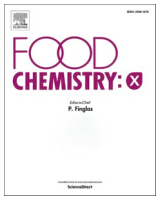
Food Chemistry: X 23 (2024) 101763

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



Innovative strategies for protein content determination in dried laver (*Porphyra* spp.): Evaluation of preprocessing methods and machine learning algorithms through short-wave infrared imaging

Eunghee Kim [a,1], Jong-Jin Park [b,1], Gyuseok Lee [a], Jeong-Seok Cho [a,b], Seul-Ki Park [a], Dae-Yong Yun [b], Kee-Jai Park [a,b], Jeong-Ho Lim [a,b,*]

[a] *Smart food manufacturing project group, Korea Food Research Institute, Wanju-gun 55365, South Korea*
[b] *Food safety and distribution research group, Korea Food Research Institute, Wanju-gun 55365, South Korea*

ARTICLE INFO

Keywords:
Dried laver
Short-wave infrared imaging
Protein content prediction
Preprocessing method
Machine learning model

ABSTRACT

In this study, we explored the application of Short-Wave Infrared (SWIR) hyperspectral imaging combined with Competitive Adaptive Reweighted Sampling (CARS) and advanced regression models for the non-destructive assessment of protein content in dried laver. Utilizing a spectral range of 900–1700 nm, we aimed to refine the quality control process by selecting informative wavelengths through CARS and applying various preprocessing techniques (standard normal variate [SNV], Savitzky-Golay filtering [SG], Orthogonal Signal Correction [OSC], and StandardScaler [SS]) to enhance the model's accuracy. The SNV-OSC-StandardScaler-Support vector regression (SVR) model trained on CARS-selected wavelengths significantly outperformed the other configurations, achieving a prediction determination coefficient (Rp^2) of 0.9673, root mean square error of prediction of 0.4043, and residual predictive deviation of 5.533. These results highlight SWIR hyperspectral imaging's potential as a rapid and precise tool for assessing dried laver quality, aiding food industry quality control and dried laver market growth.

1. Introduction

Seaweed has been a traditional dietary staple in eastern and southeastern Asia, including China, Indonesia, Korea, Philippines, and Japan. Despite its longstanding popularity in these regions, its significance and potential have often been undervalued in other regions (Msuya et al., 2022). However, recently, it has garnered significant interest as a source of various nutrients such as protein and dietary fiber (Marques de Brito, Campos, Neves, Ramos, & Tomita, 2023; Murai, Yamagishi, Kishida, & Iso, 2021). Furthermore, the excellent carbon capture ability has positioned seaweed farming as a countermeasure against global warming (Yong, Thien, Rupert, & Rodrigues, 2022). Consequently, the seaweed market has now reached an annual value of 6 billion dollars, and production has grown to 12 million tons per year (García-Poza et al., 2022).

Laver (*Porphyra* spp.) is one of the most widely cultivated seaweeds and is mainly consumed in its dried form. Seaweed is not only nutritionally excellent, as it is rich in protein, dietary fiber, and vitamin B, but is also used as a raw material for various foods such as sushi and gimbap (Wada et al., 2021). It is also processed in the form of seasoned snacks and has recently become popular not only in Asia but also in the West.

Red seaweeds, including laver, are known for their wide-ranging protein content, varying from 2.7 % to 47 % (Figueroa, Farfán, & Aguilera, 2023). This variability is primarily influenced by seasonal and climatic conditions. In the northern hemisphere, protein contents in red seaweeds are observed to decrease during the summer, while late winter and spring see an increase in protein content due to the elevated nitrogen levels associated with upwelling (Raja et al., 2020). The harvested seaweed is subjected to drying processes to obtain a moisture content of 15 % or less. This dehydration step enhances the suitability of seaweed for distribution and storage and ensures hygiene. The most important quality indicator of laver is the protein content. Laver's high protein content not only provides nutritional value, but the amino acids

* **Corresponding author at:** Korea Food Research Institute, Wanju-gun, 55365, South Korea.
E-mail addresses: kim.eunghee@kfri.re.kr (E. Kim), p.jongjin@kfri.re.kr (J.-J. Park), gslee@kfri.re.kr (G. Lee), jscho@kfri.re.kr (J.-S. Cho), skpark@kfri.re.kr (S.-K. Park), ydy0401@kfri.re.kr (D.-Y. Yun), jake@kfri.re.kr (K.-J. Park), jhlim@kfri.re.kr (J.-H. Lim).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fochx.2024.101763
Received 18 April 2024; Received in revised form 5 August 2024; Accepted 21 August 2024
Available online 23 August 2024

such as taurine, alanine, and glutamic acid derived from this protein are also the basis for its distinctive flavor (Jeong et al., 2023). Additionally, dried laver is produced by grinding raw laver and then drying it, which results in minimal visual differences based on its composition and gives it a characteristic black color. Consequently, it is challenging to determine the protein content of dried laver without specific quantative methods such as Kjeldahl Method. Traditional methods for determining protein content are characterized by their ease of use and reliability. However, they are time-consuming, laborious, and destructive. Furthermore, they do not align well with the requirements of rapid and non-destructive evaluation of dried laver. For these reasons, novel methods are necessary and should be developed for a rapid quality assessment of dried laver.

Hyperspectral imaging is a technology that combines imaging and spectroscopy, allowing simultaneous acquisition of spatial and spectral information using a single system (Özdoğan, Lin, & Sun, 2021). The spectral regions widely used for food analysis via hyperspectral imaging (HSI) include the ultraviolet (200–400 nm), VIS/NIR (400–1000 nm), and near-infrared (900–2500 nm) regions (Elmasry, Kamruzzaman, Sun, & Allen, 2012). The information obtained through hyperspectral analysis can affect the performance of the learning models due to increased data dimensionality and the extensive redundancy of information inherent in hyperspectral imaging data. Therefore, appropriate preprocessing techniques such as redundancy removal and feature selection are being utilized to enhance the efficiency of the models (Nagy, Wang, & Farag, 2022).

In recent years, short-wave infrared (SWIR) hyperspectral imaging has emerged as a powerful non-destructive analytical tool for the quantitative and qualitative assessment of food. This technique, particularly when combined with various machine-learning methods, has been effectively used to evaluate the nutritional and hygienic indicators of various foods. Notable studies have demonstrated the effectiveness of SWIR in conjunction with machine-learning techniques for analyzing food products, offering significant insights into their quality and safety (Kang et al., 2022; Ozturk, Bowler, Rady, & Watson, 2023). Hyperspectral imaging has great potential and has been used for quality assessment of a variety of agricultural products and foods, including determination of chemical components such as moisture, protein, ash, oil, reducing sugar, etc. (Fatemi, Singh, & Kamruzzaman, 2022; He et al., 2022). It has also been applied in predicting microbial spoilage (Manthou et al., 2022), analyzing textural profiles (de Souza Zangirolami, Moreira, Leimann, Valderrama, & Março, 2023), and detecting food adulteration (Amirvaresi, Nikounezhad, Amirahmadi, Daraei, & Parastar, 2021). Given these research findings, SWIR and machine learning have been widely applied in the analysis of various foods. However, studies specifically focusing on predicting the composition of seaweeds such as laver using these technologies are notably rare despite their importance. Additionally, data preprocessing methods have often been applied without consideration of the specific characteristics of machine learning models. Therefore, research is needed to compare the effects of different preprocessing techniques and explore combinations of various preprocessing methods in hyperspectral image analysis to optimize the impact on machine learning model training.

Therefore, the objectives of this study were as follows: (1) To compare the effects of different preprocessing techniques, both individually and in combination, on the performance of various machine learning models (Partial Least Square Regression, Support Vector Regression, Elastic Net Regression, Gradient Boosting Regression, and Random Forest Regression) using SWIR hyperspectral imaging data in the spectral region of 900–1700 nm, with preprocessing techniques including Standard Normal Variate (SNV), Savitzky-Golay filtering (SG), Orthogonal Signal Correction (OSC), and StandardScaler (SS), (2) To develop prediction models for protein content using both the complete spectral region and effective wavelengths selected via CARS (Competitive Adaptive Reweighted Sampling) with SWIR hyperspectral imaging in the spectral range of 900–1700 nm, and (3) To apply the models to each pixel of the images to generate chemical maps for visualizing the distribution of protein content. In this study, we aimed to demonstrate the feasibility of using SWIR-based spectroscopy combined with machine learning for the effective non-destructive prediction of protein content in dried laver.

2. Materials and methods

2.1. Sample preparation

Ninety dried laver samples were harvested and processed in Jangheung, Jeollanam-do, and Wido, Jeollabuk-do, South Korea from December 2021 to February 2022 and then stored in the refrigerator (−18 ± 0.5 °C) for further hyperspectral image collection and analysis. The width, depth, and thickness of the samples were 26.67 ± 0.29 cm, 19.37 ± 0.15 cm, and 0.09 ± 0.01 mm, respectively. For the purpose of model training and validation, the samples were evenly divided into a calibration set and a prediction set in an 8:2 ratio.

2.2. Determination of protein content

The technique developed by the Association of Official Analytical Chemists was used to examine the proximate composition of the samples (Jeong et al., 2023). The moisture content was determined by air-drying at 105 °C. The protein content was measured using the Kjeldahl method. Briefly, each sample (0.5 g) was added to a digestion flask along with 10 mL of sulfuric acid (96–98 %) and selenium tablets. Digestion was performed using the meat AOAC program from the Digestor™ auto 2508 (Foss Analytic), and the distillation and titration were conducted using an automatic Kjeltec™ 8400 (Foss Analytic) unit. The measured nitrogen content (%) was converted to protein content (%) using a conversion factor of 6.25. Each parameter of each sample was measured thrice and averaged for use as the reference value.

2.3. Hyperspectral image acquisition

The custom-designed hyperspectral system comprised a spectrograph (N17E, Specim, Oulu, Finland), vision dome light (VTDL550*240, Vision Technology, Cheonan-si, Korea) with six halogen lamps (150 W power), a SWIR camera (PA320F300TCL, OZRAY, Korea), and a linear sample stage (FBL80E1400, FUYU, Sichuan, China). The optical module (SWIR camera, spectrograph, and vision dome light) of the SWIR system was fixed 460 mm above the sample, and SWIR spectral images (hypercube) were acquired in the line-scan mode while moving the module at a constant velocity of 275 mm/s using the linear sample stage. The SWIR spectral images were recorded from 900 to 1700 nm, and the reflectance intensities of the images were measured at an average interval of 3.45 nm. The resolution of the SWIR spectral images was 320 pixels in the horizontal direction and 256 pixels in the vertical direction, and the spectral band comprised 256 channels. White and black background images were obtained by scanning a white tile (99.99 % reflectance) and completely turning off the lens using a cap (0.00 % reflectance).

2.4. Spectral extraction and preprocessing

After image calibration, the spectral information within the region of interest (ROI) of the sample image was extracted and averaged into one spectrum to represent the sample. This process was implemented using the hyperspectral imaging software Breeze (Prediktera AB, Umea, Sweden). To improve the accuracy of the SWIR quantitative analysis, a methodical spectral preprocessing strategy was adopted to highlight crucial information and reduce irrelevant background noise and scattering effects. Spectral data preprocessing was accomplished using the Unscrambler X version 10.4 software (CAMO Software, Oslo, Norway).

To determine the impact of these preprocessing techniques on model

performance, they were applied both individually and in various combinations. For instance, SNV followed by SG filtering was used to correct scatter effects and then smooth the data, while SNV combined with OSC was employed to normalize the spectra and remove uncorrelated noise. Another combination included applying StandardScaler after SNV to ensure both scatter correction and standardized scaling across features. Including the raw (unprocessed) data, a total of 12 preprocessing techniques were used in this study. These included the four preprocessing methods (SNV, SS, SG, OSC) applied individually and in various combinations, as well as the raw data. The spectra processed with these individual and combined preprocessing techniques are illustrated in Fig. 1, providing a visual overview of the methods applied to ensure optimal data quality for subsequent analysis.

2.4.1. Standard Normal Variate (SNV)

SNV is a scatter correction method used to reduce the effects of particle size and surface irregularities on the spectral data. This method normalizes each spectrum individually by centering the data to have a mean of zero and scaling it to have a standard deviation of one. By doing so, it ensures consistent spectral intensity across samples, thereby reducing variability caused by scatter effects. The transformed spectrum provides a more accurate representation of the sample's true spectral features, making it easier to compare different samples and improving the robustness of subsequent analysis.

2.4.2. StandardScaler (SS)

In contrast to SNV, which normalizes each spectrum individually, StandardScaler standardizes the data across all samples. It normalizes the data by transforming it to have a zero mean and unit variance for each spectral feature across the entire dataset. This means that each wavelength is scaled based on the mean and standard deviation calculated from all samples, ensuring that each spectral feature contributes equally to the analysis. StandardScaler addresses biases caused by different scales of measurement and can enhance the performance of machine learning algorithms sensitive to input scale differences.

2.4.3. Savitzky-Golay filtering (SG)

Savitzky-Golay filtering aims to smooth noisy spectral data while preserving the important spectral features such as peaks and troughs. This method applies a polynomial smoothing algorithm by fitting successive subsets of adjacent data points with a low-degree polynomial through the method of linear least squares. By moving the window across the data points and recalculating the polynomial coefficients, SG filtering effectively reduces random noise without significantly distorting the signal. This results in a smoother spectrum that retains the essential characteristics needed for accurate analysis.

2.4.4. Orthogonal signal correction (OSC)

OSC is a preprocessing technique designed to remove variations in the spectral data that are orthogonal (i.e., unrelated) to the response variable of interest. This method enhances the focus of the model on pertinent spectral features by eliminating uncorrelated noise and irrelevant information. The OSC process involves projecting the original spectral data onto a subspace orthogonal to the response variable, resulting in a corrected data matrix that highlights the relevant spectral information while suppressing the background noise. This improves the robustness and accuracy of the predictive models.

2.5. Predictive modeling approaches

Five regression models (Partial Least Squares Regression, Support Vector Regression, Elastic Net Regression, Gradient Boosting Regression, and Random Forest Regression) were developed and individually tested for quantitative analysis based on the absorbance SWIR spectra, following the methodology outlined in Fig. 2. The best settings for each machine learning model were thoroughly searched for using the Grid Search approach (GridSearchCV) with cross-validation (cv = 5), trying out various specified values for its hyperparameters. Each model was trained using the calibration set that had been preprocessed with each of the 12 different preprocessing techniques. The optimal hyperparameters were identified by exploring a range of values, as listed in Table S1, to

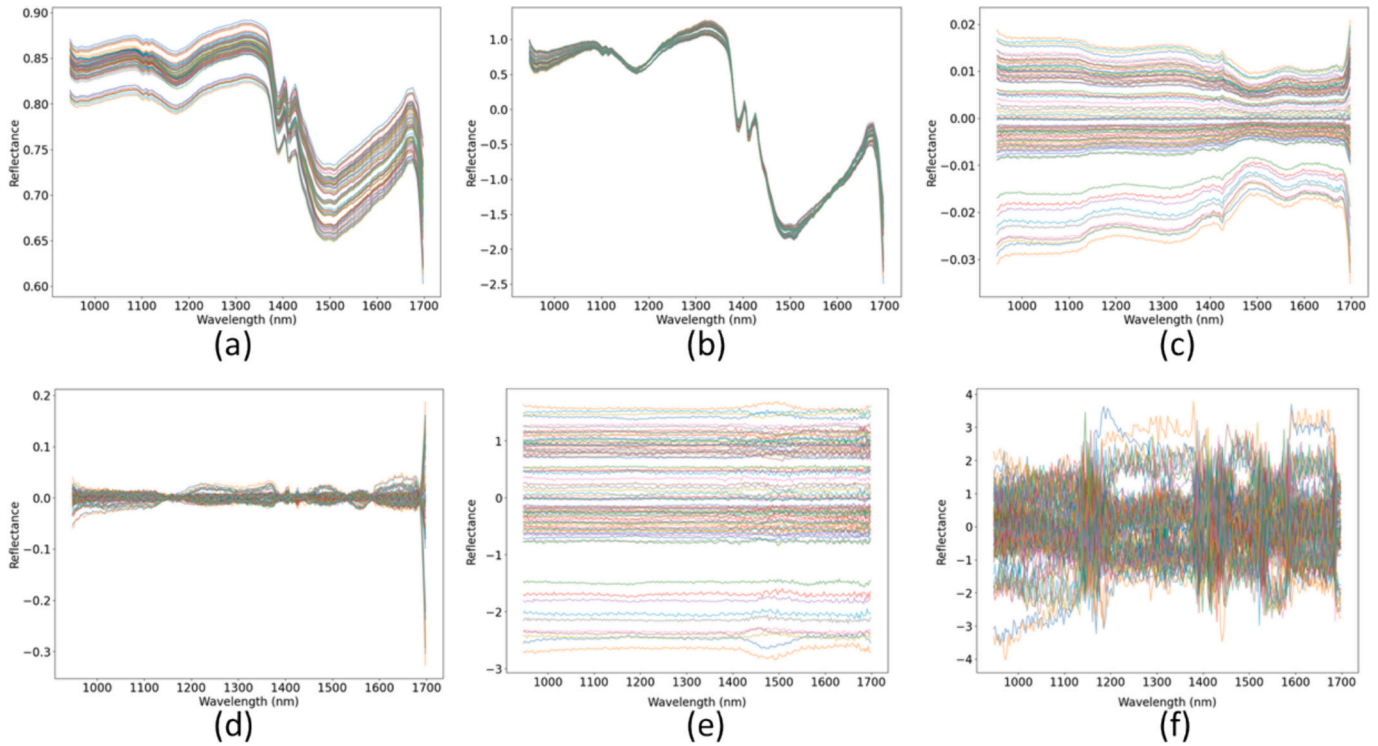


Fig. 1. SWIR spectra of dried laver by different preprocessing method including standard normal variate (SNV), Savitzky-Golay filtering (SG), Orthogonal Signal Correction (OSC), and StandardScaler (SS). (a) Raw spectra, (b) SNV, (c) OSC, (d) SNV + OSC, (e) OSC + SS, (f) SNV + OSC + SS.

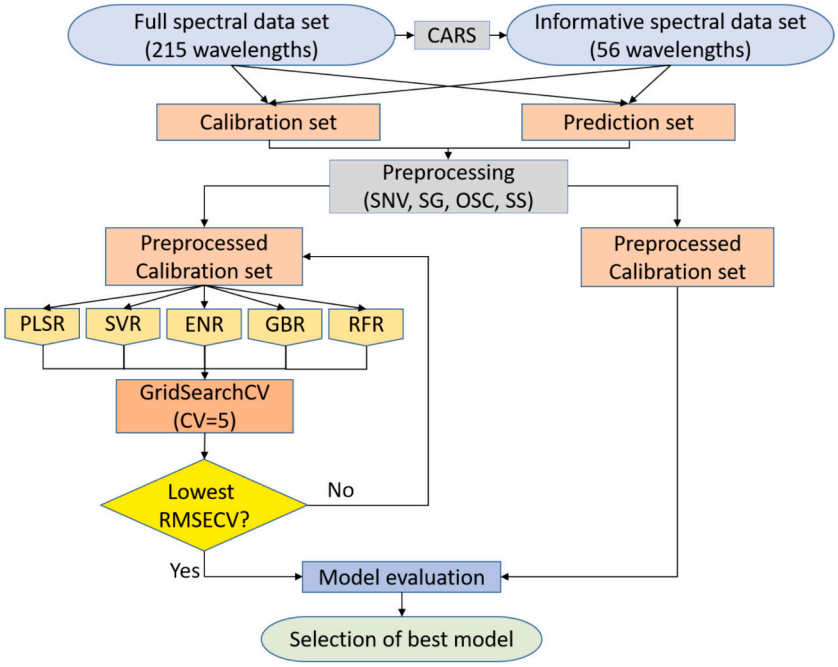


Fig. 2. Workflow of overall methodology for proposed the evaluation models.

find the combination that enhances the performance of each model. The models trained with the optimal hyperparameter conditions were then validated by estimating the protein content using the prediction set.

2.5.1. Partial least squares regression (PLSR)

Partial Least Squares Regression (PLSR) is tailored for scenarios where the relationship between variables is complex with a high degree of multicollinearity. By transforming a large set of variables into a smaller set of uncorrelated latent variables, PLSR simplifies analysis without sacrificing essential information. This model is particularly suitable for linear relationships in spectral analysis and offers a straightforward method to model substantial amounts of data with fewer samples (Cheng & Sun, 2017).

2.5.2. Support vector regression (SVR)

Support Vector Regression (SVR) stands out for its unique approach to regression challenges, emphasizing the fitting of errors within a specific threshold known as the epsilon margin. This methodology renders SVR exceptionally adept at managing outliers and ensuring that predictions are both robust and within a defined tolerance level. The flexibility afforded by the use of kernel functions allows SVR to effectively capture complex, nonlinear relationships within the data, which is particularly beneficial for the analysis of spectral data, where such patterns are common (Bermolen & Rossi, 2009).

2.5.3. Elastic Net regression (ENR)

Elastic Net combines the best ridge and LASSO regression, offering a balanced solution for regularization and variable selection. This method is particularly useful for spectral analysis, where numerous predictors can be highly correlated. Elastic Net simplifies model complexity and enhances interpretability by selecting a relevant subset of variables, effectively addressing multicollinearity, and reducing the risk of overfitting (Z. Zhang et al., 2017).

2.5.4. Gradient boosting regression (GBR)

Gradient Boosting distinguishes itself through its sequential model-building approach, which focuses on correcting errors in preceding models. This technique, which leverages the strengths of multiple weak learners, is particularly effective for modeling complex nonlinear data relationships. Its adaptability extends to handling missing data and incorporating various loss functions, making it a versatile tool for spectral analysis (Golden, Rothrock Jr, & Mishra, 2019).

2.5.5. Random forest regression (RFR)

Random Forest (RF) is well-known for its simplicity and scalability, utilizing an ensemble of decision trees to produce robust and accurate predictions. RF excels in dealing with high-dimensional datasets owing to its inherent feature-selection capabilities and resistance to overfitting. The model also offers valuable insights into variable importance, enhancing its utility in spectral analysis and other applications where understanding feature relevance is crucial (Ribeiro et al., 2021) .

2.6. Informative wavelength selection and model optimization

In the analysis of hyperspectral images, the presence of redundant information and multicollinearity can adversely affect model predictions, compromising accuracy, robustness, and predictive efficiency. To mitigate these issues, wavelength selection is pivotal to identify and utilize wavelengths that significantly enhance the predictive capabilities of the model. In this context, Competitive Adaptive Reweighted Sampling (CARS) was employed to discern the most informative wavelengths. CARS assesses the importance of each wavelength through the absolute values of regression coefficients, adhering to the principle of "survival of the fittest" (Dai et al., 2014). This selection process was executed using the MATLAB R2023b software (MathWorks, Inc., Natick, MA, USA), ensuring rigorous and efficient identification of the key wavelengths for our study.

2.7. Model performance index

In the evaluation of the predictive models in this study, a suite of indices was deployed to measure performance, encompassing both the calibration and prediction phases. This array of indices included the corrected correlation coefficient (R_c^2) for calibration, the root mean square error of calibration (RMSEC), and the root mean square error of cross-validation (RMSECV), which are all pivotal in the model calibration phase. Higher R_c^2 and lower RMSEC and RMSECV values indicate a model with better stability and calibration performance. The prediction

determination coefficient (R_p^2), root mean square error of prediction (RMSEP), and residual predictive deviation (RPD) are the benchmarks during transitioning to the prediction phase. R_p^2 gauges the accuracy of the model in terms of predictive capacity, reflecting the correlation between the observed values and the model's predictions, with values closer to 1 indicating greater accuracy. The RMSEP quantifies the precision of the model predictions by measuring the average discrepancy between the predicted and observed values; a lower RMSEP denotes enhanced precision. Finally, the RPD serves as a metric for the accuracy of the model by comparing the standard deviation of the reference laboratory values against the RMSEP, with values greater than three indicating high predictive accuracy and model robustness. These parameters were determined with the following formulas:

$$R_c{}^2, R_{cv}{}^2, R_p{}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - Y_i)^2}{\sum_{i=1}^{n} (y_i - y_m)^2} \quad (1)$$

$$\mathrm{RMSEC}, \mathrm{RMSECV}, \mathrm{RMSEP} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - Y_i)^2} \quad (2)$$

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSEP}} \quad (3)$$

where y_i and Yi denote the measured and predicted values for the i-th sample in the calibration or prediction set, respectively, y_m is the mean of the protein concentration of all samples in the calibration or prediction set, and n is the number of samples in the calibration or prediction set. And, SD is the standard deviation of the reference protein concentrations in the calibration or prediction set.

All chemometric analyses were conducted using the Scikit-learn machine-learning package in Python version 3.10.12. Scikit-learn is an open-source package that operates on top of the scientific and numerical libraries Scipy and Numpy.

2.8. Chemical visualization

Hyperspectral imaging offers a distinct advantage, namely, a spatial imaging capability, which is superior to conventional spectroscopy. To comprehensively and directly understand the differences in protein content from spot to spot in the same dried laver, the optimized model for each quality index was applied to individual pixels of the original ROI image. The protein distribution in a dried seaweed sample can be determined by calculating the dot product between the spectrum of each pixel in the image and the regression coefficient of the best model to generate a color map, where pixels exhibiting similar spectral features at the information wavelength are assigned quality index values corresponding to the same predicted value. This was visualized with the naked eye. Data analysis and visualization were conducted using Python version 3.10.12. We utilized Pandas for data handling, NumPy for numerical computations, and Matplotlib for generating visualizations. File operations were managed with the os module.

3. Results and discussion

3.1. Statistical protein contents

In this study, we predicted the protein content of 90 dried laver samples. To do so, a total of 72 samples (80 % of the total) were randomly chosen to form a training set for the calibration and internal validation of the model. Table S2 presents statistical data for both calibration and prediction sets, detailing mean, maximum, minimum, median, range, and standard deviation (SD). The protein values' range in the calibration set encompasses the range in the prediction set, and the mean, median, and SD of protein values in the calibration set closely align with those in the prediction set. This indicates a suitable distribution of samples between the calibration and validation sets. This training set underwent five-fold cross-validation to assess the model's performance across a range of preprocessing types and hyperparameters for each machine learning algorithm. The results from the cross-validation represented the average outcomes from all five folds. The remaining 18 samples (20 %) were designated as an independent testing set for external model validation. Based on the results of the internal validation, we selected the model that demonstrated the best performance in terms of preprocessing and hyperparameters. The selected model was then evaluated using the independent test set.

3.2. Predicting protein content based on full wavelength

A comprehensive study of dried laver samples using SWIR hyperspectral imaging within the 900–1700 nm range was conducted, with a specific emphasis on the protein content. This study encompassed a large dataset comprising 215 spectra. We applied a variety of preprocessing techniques along with sophisticated machine learning models to decode intricate relationships within the data. Through a detailed examination, this study explored the interplay between preprocessing methods and five cutting-edge algorithms: PLSR, SVR, ENR, GBR, and RFR. It meticulously navigated through 12 spectral variations, including Raw, SNV, SG, OSC, SNV-SG, and SNV-OSC, evaluated both with and without the application of StandardScaler. The results of the preprocessing and regression model training, which assess how various preprocessing techniques affect the performance of each model using the full spectrum, are presented in Table 1.

The impact of different preprocessing techniques varied across the models. SNV preprocessing significantly improved the performance of the SVR model because SVR is sensitive to the scales of the input variables, and SNV normalizes each spectrum to have a mean of zero and a standard deviation of one, reducing variability caused by scatter effects. However, for PLSR and ENR, SNV did not show significant effects, likely because these models are less affected by scatter-related variability. In contrast, GBR and RFR models showed a significant decrease in R_p^2 with SNV preprocessing, which could be attributed to the removal of essential information or reduced model fit, as observed in studies by Jin et al. and Aheto et al. (Aheto et al., 2020; Jin et al., 2022). OSC demonstrated significant performance enhancement in all models since OSC removes variations in the spectral data that are orthogonal to the response variable, focusing the model on pertinent spectral features. This normalization helps all models to better capture the spectral signatures relevant to protein content, leading to improved accuracy and robustness in predictions (J. Zhang et al., 2023; Zhu et al., 2021). SS preprocessing showed significant performance improvements, especially in SVR and ENR. SS standardizes the data across all samples, ensuring each spectral feature contributes equally to the analysis. This normalization is crucial for models like SVR and ENR, which are sensitive to the scales of the input variables. By balancing the scales, SS enhances the regularization applied by ENR and improves model performance (Zou & Hastie, 2005). SG filtering improved performance across most models but did not result in substantial gains. This method smooths noisy spectral data while preserving important spectral features, which is beneficial but not transformative for all models. For decision tree-based ensemble models like GBR and RFR, SG filtering can negatively impact performance by removing critical variations needed for accurate predictions. This phenomenon was also observed in the study by Loggenberg et al., where SG filtering led to decreased performance in decision tree-based models due to the loss of important signal variations (Loggenberg et al., 2018). When SNV and SS were combined, the performance sometimes decreased, similar to the effect observed with SG filtering. This is likely because both SNV and SS normalize the data, which, when combined with other preprocessing techniques, could overly smooth the data and remove essential variations needed for complex models like GBR and RFR. In contrast, models like PLSR and SVR benefited from these

Table 1
Quantitative protein prediction performance based on full band spectra wavelengths.

Models	Preprocessing	Calibration set		Validation set		Prediction set		RPD
		R_c^2	RMSEC	R_{cv}^2	RMSECV	R_p^2	RMSEP	
PLSR	Raw	0.9943	0.1579	0.7564	1.0306	0.8644	0.8237	2.7159
	SNV	0.9927	0.1789	0.7906	0.9556	0.8517	0.8615	2.5966
	SG	0.9772	0.3152	0.7953	0.9447	0.8828	0.7658	2.9212
	OSC	1.0000	0.0003	0.9730	0.3428	0.9045	0.6913	3.2358
	SNV_SG	0.9774	0.3136	0.8175	0.8920	0.8704	0.8055	2.7773
	SNV_OSC	1.0000	0.0000	0.9791	0.3018	0.9090	0.6749	3.3146
	Raw_SS	0.9195	0.5925	0.6555	1.2256	0.8010	0.9979	2.2417
	SNV_SS	0.9700	0.3616	0.8263	0.8703	0.8785	0.7797	2.8691
	SG_SS	0.8928	0.6835	0.6915	1.1597	0.8151	0.9620	2.3253
	OSC_SS	0.9997	0.0333	0.9570	0.4331	0.9056	0.6874	3.2540
	SNV_SG_SS	0.9553	0.4413	0.8729	0.7444	0.8666	0.8171	2.7376
	SNV_OSC_SS	0.9999	0.0219	0.9600	0.4175	0.9581	0.4577	4.8875
SVR	Raw	0.6479	1.2391	0.2633	1.7922	0.7151	1.1939	1.8736
	SNV	0.9765	0.3200	0.7221	1.1007	0.8754	0.7896	2.8331
	SG	0.6450	1.2441	0.2566	1.8004	0.7144	1.1954	1.8712
	OSC	0.8424	0.8291	0.8251	0.8734	0.9044	0.6917	3.2341
	SNV_SG	0.8640	0.7699	0.7685	1.0047	0.8297	0.9231	2.4233
	SNV_OSC	0.9857	0.2496	0.9506	0.4642	0.9207	0.6298	3.5518
	Raw_SS	0.8777	0.7302	0.7151	1.1146	0.7950	1.0129	2.2084
	SNV_SS	1.0000	0.0011	0.8261	0.8707	0.8451	0.8804	2.5408
	SG_SS	0.8860	0.7049	0.7407	1.0633	0.8142	0.9643	2.3198
	OSC_SS	1.0000	0.0010	0.9789	0.3034	0.9006	0.7053	3.1718
	SNV_SG_SS	0.9978	0.0978	0.8455	0.8208	0.8410	0.8919	2.5080
	SNV_OSC_SS	1.0000	0.0010	0.9833	0.2697	0.9588	0.4539	4.9287
ENR	Raw	0.4321	1.5735	0.0632	2.0211	0.4855	1.6046	1.3941
	SNV	0.6469	1.2407	0.0905	1.9914	0.5929	1.4273	1.5673
	SG	0.4320	1.5737	0.0632	2.0210	0.4854	1.6047	1.3940
	OSC	0.7235	1.0979	0.6222	1.2835	0.7787	1.0523	2.1257
	SNV_SG	0.6191	1.2887	0.0435	2.0421	0.5805	1.4489	1.5440
	SNV_OSC	0.9082	0.6328	0.8967	0.6711	0.9480	0.5100	4.3860
	Raw_SS	0.8782	0.7289	0.6196	1.2879	0.8087	0.9785	2.2862
	SNV_SS	0.9930	0.1743	0.8287	0.8643	0.8767	0.7856	2.8473
	SG_SS	0.8623	0.7749	0.6430	1.2476	0.8076	0.9813	2.2796
	OSC_SS	0.9583	0.4266	0.9141	0.6121	0.9096	0.6725	3.3263
	SNV_SG_SS	0.9779	0.3103	0.8564	0.7914	0.8668	0.8163	2.7402
	SNV_OSC_SS	0.9999	0.0237	0.9630	0.4014	0.9443	0.5279	4.2373
GBR	Raw	0.9992	0.0592	0.1332	1.9440	0.6644	1.2958	1.7262
	SNV	0.9999	0.0233	0.2244	1.8390	0.2606	1.9236	1.1629
	SG	0.8945	0.6783	0.2185	1.8459	0.6995	1.2262	1.8242
	OSC	0.9859	0.2478	0.7957	0.9439	0.9023	0.6991	3.1998
	SNV_SG	1.0000	0.0000	0.0320	2.0544	0.3755	1.7678	1.2654
	SNV_OSC	1.0000	0.0001	0.8787	0.7271	0.9390	0.5526	4.0481
	Raw_SS	0.8962	0.6727	0.2052	1.8616	0.7018	1.2216	1.8312
	SNV_SS	1.0000	0.0001	0.2292	1.8332	0.2724	1.9081	1.1723
	SG_SS	0.8945	0.6783	0.2311	1.8310	0.6989	1.2276	1.8223
	OSC_SS	0.9859	0.2478	0.7972	0.9402	0.9028	0.6973	3.2082
	SNV_SG_SS	1.0000	0.0000	0.0165	2.0708	0.4470	1.6634	1.3448
	SNV_OSC_SS	1.0000	0.0000	0.8557	0.7933	0.9417	0.5400	4.1428
RFR	Raw	0.9277	0.5614	0.1655	1.9075	0.7107	1.2033	1.8591
	SNV	0.9568	0.4341	0.0463	2.0392	0.3663	1.7807	1.2562
	SG	0.9238	0.5765	0.1707	1.9016	0.7267	1.1694	1.9129
	OSC	0.9613	0.4110	0.7985	0.9374	0.8727	0.7982	2.8026
	SNV_SG	0.9525	0.4553	−0.1206	2.2104	0.4281	1.6916	1.3224
	SNV_OSC	0.9851	0.2547	0.9031	0.6502	0.9207	0.6301	3.5501
	Raw_SS	0.9257	0.5691	0.1842	1.8860	0.7318	1.1584	1.9311
	SNV_SS	0.9553	0.4416	0.0087	2.0790	0.3665	1.7805	1.2564
	SG_SS	0.9289	0.5568	0.1863	1.8836	0.7247	1.1738	1.9058
	OSC_SS	0.9617	0.4089	0.8030	0.9269	0.8939	0.7287	3.0699
	SNV_SG_SS	0.9548	0.4437	−0.1260	2.2158	0.3849	1.7545	1.2750
	SNV_OSC_SS	0.9842	0.2628	0.9019	0.6540	0.9206	0.6305	3.5479

PLSR – partial least squares regression; SVR - Support vector regression; ENR - Elastic Net regression; GBR - Gradient Boosting regression; RFR - Random Forest regression; SNV – standard normal variate; SG – Savitzky-Golay filtering; OSC – orthogonal signal correction; SS – StandardScaler; R_c^2 – correlation coefficient of calibration; R_{cv}^2 – correlation coefficient of cross validation; R_p^2 – correlation coefficient of prediction; RMSEC – root square error of calibration; RMSECV – root mean square error of cross validation; RMSEP – root mean square error of prediction; RPD – residual predictive deviation.

preprocessing combinations due to their ability to handle normalized and linearized data more effectively. Notably, the application of a series of redundant preprocessing using SNV, OSC, and StandardScaler led to improvements in prediction accuracy in all models. This preprocessing combination ensured that all the models achieved RPD values above 3.0, signifying their reliable predictive capacity (Mishra et al., 2022). Among these, the SVR model exhibited exceptional predictive properties after the application of SNV-OSC-StandardScaler preprocessing. With metrics such as R_p^2 of 0.9588, RMSEP of 0.4539 %, and RPD of 4.9287, the model showed paramount efficiency in predicting protein content in dried laver. The scatter plots shown in Fig. S1 visually illustrate the predictive performance of these machine-learning models against the validation

set.

The relatively inferior performance of ensemble models such as GBR and RFR could be attributed to the following reasons. The variation in effectiveness could be due to the complexity of the GBR and RFR algorithms, which seem more prone to overfitting in datasets with inherently linear relationships between spectral data and protein content. It appears that the spectral linearity characteristics of the SWIR hyperspectral imaging dataset related to protein content might not have been complex enough to fully exploit the modeling capabilities of GBR and RFR. According to the research conducted by Kästner et al., RFR showed superior performance compared to PLSR in learning from heterogeneous samples. This implies that, in some cases, simpler models may be more effective, particularly when dealing with dried laver manufactured by homogenization and subsequent drying (Kästner et al., 2022). Thus, the algorithms' complexity might not align well with the simpler, linear nature of the dataset in question. Additionally, the relatively small dataset could have impeded the ability of these models to effectively generalize, potentially making simpler models such as SVR and PLSR more suitable for this specific context (Tian et al., 2023).

These results suggest the importance of carefully selecting modeling techniques and preprocessing methods that align with the dataset's nature and the analytical objectives, indicating that such a meticulous approach can notably enhance model performance.

3.3. Informative wavelength selection by CARS

The application of CARS in this study facilitated the meticulous selection of informative wavelengths from the raw SWIR spectra of dried laver samples, focusing on protein content analysis. This precision ensures that the analysis remains aligned with the protein's chemical makeup, thereby enhancing the robustness of subsequent analyses.

The methodology adopted involved Monte Carlo sampling with a total of 50 runs. This optimization process is graphically represented in Fig. 3, which details the variation in the selection of wavelengths, the progression of RMSECV values, and the evolution of regression coefficients through the increase in Monte Carlo sampling runs. The procedure of wavelength selection via CARS was methodically divided into two stages: an initial aggressive reduction and a subsequent precise refinement, depicted in Fig. 3a. At the outset, a sharp decrease in the number of selected wavelengths was observed, particularly noticeable when the number of sampling runs was low. This reduction then transitioned to a more moderate decline. This pattern was a result of employing an exponentially decreasing function aimed at swiftly identifying and discarding the less informative wavelengths, thereby stabilizing the selection process as the number of sampling runs approached 15. The RMSECV's trajectory, as showcased in Fig. 3b, underscores a declining trend as the sampling runs progressed up to the 15th iteration, marking the point where the lowest RMSECV of 0.1348 was achieved. This optimal juncture signifies the efficacy of the wavelength selection up to this point, beyond which any additional exclusion of wavelengths risked omitting potentially significant spectral features, justifying the cessation of the selection process at this stage. Fig. 3c illustrates the critical role of the regression coefficient trajectories of each wavelength throughout the sampling process, showcasing how each contributes differently across stages. A distinct blue line highlights the optimal subset of wavelengths that yielded the minimum 5-fold RMSECV, underlining their pivotal role in enhancing model precision. Through the CARS algorithm, 56 wavelengths were identified as significant, including 921.6 nm, 978.5 nm, and extending through to 1690.8 nm,

Table 2
Effective wavelengths selected from full spectrum (900–1700 nm) using competitive adaptive reweighted sampling (CARS) for prediction of protein in dried laver.

Method	Number of wavelengths	Wavelength (nm)
Raw	215	900–1700
CARS	56	979, 982, 997, 1010, 1020, 1037, 1044, 1048, 1062, 1073, 1141, 1161, 1181, 1242, 1246, 1250, 1262, 1291, 1299, 1307, 1323, 1332, 1336, 1352, 1356, 1395, 1403, 1407, 1411, 1442, 1449, 1464, 1468, 1479, 1483, 1486, 1504, 1518, 1522, 1525, 1539, 1553, 1559, 1562, 1566, 1572, 1579, 1594, 1607, 1610, 1621, 1655, 1668, 1671, 1676, 1691

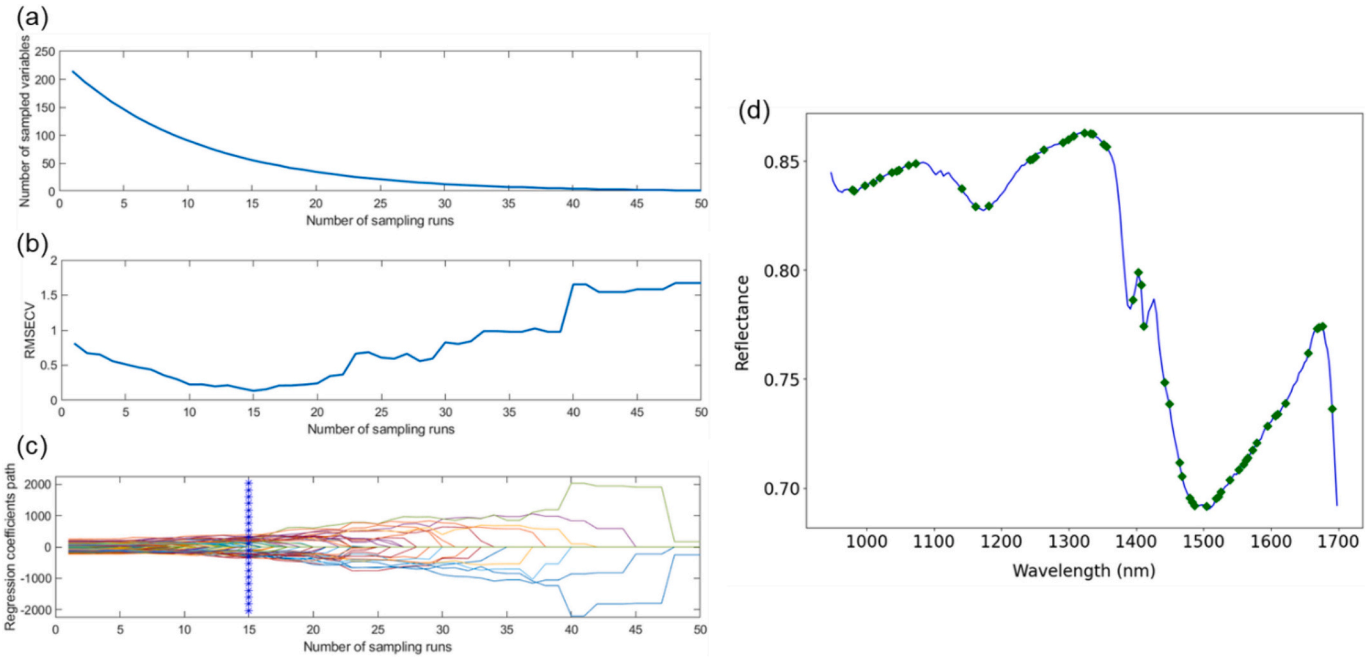


Fig. 3. Informative wavelength screening by CARS (a) change trend of the variables number in the change process diagram of CARS algorithm, (b) 5-fold RMSECV values, (c) the regression coefficient path of each wavelength with the increase of Monte-Carlo sampling runs, and (d) raw spectra with the region of wavelengths selected by CARS.

which resulted in a notable reduction of spectral data by 74.0 % (Table 2). This reduction underscores the efficiency of CARS in isolating the most pertinent spectral features for protein analysis, as depicted in Fig. 4d, where both the entire raw spectra and precise locations of the selected wavelengths are illustrated.

In the presented analysis, the absorption characteristics within the Short-Wave Infrared (SWIR) region, particularly around 1050, 1250, and 1400–1650 nm, offer significant insights into the presence and structural properties of proteins. The absorption near 1050 nm is largely attributed to the C—H stretching vibrations, common in organic compounds and indicative of the protein's structural framework. At 1250 nm, the observed absorption peaks are due to O—H and N—H bending vibrations, highlighting the specific molecular interactions within protein molecules (Rodríguez-Pulido et al., 2014). Furthermore, the broad range between 1400 and 1650 nm encompasses critical absorption bands associated with amide I and II vibrations, which are directly related to the protein's secondary structure, including aspects like α-helices and β-sheets (Golovynskyi et al., 2023; Niemi et al., 2023). These spectral regions, therefore, provide a robust basis for the detection and analysis of proteins, leveraging their unique molecular vibrations to elucidate protein content and structural information in various samples.

3.4. Predicting protein content based on informative wavelengths

In the analysis of dried laver for protein content prediction using SWIR hyperspectral imaging, the CARS methodology was pivotal in distilling the dataset to 56 significantly informative wavelengths from an initial set of 215 spectra. This strategic reduction optimized the analytical process and highlighted the efficiency of CARS in identifying the most relevant spectral features for protein estimation. The results of preprocessing and regression model training utilizing the wavelengths selected by CARS are presented in Table 3. And, In Fig. 4, scatter plots depict the comparison between measured and predicted protein content using the selected wavelengths by CARS, employing optimal preprocessing methods and regression models, including PLSR, SVR, ENR, GBR, and RFR.

The CARS implementation for wavelength selection was instrumental in enhancing the performance of the PLSR and SVR models. Notably, among the various model configurations examined, the SVR model preprocessed with StandardScaler after SNV and OSC transformations (SNV-OSC-StandardScaler-SVR) stood out, delivering superior performance when trained on CARS-selected wavelengths compared to the full spectrum approach. This refined model configuration significantly improved the key performance metrics, including R_p^2, RMSEP, and RPD. Specifically, the SNV-OSC-Standard Scaler-SVR model achieved an R_p^2 of 0.9673, indicating its accuracy in predicting protein content. Furthermore, the model attained an RMSEP of 0.4043, reflecting its precision in estimating protein levels in the dried laver samples. The RPD value of 5.533 further underscores the robustness and reliability of the model in prediction, indicating a substantial enhancement over the results obtained from the models trained on the full spectral data. Similarly, for PLSR and ENR, although R_c^2 slightly decreased, R_p^2 increased when using CARS-selected wavelengths. This indicates that using only the most effective wavelengths prevented overfitting and enhanced the robustness of the models.

The notable performance enhancement observed with the ENR model, specifically when the dataset was preprocessed with StandardScaler and informed by wavelengths selected through CARS, indicates the inherent modeling strengths of ENR. By design, ENR is adept at handling situations with high-dimensional data, where the number of predictors exceeds the number of observations. This is achieved by incorporating both L1 and L2 regularization, which facilitates feature selection and shrinkage. Applying StandardScaler to the CARS-selected wavelengths normalized the dataset, ensuring that each feature contributed evenly to the predictive capability of the model. This normalization was crucial for models such as ENR, in which the regularization terms were sensitive to the scale of the variables. By balancing the scale, StandardScaler ensures that the regularization applied by ENR is more effective, leading to improved model performance, even with a reduced feature set. This highlights the capability of the model to

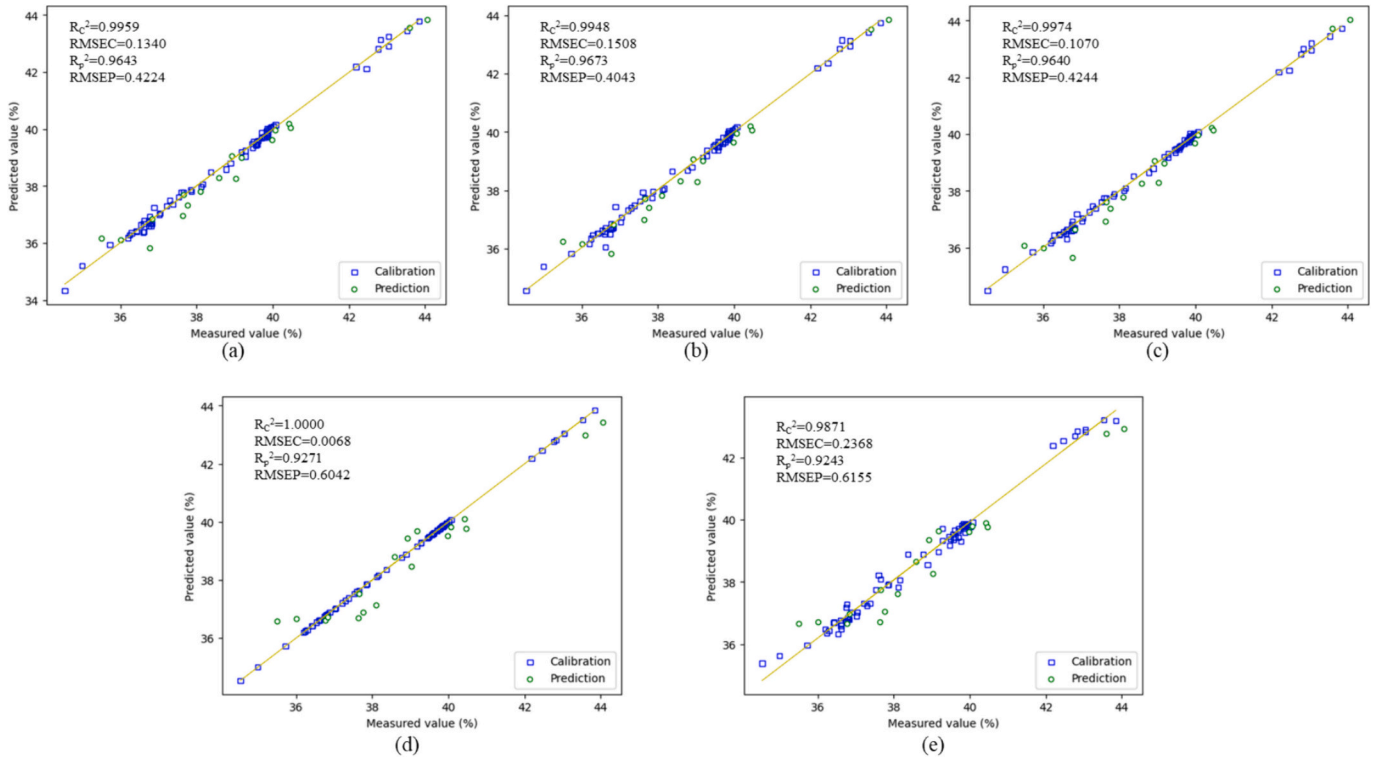


Fig. 4. Scatter plots of the measured vs predicted protein content using the selected wavelengths by CARS under the optimal preprocessing methods and regression models: (a) PLSR, (b) SVR, © ENR, (d) GBR, and (e) RFR.

Table 3
Quantitative protein prediction performance based on informative wavelengths.

Models	Preprocessing	Calibration set		Validation set		Prediction set		RPD
		R_c^2	RMSEC	R_{cv}^2	RMSECV	R_p^2	RMSEP	
PLSR	Raw	0.9999	0.0246	0.9949	0.1493	0.8125	0.9687	2.3092
	SNV	0.9874	0.2348	0.8949	0.6769	0.8993	0.7097	3.1519
	SG	0.9689	0.3684	0.8716	0.7481	0.8700	0.8064	2.7740
	OSC	0.9999	0.0163	0.9977	0.1000	0.9095	0.6729	3.3245
	SNV_SG	0.9609	0.4127	0.8576	0.7879	0.8149	0.9624	2.3243
	SNV_OSC	0.9968	0.1184	0.9762	0.3225	0.9570	0.4639	4.8217
	Raw_SS	0.9898	0.2108	0.9432	0.4978	0.7938	1.0159	2.2020
	SNV_SS	0.9811	0.2871	0.9141	0.6120	0.8805	0.7733	2.8929
	SG_SS	0.9543	0.4464	0.7762	0.9878	0.8342	0.9110	2.4556
	OSC_SS	0.9994	0.0499	0.9894	0.2145	0.9016	0.7018	3.1874
	SNV_SG_SS	0.9645	0.3936	0.8762	0.7347	0.8112	0.9719	2.3017
	SNV_OSC_SS	0.9959	0.1340	0.9748	0.3317	0.9643	0.4224	5.2956
SVR	Raw	0.5878	1.3406	0.2862	1.7642	0.6852	1.2551	1.7823
	SNV	0.9463	0.4841	0.7474	1.0494	0.8825	0.7669	2.9169
	SG	0.5886	1.3392	0.2844	1.7664	0.6864	1.2526	1.7858
	OSC	0.8395	0.8364	0.8229	0.8786	0.9024	0.6988	3.2011
	SNV_SG	0.8484	0.8129	0.6961	1.1511	0.7894	1.0266	2.1790
	SNV_OSC	0.9556	0.4400	0.9287	0.5574	0.9623	0.4341	5.1535
	Raw_SS	0.9998	0.0286	0.9955	0.1408	0.8125	0.9687	2.3092
	SNV_SS	0.9820	0.2805	0.9104	0.6249	0.8956	0.7227	3.0953
	SG_SS	0.9631	0.4012	0.8525	0.8020	0.8801	0.7747	2.8877
	OSC_SS	0.9999	0.0156	0.9964	0.1251	0.9178	0.6412	3.4885
	SNV_SG_SS	0.9567	0.4347	0.8802	0.7227	0.8415	0.8906	2.5118
	SNV_OSC_SS	0.9948	0.1508	0.9758	0.3251	0.9673	0.4043	5.5334
ENR	Raw	0.2710	1.7828	−0.0849	2.1749	0.2877	1.8879	1.1849
	SNV	0.0000	2.0881	−0.1933	2.2810	−0.0097	2.2478	0.9952
	SG	0.2708	1.7830	−0.0850	2.1750	0.2876	1.8881	1.1848
	OSC	0.5355	1.4232	0.3073	1.7379	0.5641	1.4769	1.5147
	SNV_SG	0.4808	1.5046	−0.1332	2.2228	0.4335	1.6836	1.3287
	SNV_OSC	0.6187	1.2893	0.4513	1.5467	0.7111	1.2023	1.8605
	Raw_SS	0.8291	0.8631	0.6079	1.3074	0.7781	1.0538	2.1227
	SNV_SS	0.9799	0.2958	0.8990	0.6638	0.8888	0.7458	2.9993
	SG_SS	0.8136	0.9016	0.6002	1.3203	0.7735	1.0647	2.1010
	OSC_SS	0.9446	0.4914	0.9183	0.5967	0.9293	0.5948	3.7610
	SNV_SG_SS	0.9593	0.4210	0.8635	0.7715	0.8557	0.8499	2.6321
	SNV_OSC_SS	0.9974	0.1070	0.9772	0.3155	0.9640	0.4244	5.2714
GBR	Raw	0.7273	1.0903	0.2146	1.8505	0.6140	1.3898	1.6095
	SNV	0.9997	0.0332	−0.2356	2.3211	0.2440	1.9450	1.1501
	SG	0.8785	0.7278	0.1919	1.8770	0.7064	1.2122	1.8454
	OSC	0.9839	0.2648	0.7979	0.9388	0.8925	0.7335	3.0495
	SNV_SG	1.0000	0.0045	−0.2735	2.3564	0.2023	1.9979	1.1197
	SNV_OSC	1.0000	0.0068	0.8797	0.7242	0.9271	0.6042	3.7025
	Raw_SS	0.7273	1.0903	0.1841	1.8861	0.6123	1.3929	1.6060
	SNV_SS	1.0000	0.0040	−0.2400	2.3252	0.2344	1.9573	1.1429
	SG_SS	0.8785	0.7278	0.2072	1.8592	0.7050	1.2150	1.8411
	OSC_SS	0.9839	0.2648	0.7979	0.9387	0.8924	0.7338	3.0483
	SNV_SG_SS	1.0000	0.0001	−0.2966	2.3777	0.2830	1.8942	1.1810
	SNV_OSC_SS	1.0000	0.0000	0.8753	0.7374	0.9173	0.6432	3.4778
RFR	Raw	0.9162	0.6045	0.1436	1.9324	0.6735	1.2783	1.7500
	SNV	0.9587	0.4242	−0.2457	2.3305	0.1541	2.0574	1.0873
	SG	0.9098	0.6273	0.2007	1.8668	0.6988	1.2278	1.8220
	OSC	0.9587	0.4242	0.8097	0.9108	0.8897	0.7430	3.0107
	SNV_SG	0.9516	0.4593	−0.1957	2.2833	0.0879	2.1364	1.0471
	SNV_OSC	0.9871	0.2368	0.8942	0.6792	0.9243	0.6155	3.6345
	Raw_SS	0.9237	0.5769	0.1814	1.8892	0.6886	1.2483	1.7920
	SNV_SS	0.9372	0.5232	−0.2281	2.3140	0.1597	2.0505	1.0909
	SG_SS	0.9184	0.5963	0.1801	1.8908	0.6982	1.2290	1.8202
	OSC_SS	0.9586	0.4246	0.8108	0.9083	0.8906	0.7397	3.0240
	SNV_SG_SS	0.9394	0.5142	−0.1941	2.2817	−0.0016	2.2388	0.9992
	SNV_OSC_SS	0.9873	0.2349	0.9054	0.6422	0.9210	0.6287	3.5583

PLSR – partial least squares regression; SVR - Support vector regression; ENR - Elastic Net regression; GBR - Gradient Boosting regression; RFR - Random Forest regression; SNV – standard normal variate; SG – Savitzky-Golay filtering; OSC – orthogonal signal correction; SS – StandardScaler; R_c^2 – correlation coefficient of calibration; R_{cv}^2 – correlation coefficient of cross validation; R_p^2 – correlation coefficient of prediction; RMSEC – root square error of calibration; RMSECV – root mean square error of cross validation; RMSEP – root mean square error of prediction; RPD – residual predictive deviation.

leverage distilled yet highly relevant spectral information, enhancing its predictive precision for protein content in dried laver samples. This demonstrates the nuanced interplay among feature selection, model complexity, and data preprocessing to achieve optimal model performance.

The GBR and RFR did not exhibit performance improvements across the various preprocessing techniques, including those processed with StandardScaler. In contrast, these models sometimes showed decreased

performance, as evidenced by key metrics, such as R_p^2, RMSEP, and RPD. For instance, in the case of OSC-SNV-StandardScaler-GBR, the best performance noted was an R_p^2 of 0.9417, an RMSEP of 0.5400, and an RPD of 4.1428, whereas after applying CARS-selected wavelengths and preprocessing, these performance indicators decreased to an R_p^2 of 0.9173, an RMSEP of 0.6432, and an RPD of 3.478.

The lack of performance improvement or degradation of the GBR and RFR models can be attributed to several factors. The complexity of the GBR and RFR algorithms, which are more prone to overfitting in datasets with inherently linear relationships between spectral data and protein content, could be a significant factor. The spectral linearity characteristics of the SWIR hyperspectral imaging dataset might not have been complex enough to fully exploit the modeling capabilities of GBR and RFR. Furthermore, ensemble learning methods like GBR and RFR may not align well with the reduced feature sets provided by CARS, as they often require more diverse data to generalize effectively. In contrast, SVR's application of structural risk minimization and its suitability for smaller sample sizes make it more appropriate for this context than GBR or RFR, as it can more effectively address collinearity issues (Meiyan et al., 2023). Furthermore, the potential under- or over-fitting issue with GBR and RFR when dealing with a smaller set of selected wavelengths could also contribute to the observed performance dip. Although these models handle high-dimensional data by constructing numerous decision trees to improve prediction accuracy, the narrowed-down feature set may not provide sufficient diversity in the data for them to generalize well (Shafagh-Kolvanagh et al., 2022).

This analysis underscores the importance of matching the characteristics of predictive models with the nature of the processed data, highlighting that more sophisticated or complex models are not always the most suitable choice for every dataset or analytical goal. These findings suggest the need for further research to optimize model selection and feature set refinement to enhance predictive modeling in the context of hyperspectral imaging for food quality assessment.

3.5. *Visualization of protein in dried laver*

Employing the developed imaging algorithm, the CARS-SNV-OSC-Standard Scaler-SVR model was applied across the entire pixel matrix of the image ROIs for the dried laver samples. This resulted in the creation of color maps that sharply delineated the protein distribution; illustrative examples are presented in Fig. 5. The color spectrum within these maps transitioned from purple to red, indicating ascending mean protein content values. Thus, the color maps rendered the protein content variation and distribution in the samples readily apparent, enabling direct visual assessment. In essence, this innovative combination of predictive modeling and visual mapping allows for a more objective and precise evaluation of the protein content of dried laver.

4. Conclusions

The exploratory study presented here sheds light on the efficacy of SWIR hyperspectral imaging, spanning the 900–1700 nm spectrum, as a rapid, non-destructive analytical tool for assessing multiple quality parameters in dried laver. Spectral data preprocessed using a combination of SNV, OSC, and StandardScaler proved to be highly effective for protein prediction. SVR models leveraging the SNV-OSC-StandardScaler preprocessed spectra exhibited commendable predictive capabilities with an R_p^2 of 0.9588, an RMSEP of 0.4539, and an RPD of 4.9287. Furthermore, a CARS method was utilized to select 56 informative wavelengths from the raw spectra. The refined CARS-SNV-OSC-Standard Scaler-SVR models demonstrated superior performance with enhanced predictive abilities for protein content, achieving an R_p^2 of 0.9673, RMSEP of 0.4043, and RPD of 5.533. Hyperspectral imaging within the 900–1700 nm range holds substantial promise for the quality assessment of dried laver. The results of this study pave the way for the broader adoption of SWIR hyperspectral imaging as a reliable, non-invasive method for determining the nutritional components of seaweeds, underpinning its potential as a cornerstone technology for the expansion of seaweed and marine food industries.

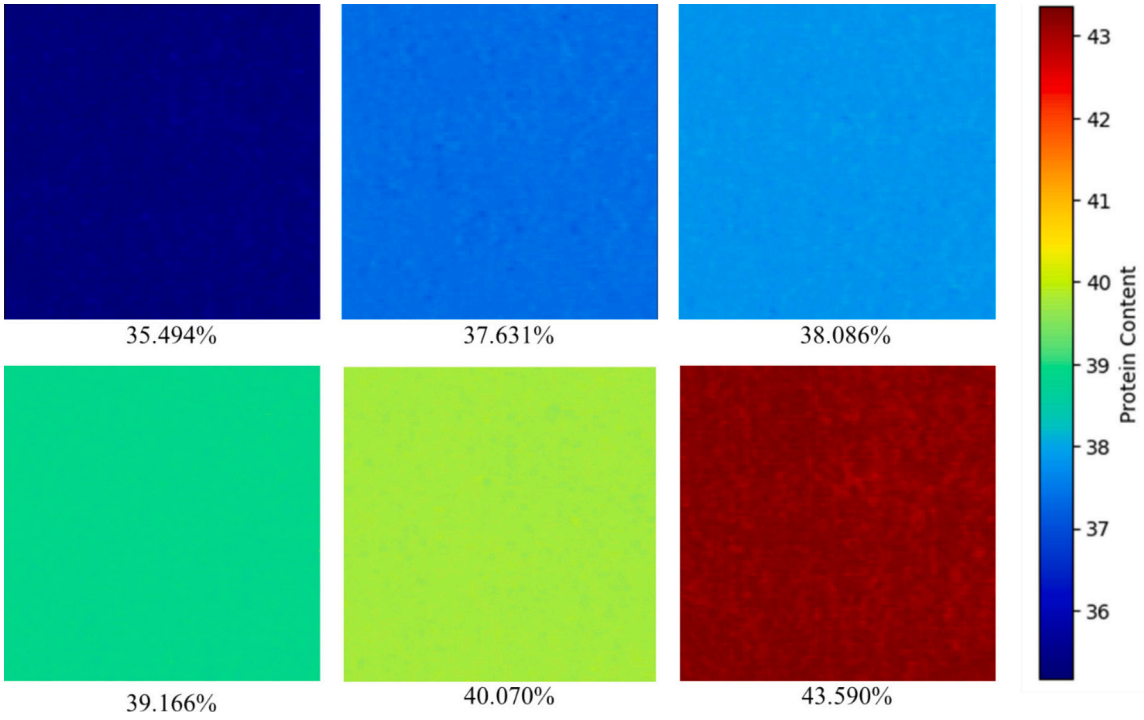


Fig. 5. Example of visualization maps of protein content in dried lavers.

CRediT authorship contribution statement

Eunghee Kim: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Jong-Jin Park:** Writing – original draft, Investigation, Conceptualization. **Gyuseok Lee:** Software, Methodology, Investigation. **Jeong-Seok Cho:** Writing – review & editing, Supervision, Methodology. **Seul-Ki Park:** Writing – review & editing, Methodology. **Dae-Yong Yun:** Visualization, Software. **Kee-Jai Park:** Resources, Project administration, Conceptualization. **Jeong-Ho Lim:** Writing – original draft, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that may have influenced the work reported in this study.

Data availability

The data that has been used is confidential.

Acknowledgements

This research was supported by the [1)]Korea Institute of Marine Science & Technology Promotion (KIMST), funded by the Ministry of Oceans and Fisheries (20210695); [2)]the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) through the High Value-Added Food Technology Development Program funded by the Ministry of Agriculture, Food, and Rural Affairs (MAFRA) (321049-5); and [3)]the Main Research Program (E0211001) of Korea Food Research Institute (KFRI) funded by the Ministry of Science and ICT.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101763.

References

Aheto, J. H., Huang, X., Tian, X., Lv, R., Dai, C., Bonah, E., & Chang, X. (2020). Evaluation of lipid oxidation and volatile compounds of traditional dry-cured pork belly: The hyperspectral imaging and multi-gas-sensory approaches. *Journal of Food Process Engineering, 43*(1), Article e13092.

Amirvaresi, A., Nikounezhad, N., Amirahmadi, M., Daraei, B., & Parastar, H. (2021). Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy based on chemometrics for saffron authentication and adulteration detection. *Food Chemistry, 344*, Article 128647.

Bermolen, P., & Rossi, D. (2009). Support vector regression for link load prediction. *Computer Networks, 53*(2), 191–201.

Cheng, J.-H., & Sun, D.-W. (2017). Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food Engineering Reviews, 9*, 36–49.

Dai, Q., Sun, D. W., Cheng, J. H., Pu, H., Zeng, X. A., & Xiong, Z. (2014). Recent advances in de-noising methods and their applications in hyperspectral image processing for the food industry. *Comprehensive Reviews in Food Science and Food Safety, 13*(6), 1207–1218.

Elmasry, G., Kamruzzaman, M., Sun, D.-W., & Allen, P. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: A review. *Critical Reviews in Food Science and Nutrition, 52*(11), 999–1023.

Fatemi, A., Singh, V., & Kamruzzaman, M. (2022). Identification of informative spectral ranges for predicting major chemical constituents in corn using NIR spectroscopy. *Food Chemistry, 383*, Article 132442.

Figueroa, V., Farfán, M., & Aguilera, J. (2023). Seaweeds as novel foods and source of culinary flavors. *Food Reviews International, 39*(1), 1–26.

García-Poza, S., Cotas, J., Morais, T., Pacheco, D., Pereira, L., Marques, J. C., & Gonçalves, A. M. (2022). Global trade of seaweed foods. In *Sustainable global resources of seaweeds volume 2: Food, pharmaceutical and health applications* (pp. 325–337). Springer.

Golden, C. E., Rothrock, M. J., Jr., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. *Food Research International, 122*, 47–55.

Golovynskyi, S., Golovynska, I., Roganova, O., Golovynskyi, A., Qu, J., & Ohulchanskyy, T. Y. (2023). Hyperspectral imaging of lipids in biological tissues using near-infrared and shortwave infrared transmission mode: A pilot study. *Journal of Biophotonics, 16*(7), Article e202300018.

He, H.-J., Wang, Y., Zhang, M., Wang, Y., Ou, X., & Guo, J. (2022). Rapid determination of reducing sugar content in sweet potatoes using NIR spectra. *Journal of Food Composition and Analysis, 111*, Article 104641.

Jeong, G.-T., Lee, C., Cha, E., Moon, S., Cha, Y.-J., & Yu, D. (2023). Determination of optimum processing condition of high protein laver Chip using air-frying and reaction flavor technologies. *Foods, 12*(24), 4450.

Jin, X., Wang, L., Zheng, W., Zhang, X., Liu, L., Li, S., … Xuan, J. (2022). Predicting the nutrition deficiency of fresh pear leaves with a miniature near-infrared spectrometer in the laboratory. *Measurement, 188*, Article 110553.

Kang, Z., Zhao, Y., Chen, L., Guo, Y., Mu, Q., & Wang, S. (2022). Advances in machine learning and hyperspectral imaging in the food supply chain. *Food Engineering Reviews, 14*(4), 596–616.

Kästner, F., Sut-Lohmann, M., Ramezany, S., Raab, T., Feilhauer, H., & Chabrillat, S. (2022). Estimating heavy metal concentrations in Technosols with reflectance spectroscopy. *Geoderma, 406*, Article 115512.

Loggenberg, K., Strever, A., Greyling, B., & Poona, N. (2018). Modelling water stress in a shiraz vineyard using hyperspectral imaging and machine learning. *Remote Sensing, 10*(2), 202. https://doi.org/10.3390/rs10020202

Manthou, E., Karnavas, A., Fengou, L.-C., Bakali, A., Lianou, A., Tsakanikas, P., & Nychas, G.-J. E. (2022). Spectroscopy and imaging technologies coupled with machine learning for the assessment of the microbiological spoilage associated to ready-to-eat leafy vegetables. *International Journal of Food Microbiology, 361*, Article 109458.

Marques de Brito, B., Campos, V.d. M., Neves, F. J., Ramos, L. R., & Tomita, L. Y. (2023). Vitamin B12 sources in non-animal foods: A systematic review. *Critical Reviews in Food Science and Nutrition, 63*(26), 7853–7867.

Meiyan, S., Jinyu, Z., Xiaohong, Y., Xiaohe, G., Baoguo, L., & Yuntao, M. (2023). A spectral decomposition method for estimating the leaf nitrogen status of maize by UAV-based hyperspectral imaging. *Computers and Electronics in Agriculture, 212*, Article 108100.

Mishra, G., Panda, B. K., Ramirez, W. A., Jung, H., Singh, C. B., Lee, S.-H., & Lee, I. (2022). Application of SWIR hyperspectral imaging coupled with chemometrics for rapid and non-destructive prediction of Aflatoxin B1 in single kernel almonds. *Lwt, 155*, Article 112954.

Msuya, F. E., Bolton, J., Pascal, F., Narrain, K., Nyonje, B., & Cottier-Cook, E. J. (2022). Seaweed farming in Africa: Current status and future potential. *Journal of Applied Phycology, 34*(2), 985–1005.

Murai, U., Yamagishi, K., Kishida, R., & Iso, H. (2021). Impact of seaweed intake on health. *European Journal of Clinical Nutrition, 75*(6), 877–889.

Nagy, M. M., Wang, S., & Farag, M. A. (2022). Quality analysis and authentication of nutraceuticals using near IR (NIR) spectroscopy: A comprehensive review of novel trends and applications. *Trends in Food Science & Technology, 123*, 290–309.

Niemi, C., Mortensen, A. M., Rautenberger, R., Matsson, S., Gorzsas, A., & Gentili, F. G. (2023). Rapid and accurate determination of protein content in North Atlantic seaweed by NIR and FTIR spectroscopies. *Food Chemistry, 404*, Article 134700.

Özdoğan, G., Lin, X., & Sun, D.-W. (2021). Rapid and noninvasive sensory analyses of food products by hyperspectral imaging: Recent application developments. *Trends in Food Science & Technology, 111*, 151–165.

Ozturk, S., Bowler, A., Rady, A., & Watson, N. J. (2023). Near-infrared spectroscopy and machine learning for classification of food powders during a continuous process. *Journal of Food Engineering, 341*, Article 111339.

Raja, R., Hemaiswarya, S., Sridhar, S., Alagarsamy, A., Ganesan, V., Elumalai, S., & Carvalho, I. S. (2020). Evaluation of proximate composition, antioxidant properties, and phylogenetic analysis of two edible seaweeds. *Smart Science, 8*(3), 95–100.

Ribeiro, M. N., Carvalho, I. A., Fonseca, G. A., Lago, R. C., Rocha, L. C., Ferreira, D. D., … Pinheiro, A. C. (2021). Quality control of fresh strawberries by a random forest model. *Journal of the Science of Food and Agriculture, 101*(11), 4514–4522.

Rodríguez-Pulido, F. J., Hernández-Hierro, J. M., Nogales-Bueno, J., Gordillo, B., González-Miret, M. L., & Heredia, F. J. (2014). A novel method for evaluating flavanols in grape seeds by near infrared hyperspectral imaging. *Talanta, 122*, 145–150.

Shafagh-Kolvanagh, J., Dehghanian, H., Mohammadi-Nassab, A. D., Moghaddam, M., Raei, Y., Salmasi, S. Z., … Gholizadeh-Khajeh, B. (2022). Machine learning-assisted analysis for agronomic dataset of 49 Balangu (*Lallemantia iberica* L.) ecotypes from different regions of Iran. *Scientific Reports, 12*(1), 19237.

de Souza Zangirolami, M., Moreira, T. F. M., Leimann, F. V., Valderrama, P., & Março, P. H. (2023). Texture profile and short-NIR spectral vibrations relationship evaluated through Comdim: The case study for animal and vegetable proteins. *Food Control, 143*, Article 109290.

Tian, H., Wu, D., Chen, B., Yuan, H., Yu, H., Lou, X., & Chen, C. (2023). Rapid identification and quantification of vegetable oil adulteration in raw milk using a flash gas chromatography electronic nose combined with machine learning. *Food Control, 150*, Article 109758.

Wada, K., Tsuji, M., Nakamura, K., Oba, S., Nishizawa, S., Yamamoto, K., … Nagata, C. (2021). Effect of dietary nori (dried laver) on blood pressure in young Japanese children: An intervention study. *Journal of Epidemiology, 31*(1), 37–42.

Yong, W. T. L., Thien, V. Y., Rupert, R., & Rodrigues, K. F. (2022). Seaweed: A potential climate change solution. *Renewable and Sustainable Energy Reviews, 159*, Article 112222.

Zhang, J., Guo, Z., Ren, Z., Wang, S., Yin, X., Zhang, D., … Ma, C. (2023). A non-destructive determination of protein content in potato flour noodles using near-infrared hyperspectral imaging technology. *Infrared Physics & Technology, 130*, Article 104595.

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing, 26*(3), 1466–1481.

Zhu, M., Long, Y., Chen, Y., Huang, Y., Tang, L., Gan, B., … Xie, J. (2021). Fast determination of lipid and protein content in green coffee beans from different origins using NIR spectroscopy and chemometrics. *Journal of Food Composition and Analysis, 102*, Article 104055.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*(2), 301–320.