Accepted 11th July 2025

DOI: 10.1039/d5ja00200a

rsc.li/jaas



## 1. Introduction

In Korean cuisine, kimchi is a traditional food consumed daily with every meal as a side dish. Moreover, it has become a global representative of Korean culture. Kimchi is a fermented vegetable dish primarily made from cabbage and radish, pickled in salt, along with other ingredients such as green onion, garlic, ginger, red pepper powder, and salted fish.[1–3] For the salting process, various types of salt are used, including purified salt, rock salt, recrystallized salt, and solar sea salt.[4] Fermentation, which follows the salting process, imparts kimchi's distinct flavor and extends its shelf life. During fermentation, probiotics are produced that promote gut health. Kimchi is also rich in calcium, iron, vitamin A, vitamin B1 (thiamine), and vitamin B2 (riboflavin), and contains a diverse array of lactic acid bacteria.[5] Recognized as one of the five healthiest foods globally, kimchi offers numerous functional health benefits, including anti-aging, anti-cancer, anti-diabetic, anti-obesity, and antioxidant effects.[6–10] Consequently, the global demand for kimchi has increased significantly. In response to this growing demand, kimchi production has expanded beyond South Korea to countries such as China, Japan, the United States, and Indonesia. However, in the South Korean domestic market—the primary focus of this study—commercial kimchi is overwhelmingly sourced from either South Korea or China. According to the Korea Customs Service, over 99% of imported kimchi in South Korea originates from China, while kimchi produced in countries such as Japan or the United States is typically intended for local consumption in those countries rather than for export to Korea.[11] Therefore, the inclusion of only South Korean and Chinese samples reflects the actual distribution dynamics relevant to origin verification in the Korean market. To maintain transparency, South Korea has mandated the labeling of the origin of kimchi sold in markets. However, the price of kimchi varies depending on the country of production due to differences in raw material and resource

[a]Plasma Spectroscopy Analysis Center, Mokpo National University, Muan, Jeonnam, 58554, Republic of Korea. E-mail: shnam@mokpo.ac.kr; yhlee@mokpo.ac.kr; Fax: +82-61-450-2339

[b]Department of Chemistry, Mokpo National University, Muan, Jeonnam, 58554, Republic of Korea

[c]Department of Food Engineering, Mokpo National University, Muan, Jeonnam, 58554, Republic of Korea

[d]Division of Applied Life Sciences (BK21 plus), Department of Food Science & Technology, Institute of Agriculture and Life Science, Gyeongsang National University, Jinju, Gyeongnam, 52828, Republic of Korea

[e]Division of Navigation Science, Mokpo National Maritime University, Mokpo, Jeonnam, 58628, Republic of Korea

[f]Department of Energy and Chemical Engineering, Mokpo National University, Muan, Jeonnam, 58554, Republic of Korea

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5ja00200a

costs. This price disparity may lead to the fraudulent mislabeling of kimchi products with false origin claims. To address this issue, appropriate spectroscopic techniques are essential for accurately classifying food products by their geographical origins, thereby ensuring product integrity and consumer safety.

Conventional elemental analysis techniques, such as atomic absorption spectroscopy (AAS), inductively-coupled plasma optical emission spectrometry (ICP-OES), and inductively-coupled plasma mass spectrometry (ICP-MS), can be used for quality evaluation and determination of the geographical origin of food products.[12–14] However, these techniques have issues that include sample preparation time, analyte loss, and potential contamination during the conversion of solid materials into liquid solutions. They also have problems with sample digestion prior to analysis.[15,16] Laser-induced breakdown spectroscopy (LIBS) is one of the rapid and simple elemental analysis techniques that requires minimal sample preparation and can be applied directly to the solid samples.[17,18] In LIBS measurements, laser pulses are focused on the sample surface that vaporize and ionize a small portion of the sample material to generate microplasmas. The laser-induced plasma then de-excites and dissipates energy, mainly *via* optical emissions, within tens of microseconds. The optical emissions from laser-induced plasma exhibit atomic and ionic lines, offering information about the sample's elemental composition and enabling both sample classification and quantification of analytes of interest. Recently, our group demonstrated the feasibility of the LIBS technique for distinguishing kimchi purchased in South Korean markets based on their counties of production, using Mg and K emission lines.[19] Recently, LIBS has been used in combination with infrared (IR) spectroscopy to grasp comprehensive information about samples, covering both elemental and molecular features. LIBS in combination with IR spectroscopy is also explored for different food products, such as the enhanced discrimination of geographical origin of soybean paste,[20] for rice varieties,[21] and milk vetch root samples.[22] However, the combination of LIBS and IR spectroscopy has not been used to distinguish kimchi products. To our knowledge, other hyphenated spectroscopic techniques have also not been applied to this purpose. This gap in the literature motivates us to explore the potential of combining LIBS and IR for classifying kimchi by country of origin.

In this study, we investigated the feasibility of combining LIBS and IR spectroscopy for improved classification of kimchi produced in China and South Korea. Spectral variables from LIBS and IR were used to model the two sample classes using *k*-nearest neighbors (*k*-NN). In the LIBS spectra, emission peaks corresponding to Na, K, Mg, Ca, C, H, and O were identified. Following the forward selection scheme based on the interclass distance values of the seven emission peak intensities, two LIBS variable sets showing the highest classification accuracy (94.4%) were identified: (i) K I (766 nm), O I (777 nm), and C I (248 nm); and (ii) K I (766 nm), O I (777 nm), and Mg II (279 nm). For IR spectroscopy, the region between 1254 and 1018 $cm^{-1}$—corresponding to characteristic carbohydrate absorption bands—showed relatively high discriminative power. Principal component analysis (PCA) was performed on this region, and the third principal component (PC3) was selected as a variable, and the IR model based on the PC3 scores from the selected spectral region showed the classification accuracy of 86.4%. Although the two LIBS models showed the same classification performances, that composed of K I (766 nm), O I (777 nm), and C I (248 nm) were found to comprise the best variable set for pairing with the IR variable. By assigning an optimal weighting factor to the IR variable, the fused LIBS-IR model achieved the classification accuracy of 96.8% outperforming the standalone LIBS and IR models. These results indicate that the combination of LIBS and IR spectroscopy provides complementary elemental and molecular information of kimchi products and demonstrates strong potential for accurately verifying the country of origin of kimchi products distributed in the market.

## 2. Materials and methods

### 2.1 Sample preparation

In 2021, the National Agricultural Products Quality Management Service (NAQS) of South Korea conducted the kimchi sample collection process.[13] From Korean markets, a total of 125 cabbage kimchi samples were gathered. Of them, 72 were imported from China, and 53 were domestic ones. The 12 regions of South Korea that comprise Chungbuk-do Cheongju-si, Chungnam-do Asan-si, Chungnam-do Dangjin-si, Chungnam-do Goesan-si, Daegu, Daejeon, Gangwon-do Donghae-si, Gyeongbuk-do Gimcheon-si, Gyeongbuk-do Gumi-si, Gyeonggi-do Pyeongtaek-si, Gyeongnam-do Changnyeong-gun, and Jeonbuk-do Gimje-si were the locations from which the kimchi samples were acquired directly from the sales outlets. The import certificates of the Chinese samples were thoroughly examined at nearby grocery stores prior to the separation of the Chinese and Korean kimchi samples. For each sample, 200 grams of kimchi were first taken and washed under running deionized water to remove seasonings. All samples were cleaned, dried, and then pre-frozen for eight hours at −40 °C until they weighed 50 grams. After that, the samples were freeze-dried for the whole day. A ceramic homogenizer (Pulverisette 14, FRITSCH, Idar-Oberstein, Germany) was then used to ground the dry materials. A 0.5 mm filter was used to filter the final powder. This freeze-dried powder was used directly for IR measurements without any further preparation, whereas additional pellet formation was required for LIBS measurements.

### 2.2 IR measurements

The IR spectra were recorded using a commercial Fourier-transform IR instrument (Frontier PerkinElmer, Waltham, USA) with the Attenuated Total Reflection (ATR) method (GladiATR accessory, PIKE Technologies, Fitchburg, USA). Kimchi powder samples were placed to cover the ATR diamond crystal, and measurements were repeated five times for each sample. The spectrum for each sample, obtained by averaging a total of 75 measurements in the range of 450–4000 $cm^{-1}$ was used for analysis. For the analysis of IR spectral data,

preprocessing was conducted by converting transmittance to absorbance.

### 2.3 LIBS measurements

Details of the experimental conditions for LIBS were reported previously.[19] Briefly, a 13 mm diameter pellet was prepared for each samples using the freeze-dried kimchi powder treated by the procedure described above. LIBS spectra of the pelletized kimchi samples were acquired using a commercial LIBS instrument (J200, Applied Spectra, Inc.). The LIBS instrument was integrated with a flash-lamp-pumped Q-switched neodymium-doped yttrium aluminum garnet (Nd:YAG) laser and a 6-channel charge-coupled device (CCD) spectrometer. The Nd:YAG laser emitted pulses at a repetition rate of 10 Hz, with a wavelength of 266 nm, a pulse energy of 9.4 mJ per pulse, and a pulse duration of 7 ns. The laser beam was focused on the pellet surface through a high-power focusing lens, and the focused spot diameter was set to be 35 μm. The light emitted from the laser-induced plasma was collected by two plano-convex lenses and delivered to the spectrometer *via* an optical fiber bundle. The 6-channel CCD spectrometer covered wavelengths from 180 to 1047 nm with a spectral resolution of ~0.1 nm. To alleviate unwanted continuum background and line broadening, the CCD detection gate was deliberately delayed by 0.5 μs from the laser Q-switching event and remained open for 1 μs. The sample chamber was purged with helium gas flowing at a rate of 0.7 L $min^{-1}$. Each LIBS spectrum was recorded from an 8 mm long line-scan on the sample pellet surface. Along each scan line, 81 laser pulses were launched, and the emissions were accumulated for recording a LIBS spectrum. During the LIBS measurement, the sample stage was linearly translated at a rate of 1 mm $s^{-1}$. For each sample pellet, fifteen line-scans were conducted. The distance between two adjacent line-scans was 500 μm. The analysis was carried out by taking an average of the fifteen LIBS spectra for each sample.

### 2.4 IR spectral preprocessing

The preprocessing of the IR spectra is crucial for extracting reliable variables. It typically enhances the robustness and accuracy of the following quantitative or classification analyses and also increases the interpretability of the data by correcting issues related to spectral data acquisition.[23] It ensures only the sample's chemical differences are involved in the final analysis results. In this study, the IR spectra of all kimchi samples were pre-processed following these steps. Initially, the average (of 5 spectra) spectrum recorded in the spectral range 400–4000 $cm^{-1}$ was first obtained for each sample. From the IR spectra, it is observed that the O–H band mainly predominates in the spectral range beyond 1750 $cm^{-1}$. Thus, the IR spectra of all kimchi samples in the 750–1750 $cm^{-1}$ range were chosen for the spectral pre-processing. Then, the first derivatives were obtained in the range of 750–1750 $cm^{-1}$ using the Savitzki–Golay method.[24] This method resolves overlapped bands in the complex IR spectra and also helps in eliminating slopes at lower wavenumbers. The first derivative spectra of 125 kimchi samples were then normalized by their unit vectors.

### 2.5 *k*-NN modeling and validation

The *k*-NN algorithm is a non-parametric, instance-based method that classifies a sample by identifying the majority class among its *k* closest training objects in the feature space.[25] It does not rely on any assumptions about the underlying data distribution and is particularly suited for simple, flexible implementation. The parameter "*k*" in the *k*-NN algorithm determines how many of the nearest neighbors are considered when assigning a class to a sample. A small *k* (*e.g.*, $k = 1$) often reflects local variations and may capture subtle distinctions, but can be more susceptible to noise or outliers. In contrast, larger *k* values provide more general predictions by averaging over more samples, but may smooth out meaningful differences, especially when the data exhibits local structure. Therefore, the choice of *k* greatly influences classification performance. To determine the optimal value of *k* in our modeling, we systematically evaluated classification accuracy across a wide range of *k* values—from 1 up to one less than the total number of samples—using leave-one-out cross-validation (LOOCV).[26] In LOOCV, each sample in the dataset is sequentially used as a test object, while the remaining objects form the training set. This process is repeated for all objects, ensuring that each one is used for validation exactly once. Classification accuracy is then calculated as the proportion of correctly predicted objects across the entire dataset. LOOCV is especially well-suited for relatively small datasets, as it maximizes data utilization and provides a nearly unbiased estimate of model performance. In this study, LOOCV was applied to 125 cabbage kimchi samples, including 53 produced in South Korea and 72 imported from China. For each iteration, one sample was held out as the test sample, and the remaining 124 were used to train the *k*-NN model, which then predicted the country of origin of the test sample. This procedure was repeated 125 times—once per sample—and the final accuracy was determined by the percentage of correct predictions. To evaluate the effect of the number of neighbors, *k*, on classification performance, LOOCV was performed from 1 to 124. For each *k*, the entire LOOCV procedure was conducted, and the corresponding accuracy was recorded. This approach allowed us to identify the optimal *k* value for each modeling scenario. Euclidean distance was measured between a test object to each of training objects and used to assign the sample class of the test object for LOOCV.

## 3. Results and discussion

### 3.1 LIBS modeling

The average LIBS spectra of kimchi samples produced in South Korea (53 samples) and China (72 samples) are shown in Fig. 1. The spectra primarily consist of emission peaks from Na, K, Mg, Ca, C, H, and O, which were identified using the NIST Atomic Spectra Database.[27] The presence of metallic elements such as Na, K, Mg, and Ca is associated with the plant tissues that are essential micronutrients for the growth of cabbage. Additionally, the salting process during kimchi preparation further enhances their concentrations in cabbage. Although the LIBS spectra of the two groups appear similar in overall shape,

 

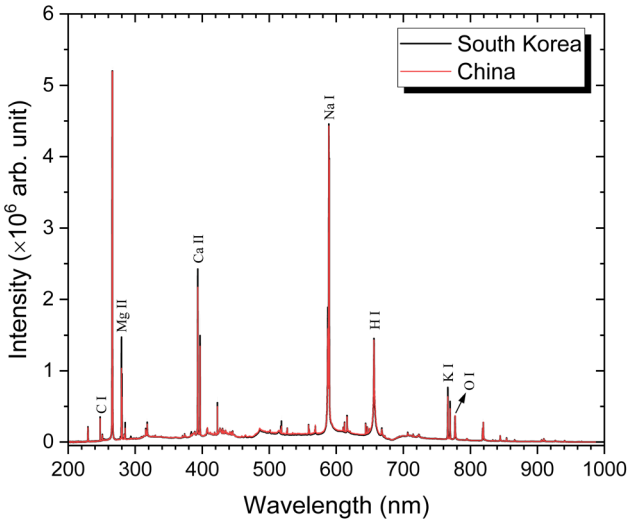


Fig. 1 LIBS spectra of kimchi samples produced in South Korea and China within the wavelength range of 200–1000 nm. The strongest emission peaks corresponding to the elements C, Mg, Ca, Na, H, K, and O are marked.

certain emission lines exhibit interclass differences that are substantially greater than the variability within each class. In the following, rather than using the full spectral range from 200 to 1000 nm, we selected only a small number (typically 1 to 4) of the most prominent and stable emission peaks as variables. These peaks were chosen based on their individual discriminative power and their potential to complement one another when used together in a multivariate model. The emission peaks of nonmetallic elements, C, H, and O, are attributed to the cabbage and other vegetables used as main ingredients in kimchi. The strongest emission peaks of Na I (590 nm), K I (766 nm), Mg II (279 nm), Ca II (393 nm), C I (248 nm), H I (656 nm), and O I (777 nm) were marked in Fig. 1. The intensities of these peaks were calculated as baseline-subtracted peak areas. These peaks are representative of emissions from their corresponding elements. Based on the intensities of these seven emission peaks extracted from the LIBS spectra of South Korean and Chinese kimchi samples, *k*-NN models were constructed to classify the two sample classes.

The best variable sets for *k*-NN modeling using the peak intensities from LIBS spectra were identified by following the forward selection scheme.[28,29] In the forward selection scheme, modeling begins with no variables initially included. At each step, a new variable is added by selecting the one that yields the greatest improvement in model performance, as measured by cross-validated classification accuracy. This process is repeated iteratively: after each addition, the next best-performing variable is identified and incorporated into the model. The procedure terminates when the inclusion of additional variables no longer results in a significant increase in accuracy. Through this approach, variable sets that balance high classification performance with minimal model complexity were systematically identified. Herein, seven one-variable *k*-NN models were trained first and their classification performances were evaluated.

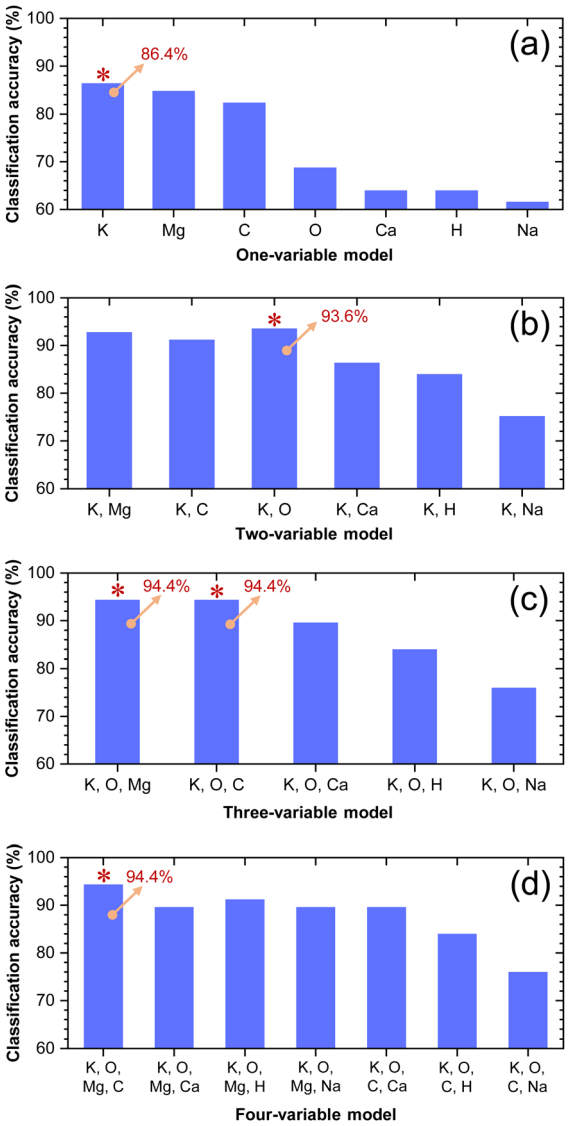


Fig. 2 Classification accuracy values from LOOCV of (a) one-, (b) two-, (c) three-, and (d) four-variable *k*-NN models. In each panel, the models showing the highest classification accuracy are marked using "*" with the corresponding accuracy values.

Fig. 2a shows the classification accuracy values of the seven one-variable *k*-NN models. Among them, the model trained using the K I emission peak intensity shows the highest classification accuracy (86.4%). Then, the K I emission peak intensity was paired with each of the others in turn to train six two-variable *k*-NN models: (K, Mg), (K, C), (K, O), (K, Ca), (K, H), and (K, Na). The classification accuracy values of these two-variable *k*-NN models are compared in Fig. 2b. When the K I emission peak intensity was paired with that of O I, the two-variable model performance was maximized with the classification accuracy of 93.6%, which is higher than that of the best one-variable model (86.4%). Thus, the O I emission peak intensity was added to the variable set where that of K I was already contained. Next, the five possible three-variable *k*-NN models were set by adding each of the Mg II, C I, Ca II, H I, and Na I emission peak

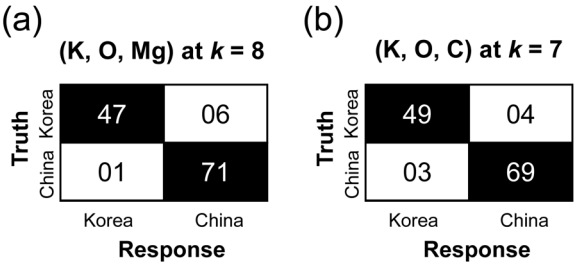


Fig. 3 Confusion matrices of three-variable standalone LIBS *k*-NN models using the emission intensities of (a) K I, O I, and Mg II, and (b) K I, O I, and C I.

intensities to the two-variable set composed of those of K I and O I: (K, O, Mg), (K, O, C), (K, O, Ca), (K, O, H), and (K, O, Na). Their classification performances were evaluated to select the third variable for the best variable set (Fig. 2c). Among the five three-variable models, those of (K, O, Mg) and (K, O, C) showed the equally highest performances with the classification accuracy of 94.4% outperforming the best two-variable model (93.6%). To the former three-variable model of (K, O, Mg), each of the C I, Ca II, H I, and Na I emission peak intensities were added to form four-variable models. However, none of them showed improved classification accuracy in comparison with that of the three-variable model of (K, O, Mg) (Fig. 2d). The other three-variable model of (K, O, C) also could not be extended to any four-variable models which outperform it (Fig. 2d). Therefore, the forward selection process was stopped here with the two variable sets of (K, O, Mg) and (K, O, C) as the best standalone LIBS models. Fig. 3 shows the confusion matrices of the two best standalone LIBS models. The optimal *k* values were comparably 8 and 7 for the models of (K, O, Mg) and (K, O, C), respectively. Although the overall classification accuracy values of the two (K, O, Mg) and (K, O, C) models were equal, the former showed much better performance in classifying the kimchi samples produced in China and the latter provided rather balanced classification accuracy between the South Korean and Chinese kimchi classes.

### 3.2 IR modeling

The IR spectra of the kimchi samples produced in South Korea and China in the IR region of 450–4000 $cm^{-1}$ are shown in Fig. 4. Each spectrum represents the average of 53 and 72 spectra for the samples from South Korea and China, respectively. The IR spectra for both kimchi classes exhibit similar features, primarily due to their cabbage content. The absorption band in the spectral region 3100–3700 $cm^{-1}$ is attributed to O–H and N–H stretching vibrations.[30] The absorption peaks at 2925 $cm^{-1}$ and 2853 $cm^{-1}$ observed in both classes correspond to asymmetrical and symmetrical stretching of C–H bonds in lipids and other organic compounds.[31] The absorption peak at 2321 $cm^{-1}$ is attributed to the characteristic asymmetric stretching band of $CO_2$.[32] The absorption band observed at ~1734 $cm^{-1}$ is linked to the stretching vibration of C=O, indicating the presence of organic acids, primarily lactic acid.[33] The typical amide I, II, and III absorption bands were observed

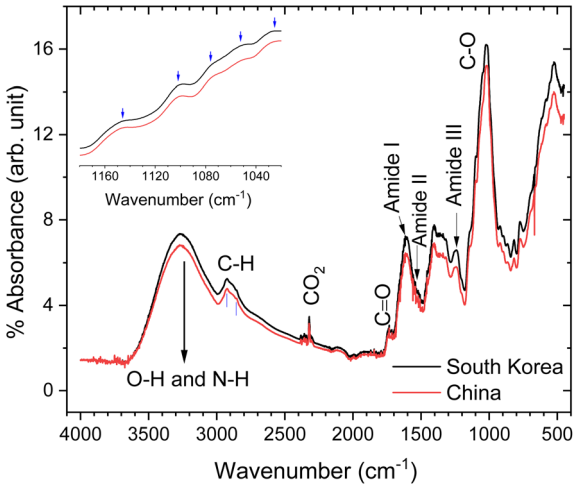


Fig. 4 Representative IR spectra of kimchi samples produced in South Korea and China in the IR range of 4000–450 $cm^{-1}$. The inset shows the expanded spectra in the region of 1180–1020 $cm^{-1}$ where the absorptions bands of carbohydrates are observed and indicated by the arrows.

in both kimchi spectra at ~1616 $cm^{-1}$ (C=O stretching), 1541 $cm^{-1}$ (N–H deformation), and 1250 $cm^{-1}$ (N–H deformation), respectively.[34] These amide absorption bands indicate the contribution of proteins or peptides to the IR spectra. Additionally, the carbohydrate absorption bands observed in the 1200–900 $cm^{-1}$ region arise predominantly from the combination of C–O–C and C–OH stretching at 1028, 1051, 1075, 1101, and 1146 $cm^{-1}$, as indicated by the arrows in the inset of Fig. 4.[35,36]

To identify a suitable range having significant discriminating power within the spectral region of 750–1750 $cm^{-1}$, the interclass distances between the two kimchi classes were calculated and compared at each wavenumber for both raw and preprocessed IR spectra. The interclass distance is actually a measure of separation between distinct classes and plays a crucial role in classification and clustering analyses. The larger interclass distances indicate the better separation between classes along a certain variable axis. Relying on interclass distances, variables with high discriminating power can be identified. The comparison of interclass distances at each wavenumber for both raw and preprocessed IR spectra offers two main advantages: (i) it demonstrates the superior class-separation capability of the preprocessed spectra compared to the raw spectra, and (ii) it enables the identification of the optimal spectral region within the preprocessed IR spectra that provides the highest discriminatory power between the two classes. The interclass distance, $d_{K-C}$, between the two kimchi sample classes was calculated using eqn (1) at each wavenumber from 750 to 1750 $cm^{-1}$.[14]

$$d_{K-C} = \frac{|m_K - m_C|}{s_{pooled}} \quad (1)$$

In the above equation, the subscripts "K" and "C" represent the two sample classes—kimchi produced in South Korea and China, respectively, and $m_K$ and $m_C$ denote the means of the

 

absorbance values for the South Korean and Chinese classes, respectively. The absolute difference, $|m_K - m_C|$, is simply the mean-to-mean distance between two classes. However, this is insufficient to describe how far the two classes are separated along the given variable axis because each class has its own variance. This can be addressed by scaling the mean-to-mean distance by the pooled standard deviation, $s_{pooled}$, calculated as the square root of a weighted average combining variances across several classes, as expressed in eqn (2).[37]

$$s_{pooled} = \sqrt{\frac{(n_K - 1){s_K}^2 + (n_C - 1){s_C}^2}{n_K + n_C - 2}} \quad (2)$$

In eqn (2), $n_K$ and $n_C$ represent the numbers of objects in the South Korean and Chinese kimchi classes, respectively, while $s_K$ and $s_C$ denote the standard deviations of the corresponding classes. Fig. 5a and b illustrate wavenumber-dependent interclass distances between South Korean and Chinese kimchi classes for both raw and preprocessed IR spectra, respectively. It is evident that for each wavenumber, the value of $d_{K-C}$ for the raw spectra is far below 1 (Fig. 5a). Contrary to the raw spectra, the $d_{K-C}$ values show significant improvement across most wavenumbers for the pre-processed IR spectra. This emphasizes how crucial spectral preprocessing is when working with complex IR data. For the preprocessed spectra, the $d_{K-C}$ values are found to be particularly high between 1254 and 1018 $cm^{-1}$. The interclass distance exceeds 1 for several bands in this spectral range, indicating a greater potential for distinguishing between the two kimchi classes compared to other ranges. Thus, modeling with the data in this spectral range would be essential. The shaded region in Fig. 5b shows relatively high discriminative power with $d_{K-C}$ values greater than 1. This shaded spectral range 1260–1010 $cm^{-1}$ is mainly associated with the carbohydrate absorption bands in the IR spectra of kimchi samples.

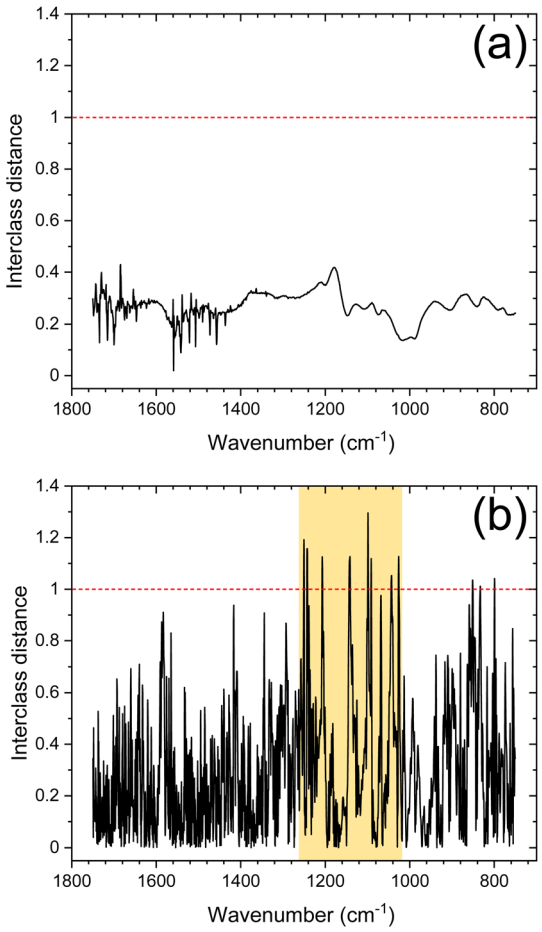


Fig. 5 Interclass distance values between the two sample classes of kimchi produced in South Korea and China using (a) raw and (b) preprocessed IR spectra in the region of 1750–750 $cm^{-1}$. The horizontal dashed lines indicate the interclass distance value of 1. In the yellow-shaded region in b, absorption bands with interclass distance values greater than 1 were observed.

The variables of the preprocessed IR spectra in the selected region (1254–1018 $cm^{-1}$) were reduced using PCA, and the discrimination capabilities of the resulting principal components (PCs) were subsequently evaluated. The first three PCs together explained approximately 91% of the total spectral variance. PCA extracts features in descending order of the variance they explain, prioritizing components that account for greater variability in the dataset. However, since this variance is calculated using all spectra in an unsupervised manner, the first principal component is not necessarily the most effective for classification purposes. Fig. 6a–c show the histograms of scores for the first three PCs. The score distributions reveal that PC1 provide almost no noticeable separation between Korean and Chinese kimchi samples, PC2 shows slight separation, and PC3 exhibits the greatest distinction between the two classes. This observation suggests that PC3 is the most suitable component for data fusion with the LIBS variables. The discrimination capabilities of the three PCs were quantitatively assessed based on their interclass distance values calculated using eqn (1) and (2). The interclass distances for PC1, PC2, and PC3 were found to be 0.05, 0.67, and 1.47, respectively, as shown in Fig. 6d. Accordingly, PC3, derived from the preprocessed IR data, was selected for inclusion in the fused LIBS-IR model. Prior to data fusion, the standalone IR *k*-NN model were independently trained using the PC3 scores from the IR spectra and evaluated by LOOCV. The standalone IR *k*-NN model achieved the highest classification accuracy of 86.4% at $k = 5$. The corresponding confusion matrix is shown as the inset in Fig. 6d.

### 3.3 LIBS-IR fused model

The variables extracted from the LIBS and IR spectra of kimchi samples were combined to improve the classification performance of the *k*-NN model. From the LIBS data, two variable sets, (i) the emission intensities of K I, O I, and Mg II, and (ii) those of K I, O I, and C I, were found to yield the equally highest classification accuracies of 94.4% at $k = 8$ and 7, respectively, among the standalone LIBS *k*-NN models. The standalone IR *k*-NN model trained using the PC3 scores from the IR spectra achieved the highest classification accuracy of 86.4% at $k = 5$. Accordingly, two fused LIBS-IR *k*-NN models were developed using (i) the LIBS emission intensities of K I, O I, and Mg II combined with the IR PC3 scores, and (ii) the LIBS emission

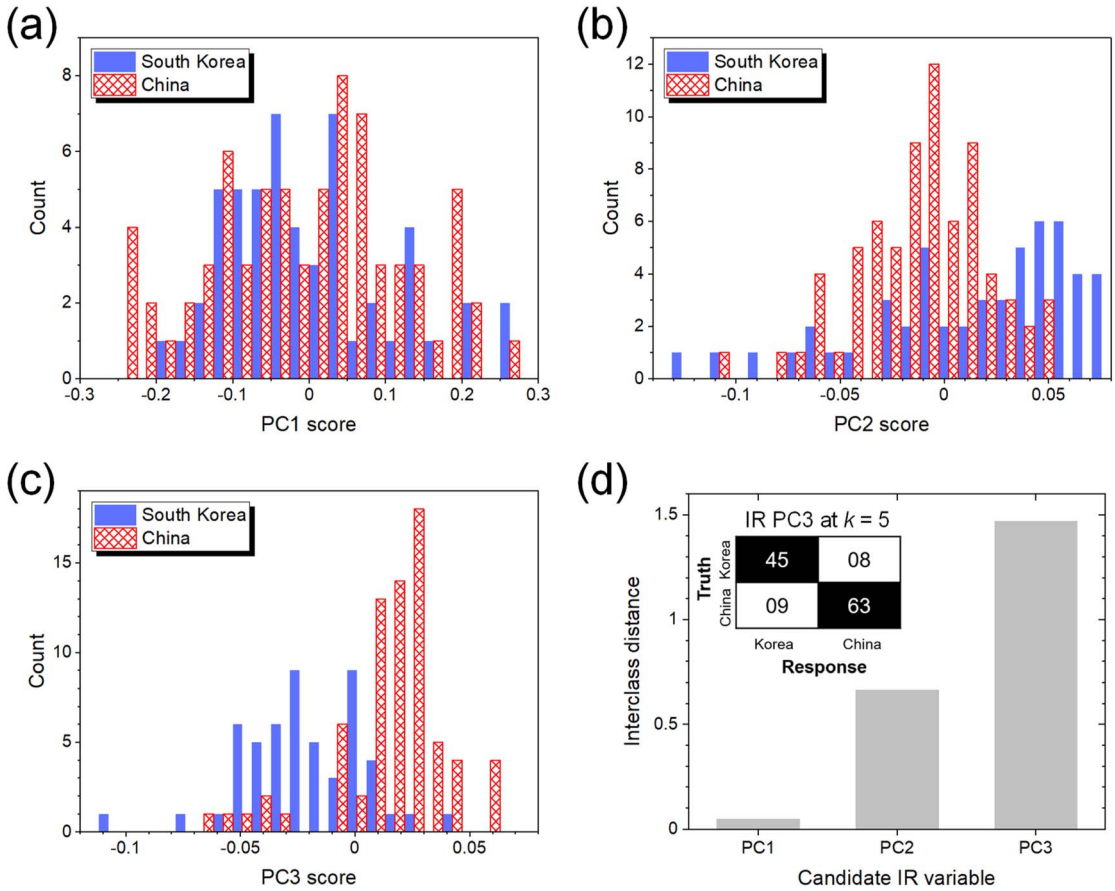


Fig. 6 Histograms of scores for (a) PC1, (b) PC2, and (c) PC3 extracted from the preprocessed IR spectra in the range of 1018–1254 $cm^{-1}$ and (d) the interclass distance values calculated for PCs 1–3. The confusion matrix of the standalone IR *k*-NN model trained using the PC3 scores extracted from the preprocessed IR spectra is shown as the inset in d.

intensities of K I, O I, and C I combined with the IR PC3 scores. However, the classification accuracies of both fused models were identical to those of the corresponding standalone LIBS models, showing the same optimal *k* values, and producing identical confusion matrices. This similarity is primarily attributed to the dominance of the score values of LIBS variables ($10^6$ to $10^7$) over those of PC3 from IR spectra (−0.1 to +0.1) and hinders the effective utilization of the discrimination power of the IR spectra. Therefore, appropriate scaling of data from different spectroscopic techniques is essential for effective combination.[38] In order for the optimal scaling, weighting factors ranging from $10^1$ to $10^9$ were applied to the PC3 scores from the IR spectra, and the classification accuracy values were investigated. Fig. 7a and b illustrate the classification accuracies of the fused LIBS-IR *k*-NN models trained using the LIBS emission intensities of K I, O I, and Mg II combined with the IR PC3 scores, and the LIBS emission intensities of K I, O I, and C I combined with the IR PC3 scores, respectively, as a function of the weighting factors applied to the IR PC3 scores. In these figures, the classification accuracy values of the standalone LIBS and IR *k*-NN models are also shown at the two extremes for comparison, and the accuracy values of the fused models lie between these extreme results. In both fused models, the classification accuracy of the standalone LIBS *k*-NN models and the fused LIBS-IR *k*-NN models remained the same until the weighting factors assigned to the IR variable became significantly large. The fused model combining the IR variable with the LIBS emission intensities of K I, O I, and Mg II showed the higher classification accuracy (96.0%) than those of the corresponding standalone LIBS and IR models with the weighting factors of $10^7$ to $5 \times 10^7$ applied to the IR variable (Fig. 7a). Also, the fused model combining the IR variable with the LIBS emission intensities of K I, O I, and C I showed the higher classification accuracy (96.8%) than those of the corresponding standalone LIBS and IR models with the weighting factors of $10^6$ to $5 \times 10^7$ applied to the IR variable (Fig. 7b). As the weighting factors increase further, the classification accuracy converged to that of the standalone IR *k*-NN models (86.4%) for both fused models.

It should be noted that the optimal weighting factors can be rationalized by considering the difference in the magnitudes of the LIBS and IR variables. *k*-NN model measures distance between a test object and training ones in the variable space, and with multiple variables, the distances along different variables are combined following Euclidean distance formulation as expressed by eqn (3).

 

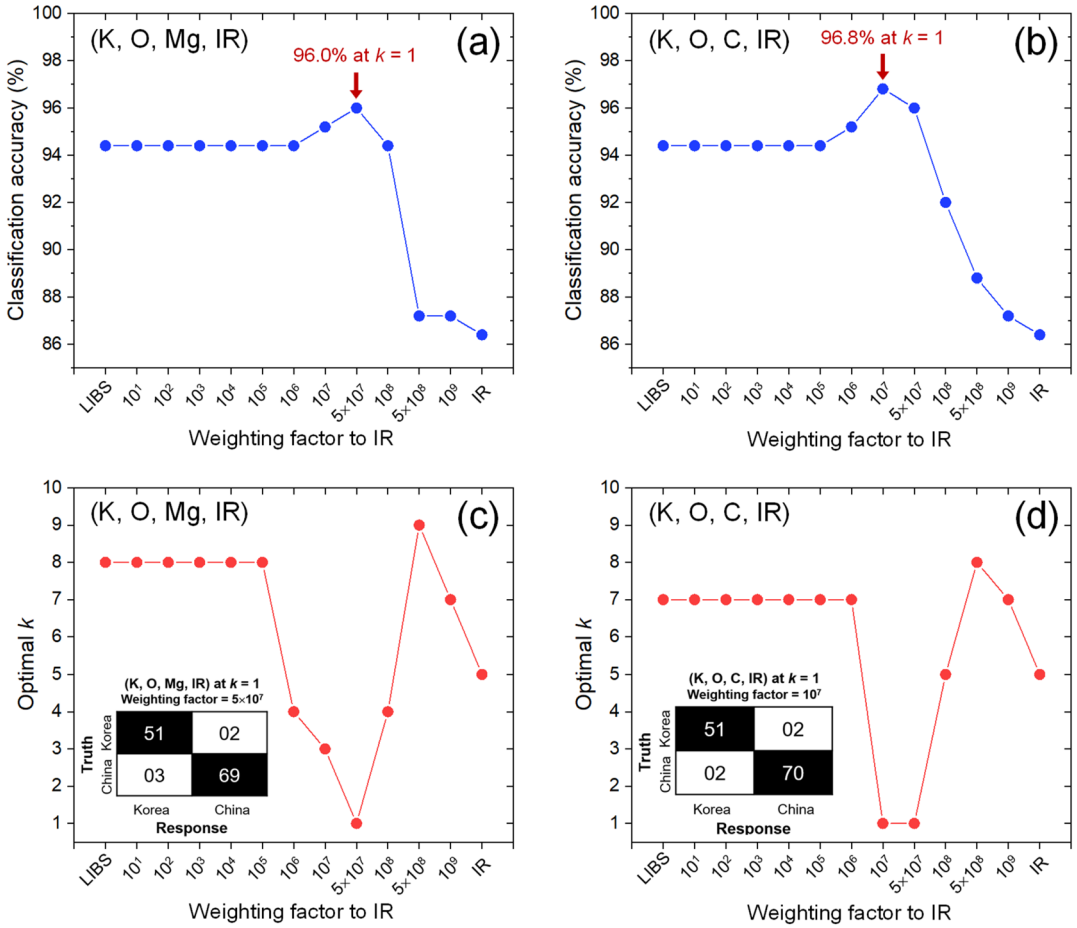


Fig. 7 Classification accuracy of fused LIBS-IR $k$-NN models trained using (a) the LIBS emission intensities of K I, O I, and Mg II and the PC3 scores from IR spectra and (b) the LIBS emission intensities of K I, O I, and C I and the PC3 scores from IR spectra and optimal $k$ values of the corresponding models with respect to the weighting factors given to the IR variable. The highest classification accuracy values are indicated by the arrows in a and b. The confusion matrices corresponding to the best fused LIBS-IR $k$-NN models trained using the two variable sets are shown as the insets in (c) and (d).

$$D = \sqrt{(v_1^{TR} - v_1^{TE})^2 + (v_2^{TR} - v_2^{TE})^2 + (v_3^{TR} - v_3^{TE})^2 + \cdots} \quad (3)$$

In a given variable space, training and test objects are represented by ($v_1^{TR}$, $v_2^{TR}$, $v_3^{TR}$,$\cdots$) and ($v_1^{TE}$, $v_2^{TE}$, $v_3^{TE}$,$\cdots$), respectively. This equation indicates that the discrimination power of $i$th variable in the multi-variable $k$-NN model originates from the absolute difference between $v_i^{TR}$ and $v_i^{TE}$ with $i$ = 1, 2, 3, … ($|v_i^{TR} - v_i^{TE}| = \sqrt{(v_i^{TR} - v_i^{TE})^2}$). Thus, the difference can be estimated by obtaining the average difference of the two scores for all possible combinations. Totally, 125 kimchi samples were used in this work. Thus, for each variable, 7750 (=125 × 124/2) score pairs can be taken for calculating the corresponding distances. The average of the 7750 distance values corresponding to the LIBS emission intensities of K I, O I, Mg II, and C I are 5.57 × $10^5$, 1.93 × $10^5$, 3.09 × $10^6$, and 1.08 × $10^5$, respectively, and that of the PC3 scores extracted from the IR spectra is 3.00 × $10^{-2}$. As can be seen from these calculations, the magnitude of the average distance of the IR variable scores is almost negligible compared to those of the LIBS variables. Under these circumstances, the distance along the IR variable can be combined as-is with those along the LIBS variables, but its contribution to the resulting distance in the fused-variable space is nearly negligible. Therefore, weighting factors are necessary to make the discrimination power from the IR variable effective in the fused models. However, the optimal weighting factor can be predicted by taking the ratio of the distance along the LIBS variable to that along the IR variable. For the fused model based on the LIBS emission intensities of K I, O I, and Mg II combined with the IR PC3 scores, this ratio was estimated to be 4.27 × $10^7$. For this calculation, the average distance of the three LIBS variables was used as a representative of the LIBS variables. The ratio of 4.27 × $10^7$ accurately predicted the optimal weighting factor to be applied to the IR variable, 5.00 × $10^7$, making the distance along the IR variable comparable to those along the LIBS variables. Likewise, for the fused model based on the LIBS emission intensities of K I, O I, and C I combined with the IR PC3 scores, this ratio was estimated to be 9.53 × $10^6$, which is very close to the optimal weighing factor, 1.00 × $10^7$.

Fig. 7c and d show the optimal $k$ values leading to the highest classification accuracy of the standalone LIBS and IR $k$-NN

models (at the leftmost and rightmost sides, respectively), as well as those of the LIBS-IR fused *k*-NN models across a range of weighting factors assigned to the IR variable. These figures highlight not only the classification performance but also the computational behavior of the fused models. When the weighting factor assigned to the IR variable is small, its influence on the overall distance metric is negligible due to its much smaller magnitude compared to the LIBS variables. Consequently, the fused models behave similarly to the standalone LIBS models, and the optimal number of neighbors, *k*, remains unchanged. However, as the weighting factor increases and the scaled IR variable begins to contribute meaningfully to the distance calculation, the class separation in the fused feature space becomes sharper. This enhanced separation allows for accurate classification using only the nearest neighbor ($k = 1$), resulting in the highest observed classification accuracy. Notably, this point of optimal class separation also corresponds to the lowest amount of computation required, since the *k*-NN model requires only one distance comparison to make a prediction. Beyond this optimal weighting range, further increases in the IR weighting factor cause the model to resemble the standalone IR model, which has lower discriminative power. Accordingly, the classification accuracy drops, and the optimal *k* value increases again, indicating a need to average over more neighbors to maintain performance. These results underscore the dual benefit of proper weighting: it enhances classification accuracy while simultaneously reducing the amount of computation required. The confusion matrices for the best-performing fused models—those using the LIBS emission intensities of K I, O I, and Mg II combined with the IR PC3 scores, and those using K I, O I, and C I with IR PC3—are shown as the insets in Fig. 7c and d, respectively.

## 4. Conclusions

This study demonstrated the effectiveness of combining elemental and molecular information obtained through LIBS and IR spectroscopy, respectively, for the accurate classification of kimchi products by country of origin. By selecting optimal variables from both spectroscopic techniques based on interclass distance and PCA, high-performing standalone models were constructed. The LIBS-based *k*-NN models achieved classification accuracies of up to 94.4%, while the IR-based *k*-NN model using PC3 reached 86.4%. To further improve classification performance, the variables from both techniques were fused within a *k*-NN modeling framework. However, due to the significant difference in magnitudes between the LIBS and IR variables, simple combination without scaling led to fused models that failed to outperform the LIBS and IR standalone models. This challenge was effectively addressed by introducing weighting factors to the IR variable, allowing its discrimination power to contribute meaningfully to the fused variable space. As a result, the classification accuracy of the fused models improved, with the best-performing model reaching 96.8%, surpassing both standalone models. Importantly, this study highlights that balancing the discrimination power of variables from heterogeneous spectroscopic sources is crucial not only for enhancing classification accuracy but also for improving computational efficiency in *k*-NN modeling. Specifically, with appropriate weighting, the optimal number of neighbors to be considered, *k*, in the fused models decreased significantly (down to $k = 1$), which in turn drastically reduced the amount of computation required for class assignment. This is particularly advantageous when working with large-scale datasets or implementing real-time classification systems. Furthermore, we proposed a rational method to predict the optimal weighting factor based on the average distance between sample pairs for each variable. This predictive approach aligns well with the empirically determined optimal weights and offers a generalizable strategy for future fusion modeling efforts involving disparate spectroscopic techniques. Overall, the proposed LIBS-IR fusion framework provides a robust, efficient, and accurate solution for verifying the geographical origin of kimchi and serves as a valuable model for broader applications in food authentication and multimodal data integration.

## Data availability

The data supporting this article are provided in the ESI.† Specifically, emission peak intensities for Mg II, K I, Ca II, C I, Na I, O I, and H I from the LIBS spectra, as well as the scores of principal components 1, 2, and 3 extracted from the IR spectra (1018–1254 $cm^{-1}$), are included.

## Author contributions

Sandeep Kumar: formal analysis, methodology, investigation, writing – original draft. Yujin Oh: formal analysis, investigation. Hyemin Jung: investigation, data curation. Kyung-Sik Hma: validation, project administration. Hyun-Jin Kim: validation, methodology. Song-Hee Han: resources, validation. Sang-Ho Nam: resources, supervision, funding acquisition, project administration. Yonghoon Lee: conceptualization, formal analysis, funding acquisition, methodology, supervision, writing – review & editing.

## Conflicts of 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.

## Acknowledgements

This research was supported by Korea Basic Science Institute (KBSI) National Research Facilities & Equipment Center (NFEC) grant funded by the Korean Government (Ministry of Education) (No. 2019R1A6C1010005 and 2023R1A6C103A019) and Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1F1A1063812). Prof. Yonghoon Lee gratefully acknowledges Ms. Hyang Kim for her technical assistance.

 

## References

1 J. H. Chang, Y. Y. Shim, S. K. Cha and K. M. Chee, *J. Appl. Microbiol.*, 2010, **109**(1), 220–230.
2 K. Y. Park and J. Ju, Kimchi and its health benefits, in *Korean Functional Foods: Composition, Processing and Health Benefits*, ed. K. Y. Park, D. Y. Kwon, K. W. Lee and S. Park, CRC Press, Boca Raton, FL, USA, 1st edn, 2018, ch. 3, pp. 43–78.
3 R. Surya, A. G. Y. Lee and J. Ethn, *Foods*, 2022, **9**(1), 20.
4 T. Yu, E. S. Park, X. Zhao, R. K. Yi and K. Y. Park, *RSC Adv.*, 2020, **10**(9), 5351–5360.
5 H. S. Cheigh, K. Y. Park and C. Y. Lee, *Crit. Rev. Food Sci. Nutr.*, 1994, **34**(2), 175–203.
6 B. Kim, K. Y. Park, H. Y. Kim, S. C. Ahn and E. J. Cho, *Food Sci. Biotechnol.*, 2011, **20**, 643–649.
7 B. Kim, J. L. Song, J. H. Ju, S. A. Kang and K. Y. Park, *Food Sci. Biotechnol.*, 2015, **24**, 629–633.
8 M. S. Islam and H. Choi, *J. Med. Food.*, 2009, **12**(2), 292–297.
9 K. H. Lee, J. L. Song, E. S. Park, J. Ju, H. Y. Kim and K. Y. Park, *Prev. Nutr. Food Sci.*, 2015, **20**(4), 298–302.
10 B. K. Kim, J. M. Choi, S. A. Kang, K. Y. Park and E. J. Cho, *Nutr. Res. Pract.*, 2014, **8**(6), 638–643.
11 E. Huang, *Quartz*, 2022, **https://qz.com/1183632/the-kimchi-you-eat-outside-of-korea-is-probably-made-in-china**, accessed June 2025.
12 M. N. M. Fairulnizal, B. Vimala, D. N. Rathi and M. M. Naeem, Atomic absorption spectroscopy for food quality evaluation, in *Woodhead Publishing Series in Food Science, Technology and Nutrition, Evaluation Technologies for Food Quality*, ed. J. Zhong and X. Wang, Woodhead Publishing, Kuala Lumpur, Malaysia, 2019, ch. 9, pp. 145–173.
13 S. H. Hur, H. Kim, Y. K. Kim, J. M. An, J. H. Lee and H. J. Kim, *Food Chem.*, 2023, **423**, 136235.
14 H. Kim, S. Jeong, H. Chung, S. H. Nam and Y. Lee, *J. Food Compos. Anal.*, 2023, **123**, 105501.
15 Q. Jin, F. Liang, H. Zhang, L. Zhao, Y. Huan and D. Song, *TrAC, Trends Anal. Chem.*, 1999, **18**(7), 479–484.
16 C. A. Bizzi, M. F. Pedrotti, J. S. Silva, J. S. Barin, J. A. Nóbrega and E. M. M. Flores, *J. Anal. At. Spectrom.*, 2017, **32**(8), 1448–1466.
17 R. E. Russo, X. Mao, J. J. Gonzalez, V. Zorba and J. Yoo, *Anal. Chem.*, 2013, **85**(13), 6162–6177.
18 F. J. Fortes, J. Moros, P. Lucena, L. M. Cabalín and J. J. Laserna, *Anal. Chem.*, 2013, **85**(2), 640–669.
19 S. Kumar, H. Choi, H. Chae, H. Kim, S. H. Nam, H. Kim, H. Kim, S. H. Han and Y. Lee, *J. Food Compos. Anal.*, 2024, **136**, 106742.
20 S. Jeong, D. Seol, H. Kim, Y. Lee, S. H. Nam, J. M. An and H. Chung, *Food Chem.*, 2023, **399**, 133956.
21 H. Wang, P. Zhang, Z. Xu, W. Cheng, X. Li, Y. Yang, Y. Wu and Q. Wang, *Microwave Opt. Technol. Lett.*, 2023, **65**(5), 1176–1185.
22 C. Eum, D. Jang, J. Kim, S. Choi, K. Cha and H. Chung, *Spectrochim. Acta, Part B*, 2018, **149**, 281–287.
23 P. Lasch, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 100–114.
24 M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys and C. Hughes, *Nat. Protoc.*, 2014, **9**, 1771–1791.
25 D. T. Larose and C. D. Larose, K-nearest neighbor algorithm, in *Discovering Knowledge in Data: an Introduction to Data Mining*, ed. D. T. Larose and C. D. Larose, Wiley, 2nd edn, 2014, ch. 7, pp. 149–164.
26 T. T. Wong, *Pattern Recognit.*, 2015, **48**(9), 2839–2846.
27 A. Kramida, *et al.*, *NIST ASD Team, NIST Atomic Spectra Database (Version 5.12)*, 2020, **https://www.nist.gov/pml/atomic-spectra-database**, accessed November 2024.
28 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
29 G. Borboudakis and I. Tsamardinos, *J. Mach. Learn. Res.*, 2019, **20**, 276–314.
30 J. B. Yang, C. W. Du, Y. Z. Shen and Z. J. Min, *Chin. J. Anal. Chem.*, 2013, **41**(8), 1264–1268.
31 B. B. Doyle, E. G. Bendit and E. R. Blout, *Biopolymers*, 1975, **14**(5), 937–957.
32 E. P. Barannik, A. A. Shiryaev and T. Hainschwang, *Diamond Relat. Mater.*, 2021, **113**, 108280.
33 P. N. Devi, J. Sathiyabama and S. Rajendran, *Int. J. Corros. Scale. Inhib.*, 2017, **6**(1), 18–31.
34 Y. Sun, N. Liu, L. Zhao, Q. Liu, S. Wang, G. Sun, Y. Zhao, D. Zhou and R. Cao, *Microchem. J.*, 2024, **204**, 111037.
35 M. Bağcıoğlu, B. Zimmermann and A. Kohler, *PLoS One*, 2015, **10**(9), e0137899.
36 N. Sharma, Y. Khajuria, V. K. Singh, S. Kumar, Y. Lee, P. K. Rai and V. K. Singh, *At. Spectrosc.*, 2020, **41**(3), 110–118.
37 International Union of Pure and Applied Chemistry, *IUPAC Gold Book pooled standard deviation*, 2005, **https://goldbook.iupac.org/terms/view/P04758**, accessed August 2024.
38 J. Park, S. Kumar, S. H. Han, S. H. Nam and Y. Lee, *Spectrochim. Acta, Part B*, 2021, **179**, 106088.