ACS FOOD SCIENCE & TECHNOLOGY

Open Access

pubs.acs.org/acsfoodscitech

Article

Rapid and Nondestructive Determination of Oil Content and Distribution of Potato Chips Using Hyperspectral Imaging and Chemometrics

Yue Sun, Nikhita Sai Nayani, Yixiang Xu, Zhanfeng Xu, Jun Yang, and Yiming Feng*

Cite This: *ACS Food Sci. Technol.* 2024, 4, 1579−1588

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Conventional techniques used to measure oil content in the food are laborious, rely on chemical agents, and have a negative environmental impact. In this study, near-infrared hyperspectral imaging was used as a rapid and nondestructive tool to determine the oil content and its distribution in commercial flat-cooked and batch-cooked potato chips. By evaluating various algorithmic models, such as partial least-squares regression (PLSR), ridge regression, random forest, gradient boosting, and support vector regression, in combination with preprocessing methods like multiplicative scattering correction, standard normal variable (SNV) transform, Savitzky−Golay filtering, normalization, and baseline correction, the most effective preprocessing method and model combination was determined to be SNV-PLSR. Moreover, by employing the optimized PLSR model, a highly accurate oil content prediction model was developed, achieving a coefficient of determination (R^2) of 0.95. To identify the wavelengths that contributed most significantly to the model's predictive power, variable importance in projection (VIP) analysis was utilized. A dimensionally reduced PLSR model using only 68 selected wavelengths was developed based on the VIP analysis. This simplified model maintained similar performance to that of the full-spectrum model while using a smaller data set. The model was also used to apply the hyperspectral images of potato chips at the pixel level to visualize the oil distribution in potato chips with the intent to provide a real-time approach to quality control for the potato chip industry.

KEYWORDS: *hyperspectral imaging (HSI), potato chips, oil content distribution, nondestructive testing, machine learning (ML), food quality control*

1. INTRODUCTION

Potato chips, one of the most popular snacks across the world, had a global market size of $33.3 billion in 2022 and will potentially reach $40.0 billion in 2028.[1] It appeals to individuals of all ages, and its popularity is mostly attributed to its distinct flavor and convenient use.[2] Despite its popularity, the high oil content of potato chips, which contributes to almost 60% of their calories, causes health concerns including heart disease and obesity that are related to oil consumption.[3] Under the current trend toward healthy diets, it is important for the food industry to provide products with low oil content in response to market demand.[4] However, lowering the oil content in potato chips has been challenging due to the complex physical process of oil transportation during the frying process. The oil content in each piece of potato chip could also be affected by the potato chip size, thickness, frying time, oil temperature, chip pore characteristic, bulk density, and porosity,[5] resulting in high variability piece to piece. Therefore, the snack food industry has been looking for a nondestructive and real-time approach to monitor the oil content in potato chips to ensure consistency of product quality.

Soxhlet extraction and gas chromatography are traditional methods to determine the total lipid content of food products, but those analytical techniques can only be used on a laboratory scale and are time-consuming, destructive, and involve toxic and expensive chemical solvents,[6] which are not feasible for real-time monitoring of oil content in production lines.[7] To date, nondestructive and real-time monitoring technologies have been focused on near-infrared (NIR) spectroscopy, Raman spectroscopy (RS), Fourier-transform infrared (FTIR) spectroscopy, and hyperspectral imaging (HSI).[8,9] Among these methods, NIR, RS, and FTIR probe spectral information from a single point,[10] and they are incapable of analyzing spatial component distribution and heterogeneity to detect regional quality and safety issues such as molding, hygroscopicity, and oxidation.[11] Attributed to the capability of capturing spectral information at every pixel, HSI has been utilized in various food products including soybeans, rapeseed, and beef as a rapid and nondestructive for chemical and physical properties' analysis.[12−15] Despite HSI's promising features, challenges include large data volumes, spectral peak shifts, unstable output, and slow speed.[10,16,17] High-dimensional data sets can be difficult to manage and process efficiently, leading to increased computational complexity and

Received: March 21, 2024
Revised: May 22, 2024
Accepted: May 22, 2024
Published: June 3, 2024

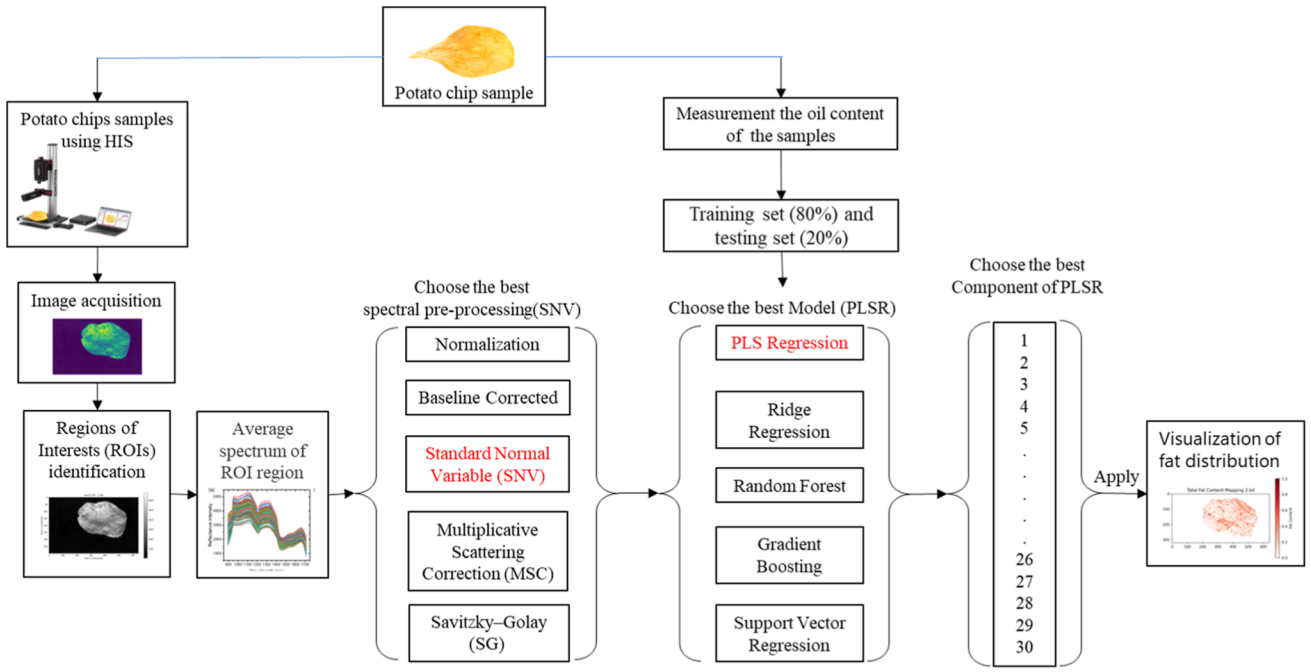

Figure 1. Process of HSI for predicting oil content in potato chip samples.

hindering real-time decision-making.[18] The high dimensionality of HSI data can also lead to collinearity, redundancy, and the presence of irrelevant or noisy variables, negatively impacting traditional data analysis methods and reducing the efficiency of quality prediction models.[19]

Combining HSI with preprocessing approaches and machine learning techniques offers a promising approach to address these limitations. Machine learning algorithms, such as partial least-squares regression (PLSR), support vector machines, and artificial neural networks, can effectively handle high-dimensional data and extract relevant features from the spectral information.[20] For instance, SVR has been successfully applied for predicting the quality of coffee and chicken meat,[21,22] random forest (RF) has demonstrated excellent performance in predicting carbonization characteristics of kraft lignin-derived hydrochar[23] and soil nitrogen and carbon measurements,[24] and SVR was effective in predicting crude protein content of alfalfa.[25] While some of these methods may be considered "black box" approaches due to their complex internal workings, their performance and suitability for various applications have been well-established through extensive research. Comparing linear and nonlinear methods allows for assessing the trade-off between model complexity and performance in predicting oil content in potato chips using HSI.

Despite the potential of HSI and machine learning algorithms in food quality assessment, limited studies have focused on estimating the oil content and distribution in potato chips. This study aims to address this gap by optimizing an HSI system for data collection, developing a preprocessing pipeline and comparing various machine learning models for predicting oil content in potato chips (Figure 1). The most robust model was identified, and its hyperparameters were tuned for optimal prediction accuracy. Finally, the model was applied to visualize the oil distribution in potato chips through pixel-by-pixel image reconstruction. This approach provides a comprehensive framework for utilizing HSI and machine learning to assess oil content and distribution in potato chips, contributing to the advancement of nondestructive quality control methods in the food industry.

2. MATERIALS AND METHODS

2.1. Materials. Nine different Potato Chips with different flavors were purchased through conventional retail channels, either local grocery stores (Hampton, VA, USA) or online shopping platforms (Amazon.com) as presented in Table 1. A total of 224 potato samples without visible damage or irregular shape were selected. The oil content of the selected products was calculated based on their nutritional labels. Hexane was obtained from Thermo-Fisher USA. Two types of Potato Chips have been used for this study, flat-cooked potato chips (conventional processing) and batch-cooked Potato Chip (kettle processing). The batch-cooked potato chips have been recently adapted by the industry in producing Potato Chips with reduced fat. Intact potato chips weighing between 0.5 and 2 g were used.

2.2. Determination of Total Oil Content. The oil content of potato chips was determined following a method described by Kadamne and Proctor.[26] Briefly, 224 pieces of potato chips (1.5−2 g) were weighed and recorded as w_1 (g). The potato chip was crushed and then placed into 50 mL conical centrifuge tubes (Corning Incorporated, Tewksbury, MA, USA) to mix with 30 mL of hexane. The solvent and chip mixture was then centrifuged at 5000 rpm for 10 min (Eppendorf, 5430, Hamberg, Germany), and the supernatant was decanted after centrifugation. The potato chips remaining in the centrifuge tube were placed in a chemical hood overnight until they reached a constant weight, w_2. The weight difference before and after the oil extraction will be used to calculate the total oil content (1).

$$\text{Total oil content (\%)} = \frac{w_1 - w_2}{w_1} \times 100\% \quad (1)$$

where w_1 is the weight of the sample before hexane extraction and w_2 is the weight of the sample after hexane extraction.

2.3. Spectral Acquisition. The spectra and images related to potato chips were acquired by using a line-scanning hyperspectral camera (Pika IR+. Resonon, MT, USA), covering the spectral range from 900 to 1700 nm (NIR to SWIR) with a spatial resolution of 640 pixels and a spectral resolution of 2.3 nm (344 spectral wavelengths in total). The hyperspectral camera was equipped with SpectrononPro

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

Table 1. Selected Potato Chips Samples' Information Including Labeled Oil Content and Flavors from Two Major US Brands

brand	brand I					brand II		
flavor of potato chips	flat potato chips (salt and vinegar)	flat potato chips (original)	flat potato chips (Barbeque)	flat potato chips (sour cream & onion)	batch-cooked potato chips (original)	batch-cooked potato chips (sweet mesquite barbeque)	batch-cooked potato chips (original sea salt)	batch-cooked potato chips (sea salt & vinegar)
oil content from nutrition label	35.34%	35.34%	31.8%	35.34%	17.86%	19.05%	21.43%	19.05%

(64bit, version 3.411) software for data and image collection, control of travel speed, and cube data processing. Dark/white correction was performed prior to the image collection. The dark chromatic reference was obtained by masking the camera with a lens cap, while the white reference was obtained with Spectralon standard material. The hyperspectral camera was configured with a scanning speed of 1.19 cm/s, a frame rate of 81.94 frames per second, and an integration time of 9.94 ms. Each hyperspectral image is a volumetric image cube with 400 × 640 pixels (x-dimension and y-dimension) and 334 spectral wavelengths (λ/z-dimension) in order to store spatial and spectral information about the sample.

2.4. Spectral Preprocessing. Before model construction, preprocessing of the average spectrum of all pixels within the identified region of interest (ROI) is necessary to reduce noise, light scattering, and other undesirable effects in the spectrum.[27] Commonly used preprocessing techniques include baseline correction, normalization, standard normal variable (SNV), multiplicative scattering correction (MSC), and Savitzky−Golay (SG) smoothing. Among these techniques, SNV and MSC are particularly effective in removing additive and multiplicative light scattering effects from inhomogeneous sample surfaces.[28] The preprocessing of all spectral images was performed in Python 3.11 in a Spyder 5.4.3 environment.

2.5. ROI Identification. To eliminate noise, remove redundant background, reduce boundary blur, and extract useful spectral information, we performed image segmentation. To extract the spectrum of each potato chip, a mask was generated under the maximum background discriminant metric using the Otsu threshold algorithm and morphological processing.[29] This process isolated the potato chips and generated an image containing only potato chips, avoiding any background interference. The pixels in the mask were then defined as ROIs, and the spectral data of these ROIs were extracted from the calibrated hyperspectral images. Through this segmentation technique, images were divided into various ROIs with similar properties, including nonredundant features that provide meaningful data.[30]

2.6. Data Modeling. *2.6.1. Chemometrics.* Description and features of five different machine learning regression models, namely, ridge regression (RR), RF, gradient boosting (GB), PLSR, and support vector regression (SVR) are presented in Table 2. Model building and hyperparameter tuning were performed using grid search for each of the five models.[31] The grid search technique involves defining a range of values for each hyperparameter and exhaustively evaluating all possible combinations to identify the optimal set of hyperparameters that yield the best model performance. The grid search was performed using a 5-fold cross-validation approach on the training set to ensure the robustness of the selected hyperparameters and to avoid overfitting. The best-performing hyperparameters for each model were then used to train the final models on the entire training set and evaluate their performance on the independent test set.

2.6.2. Model Evaluation. HSI spectroscopy (344 bands) was used to predict the oil content of the different potato chips. To evaluate the model, the data set (224 potato chips) was randomly divided into an 80% calibration set and a 20% validation set. The data were split using the "train test split" function in the Sklearn model. To verify the representativeness of the random sampling method, multiple splits using different random states were performed, and the consistency of model performance across these splits was checked. To avoid overfitting and optimize the model performance, the cross-validation method was used to analyze the calibration set and determine the ideal number of latent variables. In the process, the root-mean-square error (RMSE) was first calculated based on the cross-validation set, and then the most suitable number of PLSR factors was selected from the minimum RMSE. Next, external validation of the test set was performed to assess the model's predictive ability by calculating the predictive correlation coefficient, RMSE of prediction, and residual prediction deviation. The R^2 and RMSE values will be used as the key to determine the model prediction performance.[32]

2.6.3. Variable Importance Projection. Variable Importance in Projection (VIP) is a measure used to assess the importance of each

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

Table 2. Descriptions and Features of Five Different Machine Learning Regression Models[a]

model	equation	description	feature	references
PLSR	$X = TP^T + E$, $Y = TQ^T + F$	T is the latent component matrix, P^T and Q^T are the loading matrices, E and F are the residual terms	PLSR is widely used as a standard chemometric technique for predictive analysis of spectral data	44
RR	$\widehat{\beta}(k) = (x'x + kI)^{-1}$ $x'y$, $k \geq 0$	I is the identity matrix and k is the ridge parameter. $x'x$ is equivalent to adding the value of k to the diagonal elements of $x'x$	RR adds new factors to the least-squares objective to reduce the overfitting associated with linear regression	45
RF	$\widehat{m} = (x)\frac{1}{M}\sum_j \widehat{m}_j(x)$	$\widehat{m}_j$ is a tree estimator, M is a tuning parameter	RF model can be integrated to learn and process complex systems for efficient and accurate prediction	46
GB	$f(x) = \sum_{m=1}^{M} \beta_{jm} b(x; \tau_{jm})$	$b(x;\tau_{jm})$ is a weak parameter. τ and β_j the coefficient of the j weak learner. β_{jm} and τ_{jm} are as an adaptive fashion (can improve the data fidelity)	GB is based on the fact that it is possible to construct gradient-oriented learners in a short time to obtain good results	47
SVR	$f(x) = \phi(x)\, w + c$	x is a variable vector, ϕ is the nonlinear function, w is the weight vector, and c is a constant	SVR can be applied to minimize structural risk to ensure that the global optimum is achieved and generalization errors are reduced	48

[a]Abbreviations: regression (RR), random forest (RF), gradient boosting (GB), partial least squares regression (PLSR), and support vector regression (SVR).

variable (wavelength) in the PLSR model.[33] VIP scores summarize the contribution of each variable to the model, taking into account both the explained variance in the response variable (fat content) and the explained variance in the predictor variables (spectral data). Variables with higher VIP scores are considered more important for the model's predictive ability. VIP calculates the significance of each wavelength by using the VIP score, as shown in eq 1.

$$\mathrm{VIP}_i = \sqrt{\frac{a\sum_{n=1}^{N}(\omega_{in}^2 \cdot \mathrm{SS}_n)}{\sum_{n=1}^{N}\mathrm{SS}_n}} \quad (1)$$

In the equation, VIP_i represents the VIP score of the ith variable, where a is the number of input variables, N is the number of latent variables, ω_{in} is the weight of the ith input on the nth latent variable, and SS_n is the sum of squares of the nth latent variable.[33]

In this study, VIP scores were calculated for each wavelength using the optimized PLSR model. Wavelengths with VIP scores greater than 1 were considered important, as they contribute more than the average variable to the model's predictive ability. The stability of the VIP scores was tested for different random states. Once the important wavelengths are identified, a simplified PLSR model will be developed by using the selected wavelengths.

3. RESULTS AND DISCUSSION

3.1. Sample Characteristics. Oil content in each single piece of potato chip was determined by the methodology from the literature.[26] As expected, a large piece-by-piece variation in oil content was found for all of the potato chip brands. The average oil content varied from 8.13 to 25.66% among different products. A histogram of 224 pieces of potato chips from nine products is shown in Figure 2a, which indicates that the data were somewhat normally distributed. The most common range of oil content was between 10 and 15%. The wide distribution of oil content across each piece of potato chips was expected due to the nature of fried food products, and it also provides a wide range of training and testing data sets for this study. The results of each potato chip product (Figure 2b) agreed well with the oil content calculated from the nutritional labels.

3.2. Preprocessing. The HSI image data of each potato chip were collected before chemical analysis. A list of preprocessing algorithms was deployed to improve the data quality, including MSC, SNV, SG, baseline correction, and normalization. Table 3 shows a comparison of different preprocessing methods on the preliminary fitting results using the PLSR model (N = 10) as the universal fitting algorithm to simplify the comparison. The results demonstrate that SNV preprocessing yielded the best performance, with a high coefficient of determination (R^2) of 0.841 for the training data set (n = 179) and 0.863 for the test data set (n = 45), along with a low mean absolute error (MAE) of 0.029 and 0.026, respectively. Therefore, SNV was chosen as the standard preprocessing method for this study in the downstream machine learning tasks.

The effect of SNV preprocessing on the average spectra of the 224 potato chip samples indicated excellent consistency in scaling the spectra (Figure 3). Distinct features in characteristic peaks presented in the postprocessing spectra are in agreement with the literature.[34] Based on the chemical composition characteristics of the potato chip samples, absorption peaks, and valleys were observed in the spectrum with the main strong bands at the 950, 1100, 1200, 1450, and 1700 nm. The absorption near 1200 nm (C−H) is related to the oil content, and the presence of a wide absorption band near 1220 nm may be due to the secondary overtones of C−H and CH═CH stretching vibrations in the oil.[35] The absorption (−OH) near

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

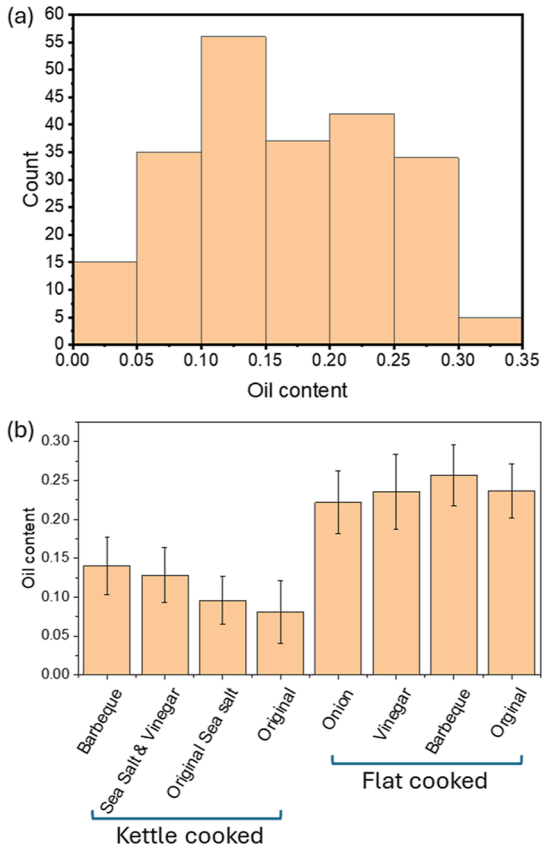

Figure 2. Oil content in potato chips. (a) Histogram of total oil content in potato chips. (b) Total oil content of potato chips of different products determined by chemical method (each error bar is constructed using standard deviation from the mean).

1450 nm was reported to be attributed to moisture.[36] The data suggest that the spectral features in this region are highly informative for predicting the oil content in potato chips. The prominence of these absorption bands highlights the importance of the NIR spectral region in capturing the chemical composition of the samples, particularly the oil-related compounds. The successful application of SNV preprocessing in enhancing the spectral data quality and the identification of key absorption bands related to oil and moisture content demonstrates the potential of HSI as a rapid and nondestructive tool for analyzing the chemical composition of potato chips. Through the SNV preprocessing, the data quality was improved, and the strong correlation between the spectral features and the chemical properties of interest lay a solid foundation for the development of robust prediction models using advanced machine learning techniques.

3.3. Comparison of Five Common Machine Learning Algorithms in Fitting the Sample. Five different regression models, namely, RR, RF, GB, PLSR, and SVR, were compared for their performance in predicting the oil content of potato chips based on the HSI data. The data set, consisting of 224 potato chip samples, was randomly divided into a training set (80%) and a testing set (20%) to evaluate the models' predictive capabilities.

Table 4 presents the best predictions for each learning method with different parameter settings. PLSR demonstrated the highest R^2 values among the models, indicating its superior performance in fitting the data. With 15 components, PLSR achieved an R^2 of 0.9243 for the training set and 0.8280 for the testing set, along with the lowest RMSE values of 0.0297 and 0.0352, respectively. The optimized full-wavenumber PLSR effectively captures the linear relationship between the near-infrared spectral data and the oil content in potato chips, enabling an accurate prediction of the relative oil content. These findings suggest that PLSR is a powerful tool for extracting relevant information from the spectral data and building a robust predictive model for oil content estimation.

RR also showed promising results, with the best performance observed when the regularization parameter (alpha) was set to 10. Under this setting, RR obtained an R^2 of 0.9843 for the training set and 0.8109 for the testing set, with corresponding RMSE values of 0.0168 and 0.0358. The relatively high performance of RR indicates that the regularization technique effectively mitigates overfitting and improves the model's generalization ability, yet the significantly higher R^2 and lower RMSE in the training data set than the testing data set suggest a slight overfitting. RF and GB exhibited moderate performance compared to PLSR and RR. The best results for RF were obtained with 100 estimators, achieving an R^2 of 0.9303 for the training set but only 0.5428 for the testing set. Similarly, the GB with 300 estimators reached an R^2 of 0.8964 for the training set and 0.5714 for the testing set. The discrepancy between the training and testing performance suggests potential overfitting issues with these ensemble models. While they can capture complex relationships in the training data, their generalization ability may be limited, leading to suboptimal performance on unseen data. SVR had the lowest performance among the compared models, with the best results obtained using a cost parameter of 10. Even with this setting, SVR only achieved an R^2 of 0.6445 for the training set and 0.6136 for the testing set, indicating its limited predictive power for this specific data set. The inferior performance of SVR compared to other models may be attributed to its sensitivity to the choice of kernel function and the difficulty in optimizing its hyperparameters for this particular application.

Table 3. Comparison of Different Pre-Processing Algorithms on the Preliminary Fitting Results[a]

preprocessing	PLSR component number	training R^2	testing R^2	training MAE	testing MAE	training RMSE	testing RMSE
raw	10	0.849	0.819	0.028	0.030	0.036	0.041
MSC	10	0.833	0.825	0.030	0.029	0.037	0.040
SNV	10	0.841	0.863	0.029	0.026	0.036	0.036
SG	10	0.825	0.859	0.027	0.026	0.034	0.036
normalized	10	0.855	0.839	0.028	0.031	0.036	0.039
baseline corrected	10	0.847	0.828	0.029	0.029	0.037	0.040

[a]Abbreviations: multiplicative scattering correction (MSC), standard normal variable transform (SNV), Savitzky−Golay (SG), mean absolute error (MAE), root mean squared error (RMSE).

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

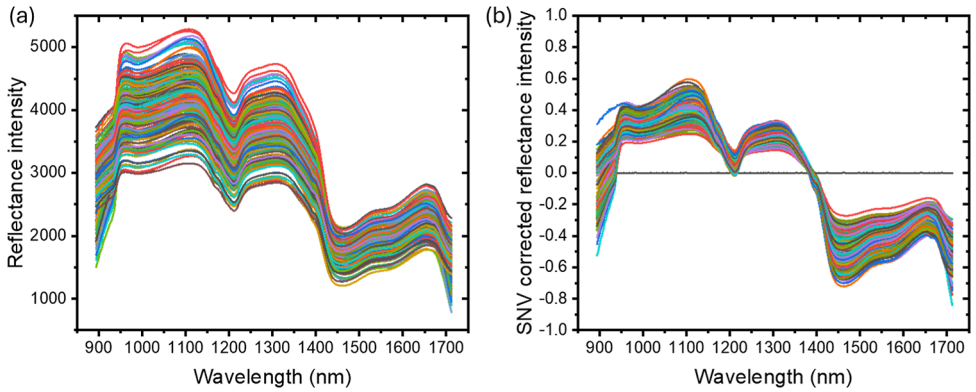

Figure 3. SNV transform preprocessing of HSI spectra of potato chips. (a) Raw spectrum and (b) SNV-processed spectra.

Table 4. Screening of Machine Learning Models for Oil Content Prediction[a]

model	parameters		training R^2	test R^2	training MAE	test MAE	training RMSE	test RMSE
PLSR	latent variables	5	0.7783	0.7540	0.0382	0.0349	0.0490	0.0459
		10	0.8794	0.8494	0.0285	0.0303	0.0365	0.0410
		15	0.9243	0.8280	0.0230	0.0252	0.0297	0.0352
RR	alpha	0.1	0.9971	0.7201	0.0016	0.0394	0.0057	0.0511
		1	0.9958	0.7517	0.0062	0.0329	0.0091	0.0439
		10	0.9843	0.8109	0.0128	0.0273	0.0168	0.0358
RF	estimators	10	0.9148	0.5014	0.0216	0.0512	0.0305	0.0646
		50	0.9258	0.5605	0.0182	0.0445	0.0245	0.0573
		100	0.9303	0.5428	0.0187	0.0470	0.0251	0.0600
GB	estimators	100	0.6647	0.4531	0.0484	0.0529	0.0615	0.0709
		200	0.8300	0.5520	0.0364	0.0482	0.0458	0.0650
		300	0.8964	0.5714	0.0299	0.0487	0.0376	0.0642
SVR	C	0.1	0.4282	0.4187	0.0595	0.0539	0.0756	0.0682
		1	0.5405	0.5239	0.0553	0.0482	0.0686	0.0610
		10	0.6445	0.6136	0.0484	0.0441	0.0586	0.0536

[a]Abbreviations: regression (RR), random forest (RF), gradient boosting (GB), partial least squares regression (PLSR), and support vector regression (SVR), mean absolute error (MAE), root mean squared error (RMSE).

Based on these results, PLSR emerged as the most suitable model for predicting the oil content of potato chips using HSI data. Its high R^2 values and low error metrics demonstrate its ability to effectively capture the underlying relationships between the spectral features and the oil content while maintaining a good generalization performance on unseen data. The success of PLSR can be attributed to its ability to handle high-dimensional spectral data, its robustness to collinearity, and its effectiveness in extracting latent variables that maximize the covariance between the spectral features and the response variable (oil content). The comparative analysis of different regression models highlights the importance of selecting an appropriate algorithm that aligns with the characteristics of the data and the problem at hand. While PLSR and RR showed superior performance in this study, it is essential to consider the trade-offs between model complexity, interpretability, and computational efficiency when deploying these models in real-world applications. Further research may explore the integration of feature selection techniques or the incorporation of domain knowledge to enhance the interpretability and scalability of the predictive models for oil content estimation in potato chips.

3.4. Hyper Tuning of the PLSR Model. The relationship between the performance of the PLSR model and the number of latent variables (number of components) is shown in Figure 4. Figure 4a shows the fitting performance as the number of components increases, the RMSE of the testing data set first decreases significantly, reaching a minimum at 14 components, and then increases. The trend is consistent with the R^2 plot using the training data set, which suggests that the model is best performing at n = 14, while incorporating too many components (>14) may lead the model to overfit or increase its complexity without further improving its performance.

The hypertuned PLSR model, with the optimal number of components, achieved an exceptional R^2 value of 0.95 when fitted to the test data set. This high R^2 value indicates the model's robustness and its ability to accurately predict the oil content of potato chips based on the HSI data. The close distribution of data points around the best-fit line in the scatterplot, without systematic deviations or obvious outliers, further confirms the stability and accuracy of the model at different levels of oil content. The outstanding performance of the optimized PLSR model can be attributed to several factors. First, PLSR is particularly well-suited for handling high-dimensional spectral data, as it projects the original variables into a lower-dimensional space of latent variables that maximize the covariance between the predictors and the response variable. This dimensionality reduction technique effectively captures the most relevant information from the spectral features while mitigating the effects of collinearity and noise.[37] Second, the rigorous optimization process, involving the selection of the appropriate number of components

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

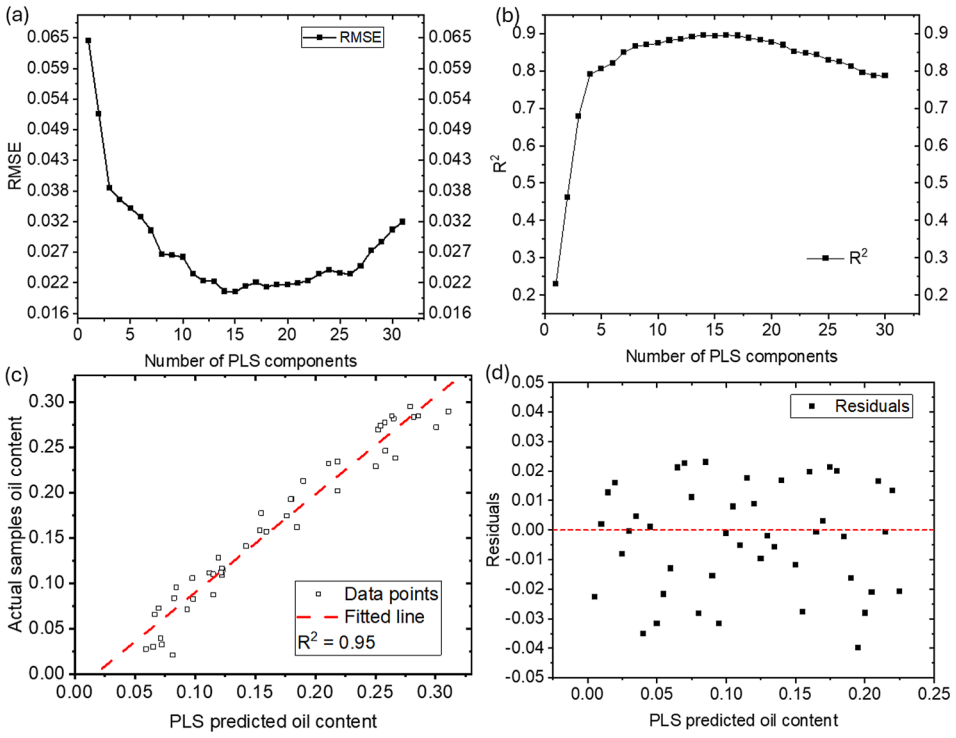

Figure 4. Fine-tuning of PLSR model to improve fitting performance. (a) RMSE as a function of the number of PLS components. (b) Relationship between PLSR model performance and the number of model components. (c) Correlation between actual and PLSR predicted oil content. (d) Residual plot of the test data set.

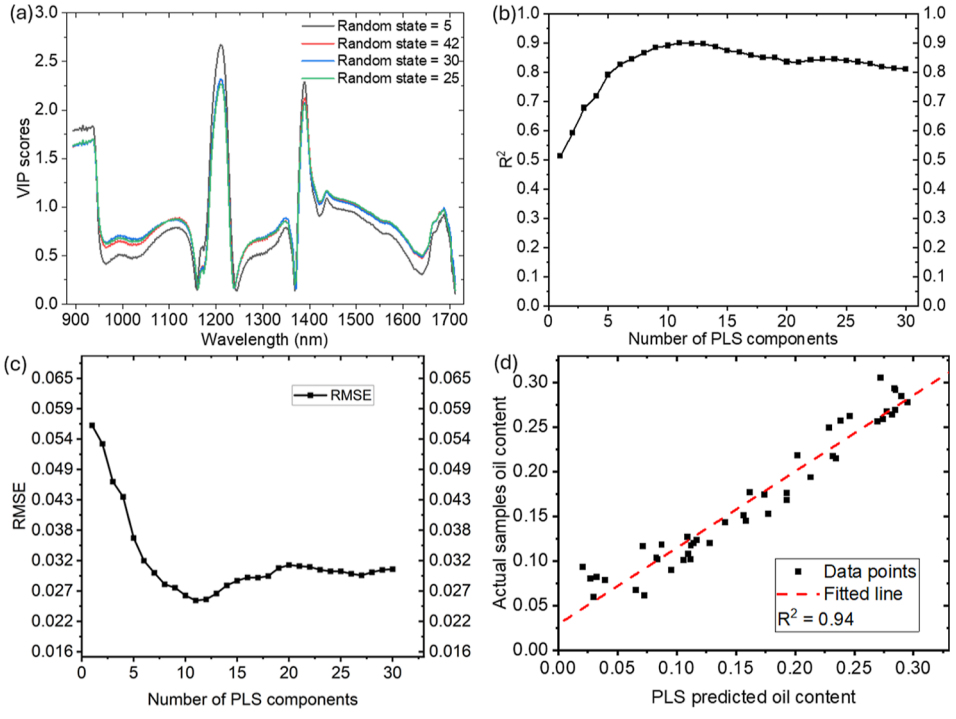

Figure 5. (a) VIP of different wavelengths at different random states. (b) R^2 of PLSR optimization using the dimension-reduced (68 wavelengths) data set. (c) RMSE of PLSR optimization using the dimension-reduced (68 wavelengths) data set. (d) Correlation between actual and the new PLSR predicted (n = 11) oil content.

through cross-validation, ensures that the model is not overfitting to the training data and can generalize well to unseen samples.[38] Lastly, the high R^2 value and the close agreement between the predicted and actual oil content values in the test data set demonstrate the model's ability to accurately capture the underlying relationships between the spectral features and the oil content. This strong predictive performance suggests that the optimized PLSR model, in combination with HSI data, has the potential to be a reliable

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

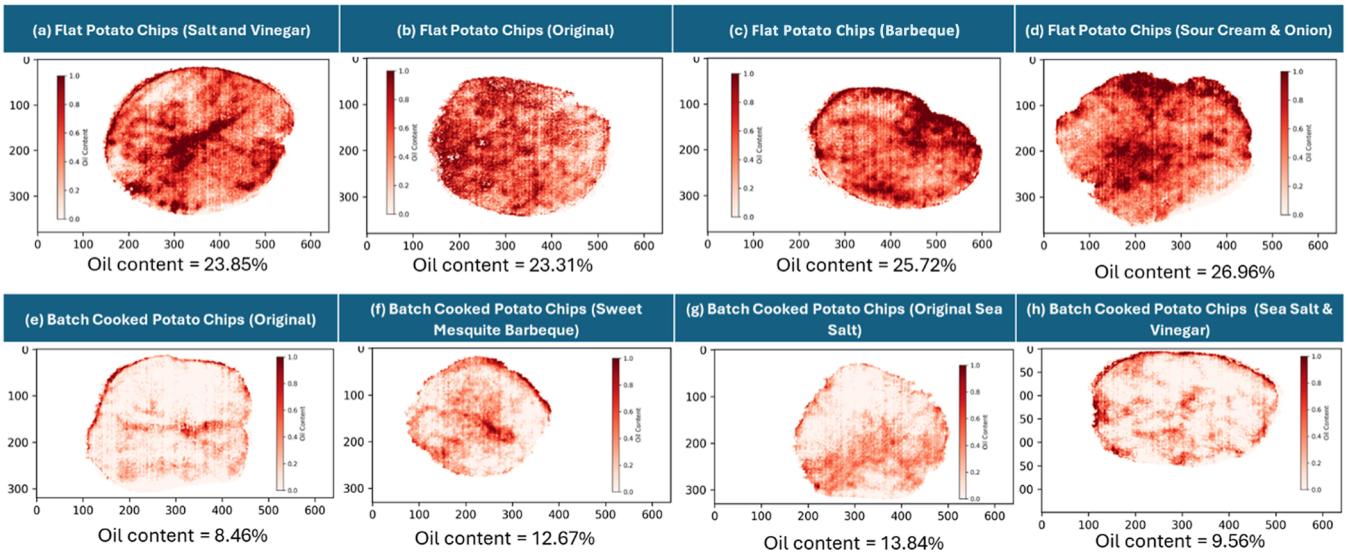

Figure 6. Visualization of oil distribution in potato chips. (a) Oil distribution visualization of the sample flat potato chips (salt and vinegar). (b) Oil distribution visualization of the sample flat potato chips (original). (c) Oil distribution visualization of the sample flat potato chips (barbeque). (d) Oil distribution visualization of the sample flat potato chips (sour cream & onion). (e) Oil distribution visualization of the sample batch-cooked potato chips (original). (f) Oil distribution visualization of the sample batch potato chips (sweet mesquite barbeque). (g) Oil distribution visualization of the sample batch-cooked potato chips (original sea salt). (h) Oil distribution visualization of the sample batch-cooked potato chips (sea salt & vinegar).

and efficient tool for rapid prediction and assessment of the oil content in food products, such as potato chips.

3.5. Variable Importance in Projection. In NIR spectroscopy, absorption bands primarily arise from overtones and combinations of fundamental molecular vibrations, such as stretching and bending modes of chemical bonds containing C−H, N−H, and O−H groups.[21] Validated by four different random states, high VIP scores have been consistently found at specific wavelength regions of 900−940, 1183−1228, and 1376−1414 nm, indicating that these wavelength regions are particularly significant in the model for predicting the response variable (Figure 5a). The high VIP scores at these wavelengths suggest that the molecular interactions or constituents that correspond to these absorption features have a strong influence on the variations captured by the PLSR model in the response variable. For example, the high VIP score in the regions of 900−940 nm and 1183−1228 nm are related to the second overtone of C−H stretching vibrations (−CH2), common in fatty acids.[39] It suggests that the oil content variations or presence of organic compounds with C−H bonds are highly relevant to the model's predictions. Wavelengths between 1376 and 1414 nm are related to the secondary overtone expansion of C−H.[40−42] The minor spike around the wavelength of 1440 nm typically corresponds to the first overtone of O−H stretching vibrations, indicating the small amount of water content in the sample matrix.[39]

The new PLSR model, built using the selected 68 wavelengths (VIP > 1) as a mask, was optimized by evaluating its performance at different numbers of latent variables. The best performing PLSR model was identified at N = 11, indicating that 11 latent variables were sufficient to capture the essential information from the selected wavelengths while avoiding overfitting (Figure 5b,c). The development of a simplified PLSR model using the selected wavelengths offers several advantages, such as reducing the dimensionality of the input data and focusing only on the most informative wavelengths for predicting the oil content. In addition, the simplified model may exhibit better generalization performance as it is less prone to overfitting compared to a model using the full spectrum. By concentrating on the most relevant wavelengths, the model can capture the underlying relationships between the spectral features and the oil content more effectively. The simplified PLSR model also performed well using the testing data set, achieving R^2 = 0.94 (Figure 5d).

3.6. Visualization of Oil Distribution in Potato Chips. The optimized SNV + PLSR model was applied to map the oil distribution in each representative potato chip as shown in Figure 6. The oil content in potato chips was indicated by color, and the darker red color indicated the higher oil content. This image confirmed that a generally higher oil content in the flat cooked potato chips when compared to the batch cooked potato chips. During the frying process, pores were observed to form on the surface and edges of the potato chips. This increase in porosity facilitates greater oil absorption, which is consistent with the mechanism of oil penetration that occurs during frying.[43] A similar observation was found using the HSI technology for rapid visualization of the oil distribution in potato chips. Future studies could further explore the effects of different processing methods, formulations, and ingredients on oil distribution and how these differences affect the nutritional value and the consumer choice of foods.

AUTHOR INFORMATION

Corresponding Author

Yiming Feng − *Virginia Seafood Agricultural Research & Extension Center, Virginia Tech, Hampton, Virginia 23669, United States; Department of Biological Systems Engineering, Virginia Tech, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0002-9693-3686; Phone: (757) 727-4861 ext. 21710; Email: yimingfeng@vt.edu

Authors

Yue Sun − *Virginia Seafood Agricultural Research & Extension Center, Virginia Tech, Hampton, Virginia 23669,*

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588

United States; School of Pharmacy, University of Camerino, Camerino, Macerata 62032, Italy

Nikhita Sai Nayani − *Virginia Seafood Agricultural Research & Extension Center, Virginia Tech, Hampton, Virginia 23669, United States; Department of Computer Science, Virginia Tech, Blacksburg, Virginia 24061, United States*

Yixiang Xu − *USDA, ARS, WRRC, Healthy Processed Foods Research, Albany, California 94710, United States;* orcid.org/0000-0003-3122-6101

Zhanfeng Xu − *PepsiCo Global R&D, Plano, Texas 75024, United States*

Jun Yang − *PepsiCo Global R&D, Plano, Texas 75024, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsfoodscitech.4c00196

Notes

The authors declare no competing financial interest.

Z.X. and J.Y. are employees of PepsiCo Inc. The views expressed in this manuscript are those of the authors and do not necessarily reflect the position or policy of PepsiCo Inc. The findings and conclusions in this research are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

■ ACKNOWLEDGMENTS

Y.F. acknowledges the support from the Virginia Agriculture Experiment Station and the Hatch Program of the National Institute of Food and Agriculture.

■ REFERENCES

(1) Research and Markets. Potato Chips Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2023−2028. 2023. https://www.researchandmarkets.com/report/potato-chips#:~:text=Whathttps://www.researchandmarkets.com/report/potato-chips#:~:text=What%20is%20the%20estimated%20value,%20at%20%2433.3%20Billion%20in%202022 (accessed 3/1/2024).

(2) Abong′, G. O.; Okoth, M. W.; Imungi, J. K.; Kabira, J. N. Consumption Patterns, Diversity and Characteristics of Potato Crisps in Nairobi, Kenya. *J. Appl. Biosci.* **2010**, *32*, 1942−1955.

(3) Mai Tran, T. T.; Chen, X. D.; Southern, C. Reducing Oil Content of Fried Potato Crisps Considerably Using a "sweet" Pre-Treatment Technique. *J. Food Eng.* **2007**, *80* (2), 719−726.

(4) Zellner, D. A.; Loaiza, S.; Gonzalez, Z.; Pita, J.; Morales, J.; Pecora, D.; Wolf, A. Food Selection Changes under Stress. *Physiol. Behav.* **2006**, *87* (4), 789−793.

(5) Zhang, Y.; Zhang, T.; Fan, D.; Li, J.; Fan, L. The Description of Oil Absorption Behavior of Potato Chips during the Frying. *Lwt* **2018**, *96* (April), 119−126.

(6) Aulia, R.; Amanah, H. Z.; Lee, H.; Kim, M. S.; Baek, I.; Qin, J.; Cho, B. K. Protein and Lipid Content Estimation in Soybeans Using Raman Hyperspectral Imaging. *Front. Plant Sci.* **2023**, *14* (August), 1−12.

(7) Nielsen, S. S. *Food Analysis*; Nielsen, S. S., Ed.; *Food Science Text Series*; Springer International Publishing: Cham, 2017.

(8) Kucha, C. T.; Liu, L.; Ngadi, M. O. Non-Destructive Spectroscopic Techniques and Multivariate Analysis for Assessment of Fat Quality in Pork and Pork Products: A Review. *Sensors* **2018**, *18* (2), 377.

(9) Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J. A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Electron.* **2019**, *57* (9), 6690−6709.

(10) Su, W. H.; Sun, D. W. Fourier Transform Infrared and Raman and Hyperspectral Imaging Techniques for Quality Determinations of Powdery Foods: A Review. *Compr. Rev. Food Sci. Food Saf.* **2018**, *17* (1), 104−122.

(11) Aviara, N. A.; Liberty, J. T.; Olatunbosun, O. S.; Shoyombo, H. A.; Oyeniyi, S. K. Potential Application of Hyperspectral Imaging in Food Grain Quality Inspection, Evaluation and Control during Bulk Storage. *J. Agric. Food Res.* **2022**, *8* (September 2021), 100288.

(12) Ma, J.; Sun, D. W.; Pu, H.; Cheng, J. H.; Wei, Q. Advanced Techniques for Hyperspectral Imaging in the Food Industry: Principles and Recent Applications. *Annu. Rev. Food Sci. Technol.* **2019**, *10*, 197−220.

(13) Su, W. H.; Xue, H. Imaging Spectroscopy and Machine Learning for Intelligent Determination of Potato and Sweet Potato Quality. *Foods* **2021**, *10* (9), 2146.

(14) Garhwal, A. S.; Pullanagari, R. R.; Li, M.; Reis, M. M.; Archer, R. Hyperspectral Imaging for Identification of Zebra Chip Disease in Potatoes. *Biosyst. Eng.* **2020**, *197*, 306−317.

(15) Rady, A.; Guyer, D.; Lu, R. Evaluation of Sugar Content of Potatoes Using Hyperspectral Imaging. *Food Bioprocess Technol.* **2015**, *8* (5), 995−1010.

(16) Fernández Pierna, J. A.; Vermeulen, P.; Amand, O.; Tossens, A.; Dardenne, P.; Baeten, V. NIR Hyperspectral Imaging Spectroscopy and Chemometrics for the Detection of Undesirable Substances in Food and Feed. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 233−239.

(17) Wu, D.; Sun, D. W. Advanced Applications of Hyperspectral Imaging Technology for Food Quality and Safety Analysis and Assessment: A Review - Part I: Fundamentals. *Innovative Food Sci. Emerging Technol.* **2013**, *19*, 1−14.

(18) Feng, Y. Z.; Sun, D. W. Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review. *Crit. Rev. Food Sci. Nutr.* **2012**, *52* (11), 1039−1058.

(19) Burger, J.; Gowen, A. Data Handling in Hyperspectral Image Analysis. *Chemom. Intell. Lab. Syst.* **2011**, *108* (1), 13−22.

(20) Pu, Y. Y.; Feng, Y. Z.; Sun, D. W. Recent Progress of Hyperspectral Imaging on Quality and Safety Inspection of Fruits and Vegetables: A Review. *Compr. Rev. Food Sci. Food Saf.* **2015**, *14* (2), 176−188.

(21) Bona, E.; Marquetti, I.; Link, J. V.; Makimori, G. Y. F.; da Costa Arca, V.; Guimarães Lemes, A. L.; Ferreira, J. M. G.; dos Santos Scholz, M. B.; Valderrama, P.; Poppi, R. J. Support Vector Machines in Tandem with Infrared Spectroscopy for Geographical Classification of Green Arabica Coffee. *Lwt* **2017**, *76*, 330−336.

(22) Barbon, S.; da Costa Barbon, A. P. A.; Mantovani, R. G.; Barbin, D. F. Machine Learning Applied to Near-Infrared Spectra for Chicken Meat Classification. *J. Spectrosc.* **2018**, *2018*, 1−12.

(23) Hwang, S. W.; Chung, H.; Lee, T.; Kim, J.; Kim, Y. J.; Kim, J. C.; Kwak, H. W.; Choi, I. G.; Yeo, H. Feature Importance Measures from Random Forest Regressor Using Near-Infrared Spectra for Predicting Carbonization Characteristics of Kraft Lignin-Derived Hydrochar. *J. Wood Sci.* **2023**, *69* (1), 1−12.

(24) Nawar, S.; Mouazen, A. M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of on-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. *Sensors* **2017**, *17* (10), 2428.

(25) Liu, T. A.; Zhai, X. Y.; Zhang, Q.; Lv, W.; Lu, W. C. Using NIR with Support Vector Regression to Predict the Crude Protein of Alfalfa. In *2016 International Conference on Information System and Artificial Intelligence (ISAI)*, 2016, pp 415−418.

(26) Kadamne, J.; Proctor, A. Rapid Oil Extraction from Potato Chips. *J. Am. Oil Chem. Soc.* **2010**, *87* (7), 835−836.

(27) Unger, P.; Sekhon, A. S.; Chen, X.; Michael, M. Developing an Affordable Hyperspectral Imaging System for Rapid Identification of Escherichia Coli O157:H7 and Listeria Monocytogenes in Dairy Products. *Food Sci. Nutr.* **2022**, *10* (4), 1175−1183.

(28) Rinnan, Å.; Berg, F. v. d.; Engelsen, S. B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC, Trends Anal. Chem.* **2009**, *28* (10), 1201−1222.
(29) Otsu, N.; Smith, P. L.; Reid, D. B.; Palo, L.; Alto, P. Otsu 1979 Otsu Method. *IEEE Trans. Syst. Man Cybern.* **1979**, *100* (1), 62−66.
(30) Grewal, R.; Kasana, S. S.; Kasana, G. Hyperspectral Image Segmentation: A Comprehensive Survey. *Multimedia Tools Appl.* **2023**, *82* (14), 20819−20872.
(31) Devos, O.; Ruckebusch, C.; Durand, A.; Duponchel, L.; Huvenne, J. P. Support Vector Machines (SVM) in near Infrared (NIR) Spectroscopy: Focus on Parameters Optimization and Model Interpretation. *Chemom. Intell. Lab. Syst.* **2009**, *96* (1), 27−33.
(32) Yao, K.; Sun, J.; Chen, C.; Xu, M.; Zhou, X.; Cao, Y.; Tian, Y. Non-Destructive Detection of Egg Qualities Based on Hyperspectral Imaging. *J. Food Eng.* **2022**, *325* (March), 111024.
(33) Farrés, M.; Platikanov, S.; Tsakovski, S.; Tauler, R. Comparison of the Variable Importance in Projection (VIP) and of the Selectivity Ratio (SR) Methods for Variable Selection and Interpretation. *J. Chemom.* **2015**, *29* (10), 528−536.
(34) Fei, X.; Jiang, X.; Lei, Y.; Tian, J.; Hu, X.; Bu, Y.; Huang, D.; Luo, H. The Rapid Non-Destructive Detection of the Protein and Fat Contents of Sorghum Based on Hyperspectral Imaging. *Food Anal. Methods* **2023**, *16* (11−12), 1690−1701.
(35) Hourant, P.; Baeten, V.; Morales, M. T.; Meurens, M.; Aparicio, R. Oil and Fat Classification by Selected Bands of Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54* (8), 1168−1174.
(36) Zhang, Y.; Zhang, L.; Wang, J.; Tang, X.; Wu, H.; Wang, M.; Zeng, W.; Mo, Q.; Li, Y.; Li, J.; Huang, Y.; Xu, B.; Zhang, M. Rapid Determination of the Oil and Moisture Contents in Camellia Gauchowensis Chang and Camellia Semiserrata Chi Seeds Kernels by Near-Infrared Reflectance Spectroscopy. *Molecules* **2018**, *23* (9), 2332.
(37) Alenezi, F. N. Majority Scoring with Backward Elimination in PLS for High Dimensional Spectrum Data. *Sci. Rep.* **2021**, *11* (1), 16974.
(38) Zheng, X.; Peng, Y.; Wang, W. A Nondestructive Real-Time Detection Method of Total Viable Count in Pork by Hyperspectral Imaging Technique. *Appl. Sci.* **2017**, *7* (3), 213.
(39) da Silva Medeiros, M. L.; Cruz-Tirado, J. P.; Lima, A. F.; de Souza Netto, J. M.; Ribeiro, A. P. B.; Bassegio, D.; Godoy, H. T.; Barbin, D. F. Assessment Oil Composition and Species Discrimination of Brassicas Seeds Based on Hyperspectral Imaging and Portable near Infrared (NIR) Spectroscopy Tools and Chemometrics. *J. Food Compos. Anal.* **2022**, *107* (October 2021), 104403.
(40) Zhang, L.; An, D.; Wei, Y.; Liu, J.; Wu, J. Prediction of Oil Content in Single Maize Kernel Based on Hyperspectral Imaging and Attention Convolution Neural Network. *Food Chem.* **2022**, *395* (June), 133563.
(41) Shiroma, C.; Rodriguez-Saona, L. Application of NIR and MIR Spectroscopy in Quality Control of Potato Chips. *J. Food Compos. Anal.* **2009**, *22* (6), 596−605.
(42) Wu, N.; Zhang, Y.; Na, R.; Mi, C.; Zhu, S.; He, Y.; Zhang, C. Variety Identification of Oat Seeds Using Hyperspectral Imaging: Investigating the Representation Ability of Deep Convolutional Neural Network. *RSC Adv.* **2019**, *9* (22), 12635−12644.
(43) Ziaiifar, A. M.; Courtois, F.; Trystram, G. Porosity Development and Its Effect on Oil Uptake during Frying Process. *J. Food Process Eng.* **2010**, *33* (2), 191−212.
(44) Panda, B. K.; Mishra, G.; Ramirez, W. A.; Jung, H.; Singh, C. B.; Lee, S. H.; Lee, I. Rancidity and Moisture Estimation in Shelled Almond Kernels Using NIR Hyperspectral Imaging and Chemometric Analysis. *J. Food Eng.* **2022**, *318* (October 2021), 110889.
(45) McDonald, G. C. Ridge Regression. *WIREs Comput. Stat.* **2009**, *1* (1), 93−100.
(46) Lin, L.; Wang, F.; Xie, X.; Zhong, S. Random Forests-Based Extreme Learning Machine Ensemble for Multi-Regime Time Series Prediction. *Expert Syst. Appl.* **2017**, *83*, 164−176.
(47) Lu, H.; Mazumder, R. Randomized Gradient Boosting Machine. *SIAM J. Control* **2020**, *30* (4), 2780−2808.
(48) Kaneko, H. Support Vector Regression That Takes into Consideration the Importance of Explanatory Variables. *J. Chemom.* **2021**, *35* (4), 1−11.

https://doi.org/10.1021/acsfoodscitech.4c00196
ACS Food Sci. Technol. 2024, 4, 1579−1588