Food Control 160 (2024) 110357

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Predicting gypsum tofu quality from soybean seeds using hyperspectral imaging and machine learning

Amanda Malik [a], Billy Ram [b], Dharanidharan Arumugam [c], Zhao Jin [a], Xin Sun [b,**], Minwei Xu [a,*]

[a] Department of Plant Sciences, North Dakota State University, Fargo, ND, 58108, USA
[b] Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND, 58108, USA
[c] School of Sustainable Engineering and the Built Environment Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ, 85287, USA

ARTICLE INFO

Keywords:
Hyperspectral imaging
Tofu quality
XGBoost
Deep neural networks
Non-destructive inspection

ABSTRACT

Soybean seeds are a key ingredient for producing quality tofu. Conventional methods for assessing soybean seed quality for tofu are time-consuming and labor-intensive. This study employs hyperspectral imaging (HSI) and machine learning to rapidly predict gypsum tofu quality from soybean seeds. Two hundred soybean seed varieties were classified into four categories based on tofu quality using hierarchical clustering. Hyperspectral scans of the soybean seeds were captured in the 900–1700 nm range. Using the Extreme Gradient Boost (XGBoost) algorithm, ten critical wavelengths were identified that correlate with protein, carbohydrate, and oil contents. A Convolutional Neural Network (CNN) model was subsequently developed, trained on HSI data from the soybean categories. For new soybean seeds, this CNN model successfully categorized them into distinct quality classes with 96–99 % accuracy. Further validation through tofu production demonstrated the model's robustness in predicting key tofu quality parameters like yield, firmness, and springiness. Overall, this pioneering research enabled rapid, non-destructive prediction of tofu quality from soybean seeds using HSI and CNN. With further refinements, this approach could revolutionize soybean seed quality assessment.

1. Introduction

Soybeans are a significant nutritional source worldwide, offering a complete protein profile containing all essential amino acids, dietary fiber, vitamins, minerals, and essential fatty acids. The beans are utilized in various products, including soy sauce, miso, natto, tempeh, sufu, kinako, soymilk, tofu, abura-age, and yuba (Fukushima, 1991). Tofu, in particular, is a traditional Asian food consumed in East-Asian countries for centuries and has gained popularity in Western countries due to the rising trend of plant-based food (Ali, Tian, & Wang, 2021).

Tofu can be chemically described as a protein gel primarily composed of water, proteins, fats, and carbohydrates. Tofu production involves adding a coagulating agent to soymilk and pressing the resulting curd into a block. Two traditional coagulants for tofu are calcium sulfate and magnesium chloride, which resulted in gypsum and marinated tofu, respectively. In addition to the coagulants, the quality of tofu is closely linked to the protein, fat, and carbohydrate content. Protein quality influences tofu textures such as hardness, cohesiveness, and springiness. Fats and carbohydrates can affect tofu quality through their interaction with proteins (Ali et al., 2021). The protein content in soymilk relates to water holding and tofu yield, although the protein content of soybean seeds does not significantly correlate with tofu yield (Lim, Deman, Deman, & Buzzell, 1990). This suggests that protein quality, including protein subunit and amino acid composition, impacts tofu quality more than the protein content of soybeans (Stanojevic, Barac, Pesic, & Vucelic-Radovic, 2011). Take the protein subunits as an example, Cai and Chang (1999) found the 11S/7S ratio of protein subunits is positively related to the firmness of tofu. James and Yang (2014) found lack of the 11SA4 subunit would benefit the texture of tofu, while Meng, Chang, Gillen, and Zhang (2016) found 11SA3 subunit is an indicator for predicting the firmness of tofu.

Traditional methods for evaluating tofu quality assess yield, texture, and sensory attributes (Poysa, Woodrow, & Yu, 2006). However, these methods have shortcomings. They are labor-intensive, lack comparability due to variations in tofu processing parameters, and take a substantial amount of time, making them unfit for modern, rapid production capacities (Kurasch, Hahn, Miersch, Bachteler, & Würschum, 2018). Therefore, there is an urgent need for a swift, efficient,

* Corresponding author.
** Corresponding author.
E-mail addresses: xin.sun@ndsu.edu (X. Sun), minwei.xu@ndsu.edu (M. Xu).

https://doi.org/10.1016/j.foodcont.2024.110357
Received 20 September 2023; Received in revised form 21 January 2024; Accepted 1 February 2024
Available online 2 February 2024
0956-7135/Published by Elsevier Ltd.

standardized method for evaluating soybean quality concerning tofu products.

Hyperspectral imaging (HSI) offers a rapid, non-invasive, and cost-effective method for non-destructive seed inspection, enabling detailed analysis of the chemical composition, moisture content, and internal quality without requiring sample preparation or posing safety risks. HSI can cover a wide wavelength range from 400 to 15,000 nm, depending on the specific camera used. Near-Infrared (NIR) is a related concept, representing a specific subset of HSI that focuses on the near-infrared region, typically spanning from 900 to 2500 nm (Gao et al., 2021; Kandpal, Lee, Kim, Bae, & Cho, 2015; Kucha, Liu, Ngadi, & Claude, 2021; Medus, Saban, Francés-Víllora, Bataller-Mompeán, & Rosado--Muñoz, 2021). HSI has been used in various studies to predict seed quality and analyze the chemical composition, such as protein, fat, and carbohydrate, and functionality of seeds through spectral information. Each chemical component has a unique spectral signature that can be detected using HSI (Erkinbaev, Henderson, & Paliwal, 2017). Those chemical components are considered key factors for determining the tofu qualities. Several studies have shown the potential of HSI, especially the NIR region, as a rapid method for the evaluation of seed quality. Squeo et al. (2022) developed a method using NIR-HSI (900–1700 nm) to perform rapid, accurate, and nondestructive quality control of texturized vegetable protein (TVP). This innovative approach, a first in this context, combines spectroscopy and imaging to analyze TVP's chemical composition, including proteins, carbohydrates, lipids, ashes, and alpha-galactosides. They employed analytical techniques like Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR), achieving high predictive accuracy with models displaying R^2 values between 0.92 and 0.98. The study concluded that NIR-HSI is an effective tool for rapid and precise quality control in TVP production, promising streamlined manufacturing and consistent product quality in the plant-based meat. da Silva Medeiros et al. (2022) employed portable NIR (900–1700 nm) and NIR-HSI (900–2500 nm) to evaluate the quality of Brassicas seeds, focusing on oil content, fatty acid composition, and species classification. Using advanced chemometric techniques like PCA and PLSR, they developed models that effectively differentiated Brassicas species, achieving up to 100 % accuracy with NIR-HSI. They collected spectral data using a portable NIR spectrometer and a NIR-HSI camera, which was pre-processed to correct for light scattering and other external influences. The results revealed significant differences in the oil content and fatty acid profiles among different Brassicas species. The study concluded that NIR-HSI-based models were more effective in calibration and prediction than portable NIR models, demonstrating the robust potential of NIR-HSI technology for quality control in Brassicas seeds.

Analyzing HSI data is challenging due to its complexity and high dimensionality, making it difficult to extract meaningful information using traditional statistical methods (Iqbal, Sun, & Allen, 2014). Machine learning, however, can efficiently and accurately analyze high-dimensional data by learning patterns and relationships in data automatically (Gao et al., 2021). Within the context of hyperspectral imaging, machine learning can perform tasks such as classification, feature selection, regression, and anomaly detection. The ultimate goal of this research is to develop an NIR sensor to non-constructively investigate the quality of soybean seeds. The current research developed a machine learning model to predict gypsum tofu quality based on HSI data of soybean seeds. The objectives were to classify soybean seeds based on corresponding gypsum tofu quality, select featured wavelengths scanned by HSI, which focus on the wavelength at the NIR range (1000–1700 nm), and develop a predictive machine learning model using soybean HSI image data and corresponding tofu quality. This research has the potential to enhance the efficiency and accuracy of tofu quality evaluation, reduce waste and cost, and assist in soybean breeding and tofu manufacturing.

2. Materials and methods

2.1. Seeds and materials

Two hundred varieties of soybeans, harvested from North Dakota, Missouri, Minnesota, Illinois, and Ohio, were generously provided by the Agricultural Utilization Research Institute (Crookston, MN). Calcium sulfate was purchased from the local market.

2.2. Water uptake capacity of soybean seeds

Two kilograms of soybeans were soaked in 5 kgs of water for 16 h at 4 °C and afterward, the excess water was drained, and the soaked soybeans were weighed to estimate the water uptake of the beans (Meng et al., 2016).

$$\text{Water uptake} = (W_S - W_D)/W_D \quad \text{Equation (1)}$$

Where W_S (kg) indicates the weight of soaked soybeans and W_D (kg) indicates the weight of dry soybeans.

2.3. Preparation of gypsum tofu

The tofu process was adapted from Meng et al. (2016) with modifications. Briefly, dry soybeans (W_0) were soaked following **Method 2.2**. The soaked soybeans were milled into slurries using a grinder hopper assembled on the automatic soymilk and tofu machine (Model B003141, MASE TOFU MACHINE Co. Ltd, Japan). Ten liters of water were added during the grinding procedure. The steam cooking (95 °C) began automatically and lasted for 5 min. The soymilk exited via the catch pipe, while the okara exited through the pressure relief valve. The initially collected soymilk was weighed and recorded (W_1). In a pan, 11 kg of soymilk were weighed to make curd. The soymilk was cooled to 82 °C and placed in a pan. Then, 35 g of calcium sulfate was evenly dispersed in the soymilk. After 12 min, the curds were broken up with an edge scraper and whipped. After setting for 1 min, the curds were poured into the drain pan with a mesh cloth at the bottom. After 5 min, the curds were wrapped with the mesh cloth and moved to the assembled air presser with a rectangular plate (20 cm × 40 cm). The initial pressure was set at 0.5 MPa for 5 min, while the second pressure was set at 1 MPa for 15 min. The prepared tofu was soaked in cool water for 15 min, and the final weight (W_2) of tofu was recorded.

The formula for calculating tofu yield is:

$$\text{Tofu yield (kg/kg soybean seeds)} = W_2 \times (W_1/11)/W_0 \quad \text{Equation (2)}$$

2.4. Evaluation of tofu texture

The quality of the tofu was analyzed by a Texture analyzer using a Stable Micro System, model TA-XT2 (Texture Technologies Corp., White Plains, NY, USA). The cylinder-shaped samples (25 mm diameter) were obtained by vertically cutting the tofu using a cylindrical cutter with triplicates. The samples were pressed twice using a metal disc probe (60 mm diameter) to simulate a mouth bite. The Texture Analyzer recorded the hardness, springiness, and cohesiveness of the tofu (Beléia, Prudencio-Ferreira, Yamashita, Sakamoto, & Ito, 2004).

2.5. Classification of soybean seeds

An unsupervised pattern recognition technique, hierarchical clustering analysis (HCA), was used in order to classify the soybean seeds based on the tofu quality. Soybean seeds were classified based on the qualities of soybean seeds and tofu, such as seed water uptake rate, tofu yield, firmness, springiness, and cohesiveness. The data were standardized and processed with the Ward method. This method provides

not only the classification of samples based on the tofu quality but is also an important source of knowledge with which to create cross-validation groups used in machine learning (Xu, Jin, Lan, Rao, & Chen, 2019). Overall, soybean seeds were sorted into four classes based on this technology.

2.6. Hyperspectral scanning of soybean seeds

A total of 250 soybean seeds were arranged within a transparent circular dish. Subsequently, the hyperspectral scanning process was conducted thrice for each variety, resulting in the acquisition of images. The hyperspectral data were recorded using the camera (Specim FX17, Specim, Oulu, Finland). The sensor is a push-broom type that captures hyperspectral cube data in the range of 900–1700 nm, with a spectral resolution of 8 nm, and can record 224 bands. To record the data, the researchers used Specim's LabScanner 40 × 20 platform, which features a halogen light source, a camera mount, and a 400 × 200 mm translation sample stage (Fig. S1). To minimize external light interference, the data was recorded in a dark room with only the halogen bulbs of the platform as the light source. The researchers captured white and dark reference calibration images with each image, where the white reference was captured using a Teflon bar with >95 % reflectance, and the dark reference was captured by closing the sensor shutter. The data recording software used was Lumo Scanner. The kernels were placed in a Petri dish to minimize the inertia generated by the translation stage.

To mitigate the effects of illumination changes and dark current in the sensor, the researchers calibrated the reflectance of the hyperspectral image using the formula below:

$$R = \frac{I - I_B}{I_W - I_B} \quad \text{Equation (3)}$$

Where $\boldsymbol{R}$ is the hyperspectral image after the reflectance calibration, $\boldsymbol{I}$ is the original hyperspectral image, $\boldsymbol{I_w}$ is the white reference hyperspectral image of the diffuse reflection whiteboard with 99 % reflectance, and $\boldsymbol{I_B}$ is the dark reference hyperspectral image when the lens is covered (Feng, Makino, Oshita, & García Martín, 2018; He et al., 2022).

2.7. HSI image processing

The HSI images were imported into MATLAB 2022a (The MathWorks, Natick, Massachusetts) and stored in a 3-D array. Each pixel value was normalized along the band axis. To increase the amount of data available for classification, each image was randomly subdivided into 64 × 64 sub-pixel images. The selection of the 64 × 64 sub-pixel regions was carried out with the requirement that at least 30 % of the pixels represented soybean seeds. Data augmentation was performed by rotating these images to 90, 180, and 270°, as well as vertically and horizontally flipping the images. After these post-processing steps, a total of 25,000 images were generated for each class of soybean seeds. The processed images were saved in one document in a CSV format.

2.8. Feature selection of HSI

The feature selection method was adapted from (Yang et al., 2021) with modifications.

2.8.1. Data segregation

After image processing, the dataset (CSV file) was randomly shuffled using a uniform distribution to ensure robust cross-validation later in the process. The dataset was divided into variables (X) and labels (y). The features, stored in X, consisted of columns 1 to 224 from the dataset. The labels, stored in y, consisted of column 225, with a subtraction of 1 applied to adjust for zero-based indexing. The dataset was further split into a training set (80 % of data) and a testing set (20 % of data) with a random seed of 0 for reproducibility.

2.8.2. Model training and evaluation

In this study, three distinct machine learning algorithms were utilized: Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF), each chosen due to their unique characteristics. The SVM algorithm, configured with a linear kernel and a cost parameter set to 1, was selected for its effectiveness in high-dimensional spaces, a characteristic that makes it highly suitable for our hyperspectral data set. The XGBoost algorithm applied both with all available features and with a subset of features selected based on their importance scores (>0.008), was chosen due to its robustness to overfitting and its ability to handle a large number of features, making it ideal for feature importance analysis in our study. Finally, the RF algorithm, implemented using an ensemble of 250 decision trees, was selected for its inherent feature selection mechanism and ability to handle non-linear relationships, characteristics that are highly beneficial when dealing with complex hyperspectral data (Su et al., 2021).

The performance of the three algorithms was compared based on eight key parameters: calibration accuracy, prediction accuracy, correlation coefficients of calibration (r_c), correlation coefficients of prediction (r_p), coefficients of determination of calibration (R_c), coefficients of determination of prediction (R_p), root mean square error of calibration (*RMSEC*), and root mean square error of prediction (*RMSEP*). Following the detailed comparative analysis, the most efficient model was selected. In this chosen model, the ten most influential wavelengths were identified, and their respective importance scores were recorded. This step facilitated a deeper understanding of the spectral characteristics that significantly contributed to the performance of our most efficient model.

2.9. Model establishment using convolutional neural network (CNN)

CNN is a highly effective feed-forward network. CNN is advantageous for handling transformations such as titling, scaling, translation, and others. The CNN framework consists of two major components: the convolutional layer, which extracts features, and the pooling layer, which reduces the input data size. Using a variety of filters, the convolutional layer can extract the deep features. Using maximum or mean combinations, the pooling layer drastically reduces the number of parameters. By combining with one or more fully connected layers, the CNN outputs the highly refined features of an image.

Based on the findings from the feature selection process (as outlined in **Method 2.8**), 5760 64 × 64 × 10 dimensional images were selected to train the convolutional neural network (CNN). The network is composed of two convolutional layers with 32 and 64 filters each, both having a 2 × 2 filter size and a stride of 2. These layers were subsequently followed by a batch normalization layer, global max pooling, and four fully connected layers. The classifier was trained using 30 epochs and a randomly selected subset of 80 % of the images. The remaining 20 % of images were reserved as a test dataset for evaluating the performance model (Lv. Ming, Chen, & Wang, 2019). Ultimately, a predictive machine learning library was developed to facilitate future predictions based on this CNN model. The confusion matrix, specificity, precision, and sensitivity of this model were listed for understanding the model's performance.

2.10. External validation of predictive machine learning model

The external validation of the predictive machine learning library was conducted using four untested soybean samples. These samples underwent the hyperspectral scanning process as outlined in **Method 2.6** and the image processing procedure detailed in **Method 2.7**. The processed images were subsequently classified by the predictive machine learning library, developed in **Method 2.9**, into one of the categories defined in **Method 2.5**. Simultaneously, these four untested soybean samples were subjected to the tofu production process described in **Method 2.3** and the tofu quality evaluation method presented in **Method 2.4**. The quality of the resulting tofu was then

statistically compared to the tofu quality characteristics of the soybean category predicted by the machine learning library.

2.11. Statistical analysis

The tofu quality analysis was performed in triplicate. The data was further subjected to analysis of variance followed by Tukey's test with Statgraphics Plus 5.1 Software (Manugistics, Inc.). Differences at $p < 0.05$ were considered significant.

HCA was performed on JMP® Pro 15.0.0 (SAS Institute Inc.). ENVI 5.3 (ITT Visual Information Solutions, Boulder, UT) was used to compute the spectral values of each pixel within the region of interest. MATLAB R2022a (The MathWorks, Natick, Massachusetts) was used for image processing. A 1D CNN model was constructed utilizing Python 3.8.3 and Jupyter Notebook. The CPU-based architecture of the 1D CNN model was programmed using the well-known deep learning framework PyTorch (https://pytorch.org/).

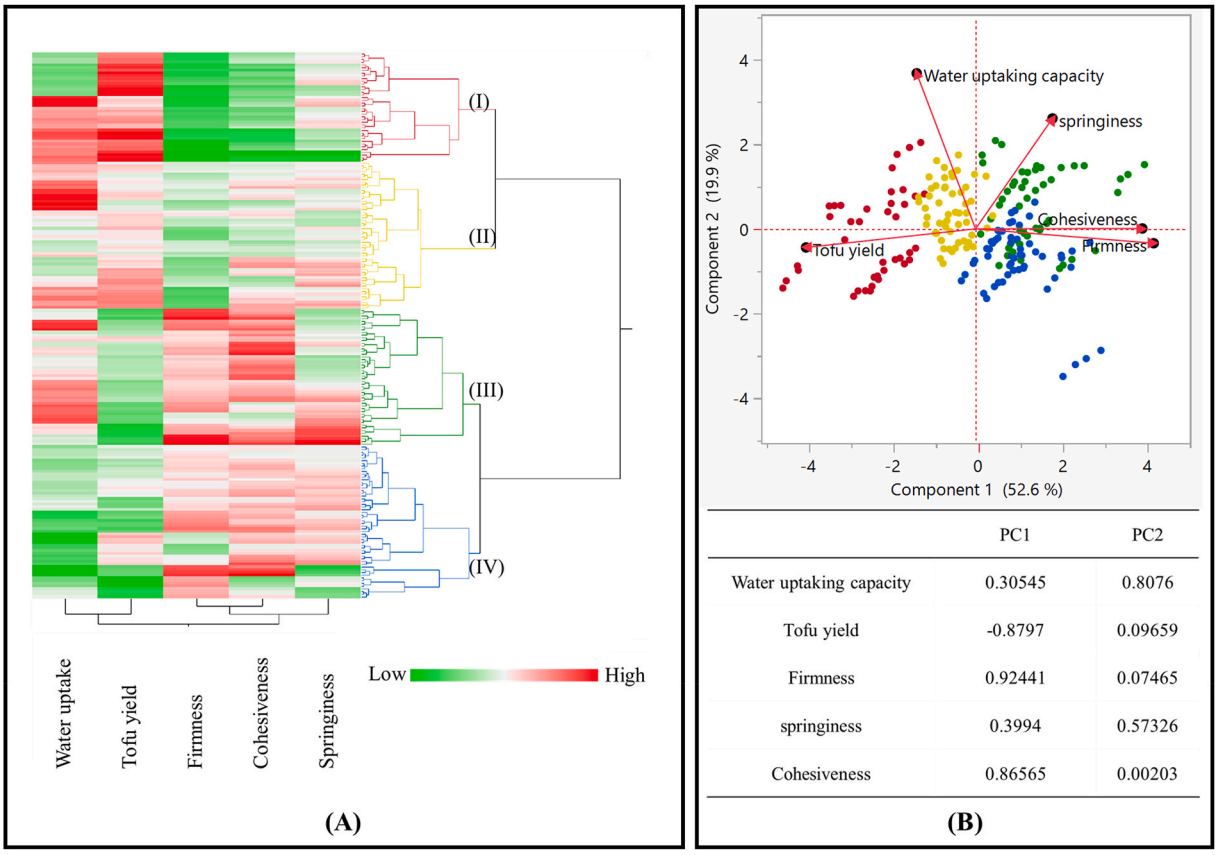
3. Results and discussion

3.1. Classification of soybean based on gypsum tofu quality

Hierarchical clustering analysis (HCA) was utilized to sort tofu samples into different classifications. Soybean seeds were divided into four classes based on the similarity between each group regarding water uptake of soybean, yield, firmness, cohesiveness, and springiness of tofu (Fig. 1A). PCA has also demonstrated similar results. The overall variance was explained by Principal Component 1 (PC1) and Principal Component 2 (PC2) by 72.5 %, with 52.6 % for PC1 and 19.9 % for PC2, respectively (Fig. 1B). Soybean seeds went from negative PC1 to positive PC1 following the group Class I, II, III, and IV. Class III and IV could not be well separated by PC1 alone; however, it is well separated by PC2. Class III was positive in PC2 while Class IV was negative in PC2. With the aid of Fig. 1B, it was observed that soybeans in Class I and II exhibited a high water-uptake capacity and yielded a high amount of tofu. Conversely, soybeans in Class III and IV displayed higher firmness, cohesiveness, and springiness of tofu. Overall, Class I soybeans had the highest water uptake and tofu yield compared to the other classes. Class II had a lower water uptake capacity and tofu yield than Class I, but higher than Class III and IV. Additionally, the results indicated a positive correlation between tofu yield and the water uptake capacity of soybean seeds. It is worth noting that Class III had a higher water-uptake capacity than Class IV, but both classes were characterized by higher values of tofu texture, such as firmness, cohesiveness, and springiness. The statistical data of Class I, II, III, and IV are listed in Table 1. The maximum tofu yield among the four classes was in Class I, about 3.6 kg/kg soybean seeds. The highest firmness and cohesiveness were found in Class III, which were 5.1 kg force and 0.67, respectively. The springiness had an insignificant difference ($p > 0.05$) in the four classes.

In chemical terms, tofu is primarily a water gel composed of protein, with smaller amounts of fats, carbohydrates, and minerals. Soybean protein goes through denaturation, coagulation, and molding to hold water and soluble in the protein gel (Chen, Hsieh, & Kuo, 2023). The chemical composition of soybean, and processing conditions, are two major factors that affect the final quality of tofu. In this research, processing conditions have been fixed while different soybean varieties indicated that different chemical compositions are considered the only factors that affect the tofu quality.

The yield of tofu is intimately tied to the water-uptaking capability of the soybean seeds, a characteristic that denotes the ability of soybeans to hydrate during tofu production. Soybeans with superior water uptaking capabilities generally produce higher tofu yields compared to their less



	PC1	PC2
Water uptaking capacity	0.30545	0.8076
Tofu yield	-0.8797	0.09659
Firmness	0.92441	0.07465
springiness	0.3994	0.57326
Cohesiveness	0.86565	0.00203

Fig. 1. Classification of soybean seeds based on gypsum tofu quality using (A) Hierarchical clustering analysis (HCA) and (B) Principal Component Analysis (PCA) and the loading score of each component. The color of Class I, II, III, and IV are indicated with red, yellow, green, and blue, respectively.

Table 1
Evaluation of gypsum tofu and soybean seed quality across various soybean classes.

	Class	Min	Max	Mean
Water uptake (kg/kg soybean)	I	1.21	1.34	1.27 ± 0.05 b
	II	1.23	1.34	1.28 ± 0.03 b
	III	1.24	1.33	1.28 ± 0.03 b
	IV	1.16	1.26	1.21 ± 0.03 a
Tofu yield (kg/kg soybean)	I	2.98	3.98	3.59 ± 0.31 c
	II	2.76	3.5	3.10 ± 0.21 b
	III	2.06	3.11	2.50 ± 0.23 a
	IV	1.87	3.26	2.62 ± 0.37 a
Firmness (g force)	I	1357	2246	1959 ± 292 a
	II	2255	3904	3082 ± 460 b
	III	3539	7189	5092 ± 1079 d
	IV	2627	6998	4483 ± 1012 c
Springiness	I	0.93	0.98	0.96 ± 0.01 a
	II	0.96	0.98	0.97 ± 0.01 a
	III	0.95	1.00	0.97 ± 0.01 a
	IV	0.94	0.98	0.97 ± 0.01 a
Cohesiveness	I	0.4	0.58	0.50 ± 0.04 a
	II	0.54	0.68	0.61 ± 0.04 b
	III	0.56	0.75	0.67 ± 0.05 d
	IV	0.51	0.77	0.64 ± 0.05 c

The sample number of classes I, II, III, IV is 40, 54, 50, and 56, respectively. Different letters indicate statistically significant differences ($p < 0.05$) by Tukey's test.

absorbent counterparts (Ali et al., 2021). This is because a higher water uptaking capacity suggests a greater water trapping capacity of the soybean protein. Ultimately, this leads to a higher yield of tofu since the weight of tofu is a sum of the weight of the solids and the absorbed water. Poysa and Woodrow (2002) investigated ten soybean lines grown at three locations for two years. They found that a higher water uptaking rate of the soybean seeds could result in a higher soymilk yield which was positively correlated with tofu yield per kilogram of soybeans.

Texture characteristics of tofu, including firmness, cohesiveness, and springiness, are fundamentally determined by the protein content and composition of the soybeans used in its production. Soybeans with a higher protein content typically produce tofu with enhanced firmness, cohesiveness, and springiness. However, it is noteworthy that as soybean seeds hydrate, the protein content becomes diluted, leading to a reduction in these texture attributes. This observation underpins the "Yield and Texture Trade-off Theory" that while high water uptaking capacity could lead to a high water content in the soymilk, resulting in a higher yield of tofu, it could simultaneously dilute the protein concentration in the soymilk. This dilution potentially diminishes tofu texture attributes such as firmness, cohesiveness, and springiness. Supporting this notion, Mujoo, Trinh, and Ng (2003) conducted a study on seven soybean varieties harvested from Michigan. Their research indicated that tofu firmness declined from 10.02 to 7.84 N as tofu yield increased from 2.93 to 3.43 kg/kg of soybeans, illustrating the balance between tofu yield and its textural attributes.

Contrarily, Class III soybeans serve as a counterexample to this theory, as their higher water-absorption capacity results in lower tofu yield and superior tofu texture. This implies that protein content is not the sole determinant of tofu quality and yield. Guan et al. (2021)underscored the influence of protein subunits on tofu yield and quality. To illustrate, soybeans with a lower 11S/7S ratio form a uniformly aggregated spherical gel, while beans with a higher 11S/7S ratio exhibit higher macroscopic phase separation, a coarser network structure, and larger pores (James & Yang, 2016). The role of amino acids in influencing tofu quality has also been reported. Coagulants such as calcium or magnesium salts are commonly used to bind the negatively charged amino acids together, forming a network-like structure (Ali et al., 2021; James & Yang, 2014). Given this information, it is plausible that the protein subunit composition and amino acid profile of Class III soybean seeds may vary significantly from those of Class IV.

In summary, soybean seeds from Class I, II, III, and IV each possess unique characteristics that influence the quality of tofu produced from them. The categorization of these soybean seeds provided valuable data for the application of supervised machine-learning techniques in the following research.

3.2. Hyperspectral images (HSI) of soybean seeds

3.2.1. Spectra of soybean HSI

The general trends of the HSI curves within the 900–1700 nm wavelength range were found to be quite similar (Fig. 2A). However, the peak intensity of each soybean seed varied, ranging from 40 to 120. To better understand the relationship between the HSI data and tofu quality, the spectra were averaged and grouped into the four previously established classes of soybeans using HCA (Fig. 2B). The intensity of the HSI spectra followed an overall order of Class I > II > III > IV, corresponding to the quality of tofu produced. These findings suggest that a predictive model could be established based on the HSI data and related parameters of tofu quality. However, with 224 wavelengths for each HSI curve, the dataset can be large, leading to potential computational complexity and noise in the predictive model. As such, the following methods will explore ways to reduce the number of wavelengths in the dataset.

Moreover, these observations imply that a predictive model could be built based on the HSI data and related parameters of tofu quality. However, as each HSI curve has 224 wavelengths, the dataset can be substantial, leading to potential computational complexity and noise in the predictive model (Loggenberg & Poona, 2020; Pal, Charan, & Poriya, 2021; Warner & Shank, 1997).

3.2.2. Selection of featured wavelengths

The HSI commonly includes highly correlated neighboring bands, causing problems of multicollinearity among closely positioned wavelength variables. To address this, featured wavelength selection is used to decrease data dimensionality and conserve storage space while preserving essential information. This strategy lessens collinearity issues, strengthens model resilience by reducing wavelength count, and potentially enhances model performance in accuracy and generalization.

Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF) were widely applied in searching for the featured wavelength of HSI (Huang et al., 2022; Pal et al., 2021). Support Vector Machine (SVM) is capable of handling high-dimensional data effectively. This is particularly important in HSI where the number of features (wavelengths) can be very large. By using a linear kernel, the SVM is looking for a linear combination of featured wavelengths that best separates the classes. This makes the interpretation of the model simpler, as the weight given to each wavelength in the final model represents its importance (Huang, Zhou, Meng, Wu, & He, 2017). XGBoost introduces a regularization term on the basis of the gradient boosting algorithm, utilizes the second-order Taylor expansion for fitting residuals, and can be calculated in parallel, so it has the advantages of *anti*-overfitting and high computational efficiency (Liao, Cao, Li, & Kang, 2019). RF is a robust, scalable, and flexible algorithm that can handle complex and noisy data, identify the most informative bands, capture non-linear relationships, and reduce overfitting in HSI analysis (Qin, Wang, Li, & Sam Ge, 2013).

In this study, both the XGBoost and RF algorithms showcased high calibration accuracy, exceeding 99 %, as represented in Table 2. However, the SVM algorithm exhibited a significantly lower calibration accuracy of just 53.8 %. Regarding prediction accuracy, despite its commendable performance, RF achieved a comparatively lower prediction accuracy of 56.4 %, suggesting that the featured wavelengths selected by this method had limited predictive power. The XGBoost algorithm stood out with a good prediction accuracy of 99.5 %, underscoring its superior capability in this context. These findings align with a

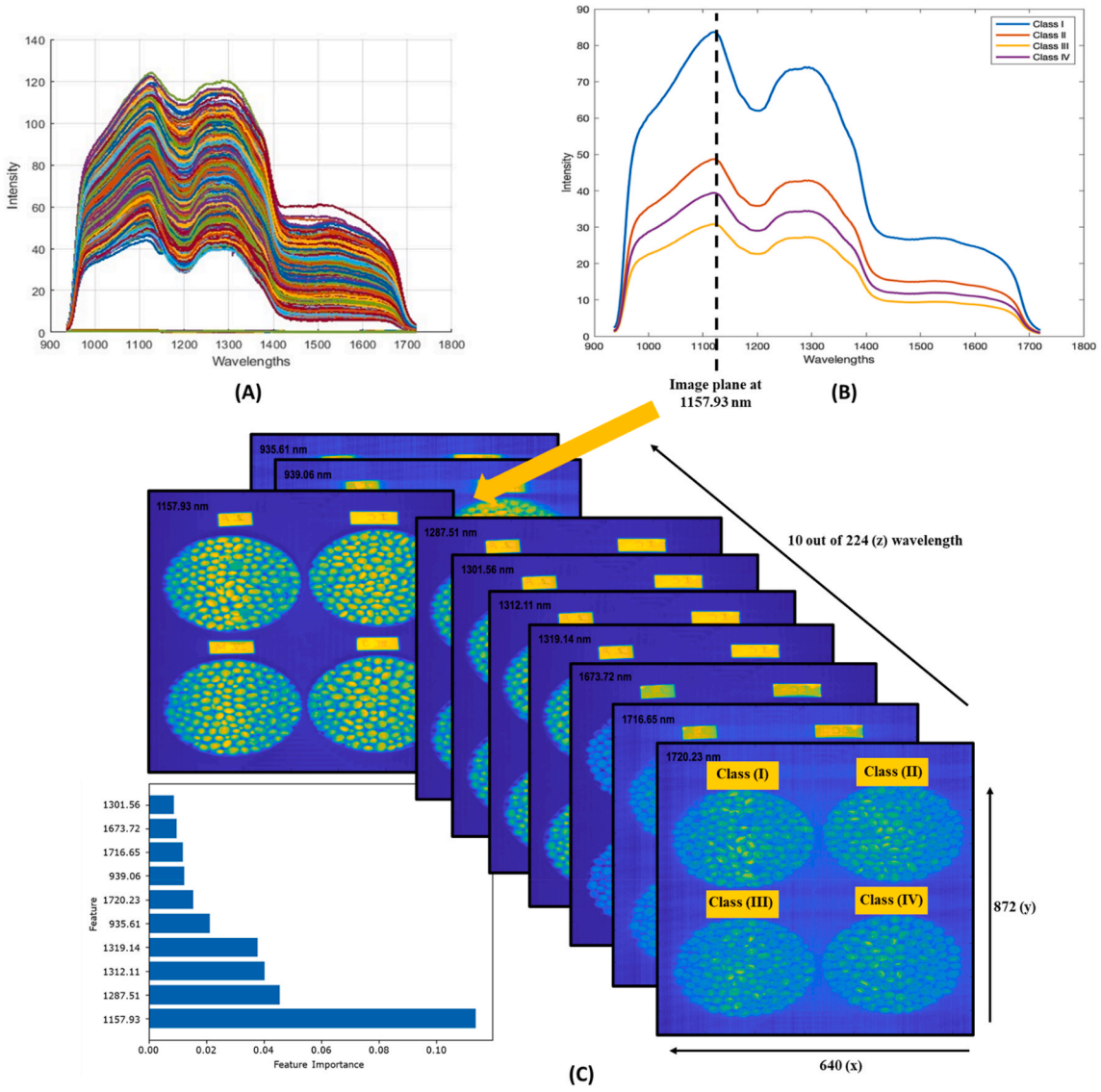


Fig. 2. Hyperspectral imaging (HSI) profile of soybean seeds at the spectral range spanned from 900 to 1700 nm. (A) The HSI wavelength profile of all the soybeans; (B) The HSI wavelength profile of classified soybeans; (C) Images of soybeans at ten featured wavelengths. The 10 featured wavelengths represented by the image planes were acquired by XGBoost with the feature importance listed.

Table 2
Performance of featured wavelength selected by different models.

Model	Calibration accuracy	r_C	R_C	RMSEC	Prediction accuracy	r_P	R_P	RMSEP
XGBoost	0.997	0.998	0.995	0.077	0.995	0.855	0.763	0.544
RF	1.000	1.000	1.000	0.000	0.564	−0.090	−0.090	1.163
SVM	0.538	0.135	0.135	1.037	0.534	0.122	0.122	1.049

Abbreviations: Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF), correlation coefficients of calibration (r_c), correlation coefficients of prediction (r_p), coefficients of determination of calibration (R_c), coefficients of determination of prediction (R_p), root mean square error of calibration (*RMSEC*), and root mean square error of prediction (*RMSEP*).

similar study by Pal et al. (2021) that also reported the superior performance of the XGBoost algorithm for feature selection in their datasets, while the RF-based approach caused a drop in classification accuracy. Upon delving deeper into the XGBoost parameters, we found that the correlation coefficients of prediction (r_p) and coefficients of determination of prediction (R_p) were 85.5 % and 76.3 % respectively. Consequently, given its remarkable prediction accuracy of 99.5 %, XGBoost was chosen as the optimal algorithm for selecting the featured wavelengths in this study.

In order to enhance model performance by reducing complexity and also offer insights into the most significant spectral bands for predicting tofu quality, wavelengths with importance scores above a defined threshold (as 0.008) are selected (Fig. 2C) in the XGBoost. The important score for each wavelength is calculated based on how much it improves the model's accuracy, typically measured by the decrease in impurity or error in the decision trees. Finally, ten featured wavelengths from XGBoost were 935.62, 939.08, 1157.93, 1287.51, 1301.50, 1312.11, 1319.14, 1673.72, 1716.65, and 1720.23 nm, respectively (Fig. 2C).

It is known that the near-infrared spectra (900–2500 nm) of soybean seeds mainly provide chemical information about the components such as protein, oil, and water with the bands of O–H, N–H, and C–H groups (Ribera-Fonseca, Noferini, Jorquera-Fontena, & Rombolà, 2016; Sun et al., 2020; Teye, Anyidoho, Agbemafle, Sam-Amoah, & Elliott, 2020). The difference in the reflectance is due to the variation in the content and structure of protein, and oil, reflecting different varieties of soybean seeds moreover, it is also due to the physical properties of the light when interacting with matter such as light scattering effects. Soybean seeds contain a lot of proteins, oils, and carbohydrates, but the chemical composition varies largely by the method of cultivation, temperature, sun, and rainfall (Song et al., 2016).

The water content is a critical factor in determining tofu quality, influencing both yield and firmness. Curran (1989) identified 1287.51 nm as a key wavelength associated with water content. Additionally, Barbin et al. (2015) and Santagapita, Tylewicz, Panarese, Rocculi, and Dalla Rosa (2016) noted that the first overtone of water corresponds to approximately 1450 nm. This wavelength is particularly notable in our study: as Fig. 2B illustrates, 1450 nm could differentiate between the four classes of soybean seeds. However, it's important to note that this wavelength does not rank among the top ten in terms of importance, as shown in Fig. 2C. This discrepancy highlights the complexity of the relationship between specific wavelengths and chemical compositions in soybean seeds, and warrants further investigation in our ongoing research.

According to Table 3, there were several compounds observed using the ten selected featured wavelengths like oil (935.62 nm, 939.08 nm), proteins (1157.93 nm, 1673.72 nm, 1716.65 nm, 1720.23 nm), cellulose (1287.51 nm), and lignin (1287.51 nm, 1673.72 nm, 1716.65 nm, 1720.23 nm) (Curran, 1989). Those results covered the major three chemical components, including protein, oil, and carbohydrates, in the featured wavelengths. Another research suggested that absorption at 1187 nm (-CH), 1496 nm (-NH), 1674 nm (-CH), 1743 nm (-CH), 1980 nm (-NH), 2055 nm (-ROH/NH), and 2167 nm (-NH) increased as the protein content increased (Ingle et al., 2016). Therefore, the ten featured wavelengths were good indicators of the protein quality of soybean seeds. Barbin, Sobottka, Risso, Zucareli, and Hirooka (2016) employed preprocessing techniques such as the first derivative, second derivative, Multiplicative Scatter Correction (MSC), and Standard Normal Variate (SNV) in conjunction with Partial Least Squares (PLS) regression models. This approach was effectively used for predicting various maize grain attributes, including protein content, water activity, moisture, and ash content. Adopting similar preprocessing methods could enhance our model's performance in future research endeavors. Specifically, these techniques might aid in elucidating the intricate relationship between wavelengths and chemical compositions, thereby refining our predictive accuracy regarding gypsum tofu quality from soybean seeds.

3.3. Predicting gypsum tofu quality based on HSI with CNN model

3.3.1. Establishment of CNN model

Spectral data is rich in complex features, making it an ideal candidate for analysis using CNN. These models, characterized by their extensive architectures, offer advantages over traditional classifiers by extracting more abstract data features, leading to heightened performance levels. Although the training time for Convolutional Neural Networks (CNN) tends to be longer compared to other models, the trade-off is a superior performance, particularly in image classification tasks, where CNNs are considered one of the most effective algorithms (Zhou, Zhang, Liu, Qiu, & He, 2019).

In this study, CNN was employed to develop a model based on ten selected spectral bands of interest. The essential parameters of this model are outlined in Table S1. A predictive model was established using the developed algorithm, which was subsequently verified by inputting random soybean seed images and evaluating the accuracy of its class predictions.

For this assessment, the confusion matrix (Fig. 3) was performed to evaluate the model performance. The diagonal cells of the matrix represent the number of correct predictions for each class. The values (293, 280, 282, 292) indicate high accuracy across all classes, suggesting that the CNN model has a strong ability to distinguish between the different quality classes of tofu. There are very few off-diagonal elements with non-zero values, indicating that misclassifications are minimal. This is evidenced by the presence of only a single '5′ in the bottom row, indicating that only five instances of class 3 were incorrectly classified as class 2.

The model's specificity, precision, and sensitivity scores of 0.9986, 0.9956, and 0.9958, respectively, indicate outstanding performance in correctly classifying the quality of soybean seeds for tofu production (Fig. 3). The high specificity suggests the model was exceptionally effective at identifying seeds that will not yield high-quality tofu, minimizing potential waste of resources on inferior seeds. The precision indicates that when the model predicts a seed will produce high-quality tofu, it is correct nearly all the time, which is crucial for maintaining the consistency of the tofu's quality. Sensitivity, or the model's ability to correctly identify high-quality seeds, ensures that superior seeds are seldom missed, maximizing the potential yield of quality tofu products. Together, these metrics underscore the robustness and reliability of the model in a practical quality control setting, balancing the need to avoid false positives (poor-quality seeds wrongly classified as high-quality) with the need to correctly identify as many high-quality seeds as possible.

In addition to the confusion matrix, one hundred images were utilized from each class to determine the prediction percentages. As indicated in Table S2, Class I achieved a 98 % prediction rate, Class II had a

Table 3
Featured wavelengths and the corresponding bonds.

Wavelength (nm)	Bond Vibration	Chemicals
935.61	C–H Stretch	Oil
939.06	C–H Stretch	Oil
1157.93	N–H Stretch	Protein
1287.51	O–H Bend, 1st Overtone	Water, Cellulose, Lignin
1673.72	C–H Stretch, 1st Overtone	Protein, Lignin, Nitrogen
1716.65	C–H Stretch, 1st Overtone	Protein, Lignin, Nitrogen
1720.23	C–H Stretch, 1st Overtone	Protein, Lignin, Nitrogen

The data is cited from (Curran, 1989).

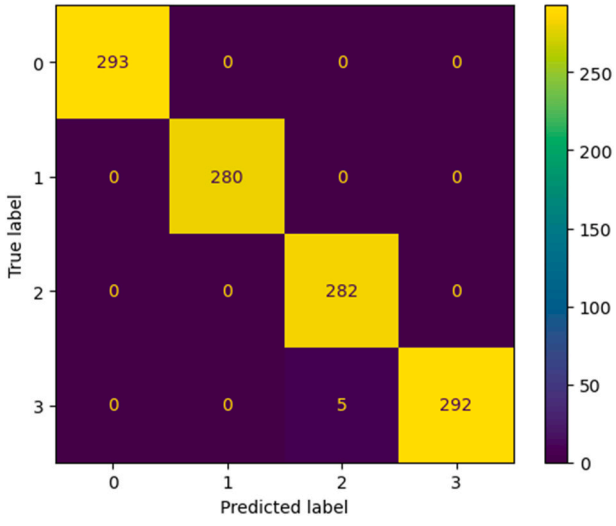


Fig. 3. Confusion matrix of the CNN model, where the numbers 0, 1, 2, and 3 on the axes correspond to Classes I, II, III, and IV, respectively. The model demonstrated high specificity, precision, and sensitivity, with respective values of 0.9986, 0.9956, and 0.9958.

99 % prediction rate, and Class III also had a 99 % prediction rate, while Class IV only attained a 96 % prediction rate. These high prediction rates serve as a testament to the testing accuracy of the model, demonstrating its efficacy and robustness in predicting soybean seed classifications.

While there have been no reported applications of CNN in predicting food processes, there is a growing body of research leveraging CNN for food quality and safety prediction. For instance, Yu, Tang, Wu, and Lu (2018) employed a deep learning model to analyze visible/near-infrared hyperspectral data from shrimps, aiming to predict their freshness. They used a Stacked Autoencoder (SAE) model to extract deep features from the samples, and then applied logistic regression to classify the freshness grade of shrimp based on these features. This novel approach yielded impressive results, with calibration and prediction set accuracies reaching 96.55 % and 93.97 % respectively, demonstrating the potential of deep learning methods in food quality assessment. Similar applications can be found in an illustrative study on the use of CNNs for HSI analysis. Qiu et al. (2018) explored the potential of CNNs to identify rice seed varieties. Significantly, the CNN model outperformed the SVM model in most scenarios, with an impressive total accuracy rate of 89.6 %, showcasing the effectiveness of CNNs in analyzing spectral data. Our research demonstrates the promising application of CNNs in hyperspectral imaging for food product prediction, especially in predicting the tofu quality based on soybean seeds. The findings suggest that the aid of rapid sample collection through CNNs and HSI is a good combination to predict food quality based on the ingredients profile.

3.3.2. *Verification of CNN model*

Four untested soybean seeds were employed to evaluate the quality of soybean seeds scanned with HSI. The resulting images were processed and fed into the CNN model, which classified the seeds into Class I, II, III, and IV. Tofu made from these soybeans was evaluated for quality using **Methods 2.4**, and the results were presented in Fig. 4. In general, the CNN model accurately predicted the quality of soybeans in Class II and Class III, and most parameters for Class I and Class IV were also well predicted. Nonetheless, there are certain limitations to these results. Specifically, each quality parameter of tofu made from Class II soybeans fell within the interquartile range (IQR), while those of Class III tofu were situated between the lower and upper whiskers. These results are considered acceptable because the predicted tofu quality remains within the range of the training dataset.

Although most quality parameters for Class I and Class IV also fell within the whisker range, there were outliers in the predicted results. The springiness of tofu (Fig. 4D) made with predicted Class IV soybeans was lower than the lower whisker, which can be considered an outlier; this may be attributed to the presence of outliers in the training dataset itself. Conversely, the firmness of Class I tofu (Fig. 4C) was higher than the upper whisker, indicating a model prediction outlier. The statistical significance of the outliers in our model's predictions is an important consideration for interpreting its accuracy. Determining whether these outliers are indicative of a systematic error or simply reflect the natural variability in soybean quality is essential for assessing the model's performance. The presence of outliers could potentially highlight areas where the model might benefit from refined calibration, especially if they stem from predictable biases within the training data. By understanding the characteristics of these outliers, we can better gauge the model's reliability and ensure its predictions are robust and consistent with the empirical quality attributes of tofu. In addition to these outliers, some predicted parameters were close to the whiskers, such as the water uptake capacity of Class I and Class III, and the tofu yield of Class IV. These results may be due to the bimodal distribution of the training dataset, as illustrated in Fig. 4A&B (Scheres, 2010; P. Xu et al., 2022).

Classifying soybeans into more categories using the training dataset might help reduce outliers and the bimodal distribution. However, increasing the number of categories may lead to fewer samples per category, potentially causing issues such as overfitting, high variance, or inappropriate model selection (Scheres, 2010). The optimal solution would be to collect additional data to enhance the performance of the machine learning model.

4. Conclusions

This study successfully determined ten featured wavelengths from

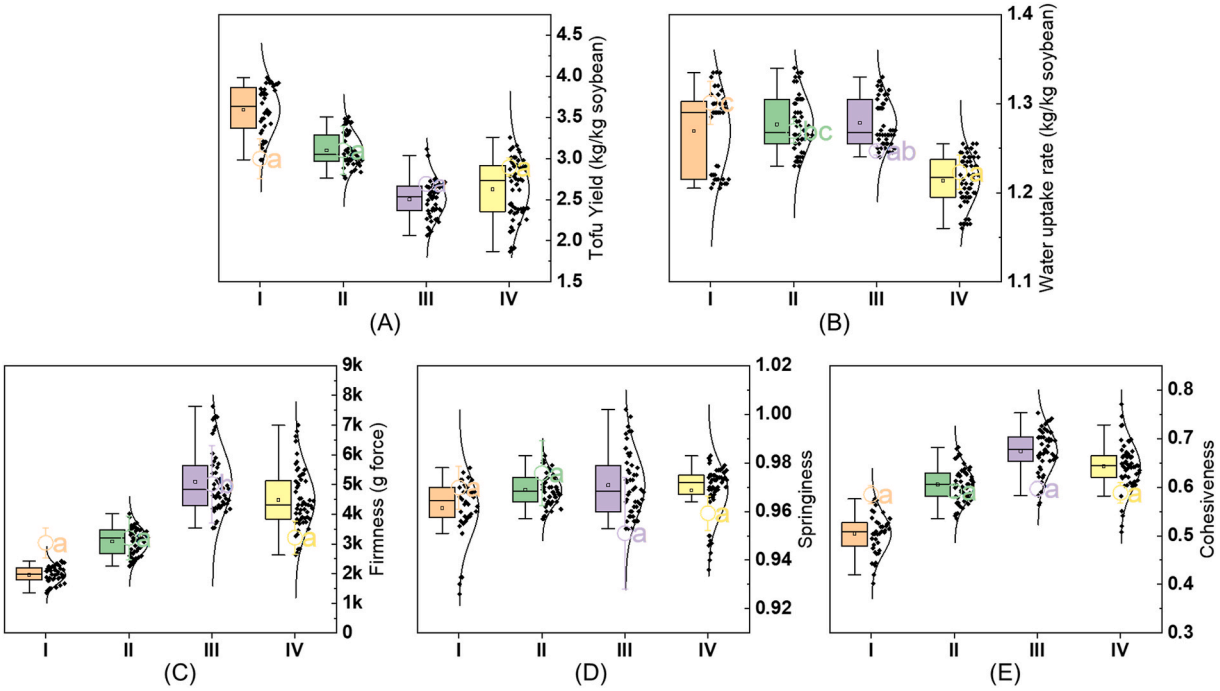


Fig. 4. Verification of soybean seeds with tofu quality (A) Tofu yield, (B) Water uptaking capacity, (C) Firmness, (D) Springiness, and (E) Cohesiveness. The circle symbol indicates the mean value of tested tofu quality. The line of each box from top to bottom indicates upper whisker, upper quartile, median, lower quartile, and lower whisker. The black dots indicate the parameter values in the training dataset. Different letters indicate statistically significant differences ($p < 0.05$).

Hyperspectral Imaging (HSI) data, spanning 200 soybean varieties, with the help of the XGBoost algorithm. These wavelengths potentially correlate with the protein, carbohydrate, and oil contents in the soybean seeds. However, further validation is needed to substantiate the relationship between these wavelengths and the respective chemical compositions.

A CNN model for predicting gypsum tofu quality has been successfully developed based on these ten featured wavelengths from the HSI data. This model, trained on data from 200 soybean varieties, is capable of classifying soybeans into four distinct classes using HSI images of individual seeds. The predictive accuracy for each class of soybeans impressively ranges from 96 % to 99 %.

The robustness of this model was further validated using untested soybean samples. These samples were accurately categorized into distinct classes, each representing a specific range of tofu quality parameters. Upon comparison, it was observed that the model accurately predicted the majority of tofu quality traits.

This research sets the groundwork for understanding the relationship between hyperspectral image, chemical composition, and tofu qualities. The feasibility of predicting tofu quality based on the hyperspectral image has been demonstrated; however, the machine learning prediction model requires further enhancements. It is recommended to collect more soybean samples to classify seeds into additional categories. This would equip the prediction model with a more comprehensive ability to accurately estimate tofu quality based on a diverse set of quality parameters. Our future research will delve deeper into understanding why HSI can predict tofu quality and identify the critical components of soybean seeds for tofu processing.

CRediT authorship contribution statement

Amanda Malik: Writing – original draft, Investigation. **Billy Ram:** Methodology. **Dharanidharan Arumugam:** Methodology. **Zhao Jin:** Conceptualization. **Xin Sun:** Resources. **Minwei Xu:** Writing – review & editing, Supervision, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by the North Dakota Agricultural Products Utilization Commission (FAR0035424) and North Dakota Agricultural Experiment Station (ND 1529). We also thank Northern Crops Institute and Agricultural Utilization Research Institute for their help with soybean sourcing and tofu processing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2024.110357.

References

Ali, F., Tian, K., & Wang, Z. X. (2021). Modern techniques efficacy on tofu processing: A review. *Trends in Food Science & Technology, 116*, 766–785. https://doi.org/10.1016/j.tifs.2021.07.023

Barbin, D. F., Sobottka, R. P., Risso, W. E., Zucareli, C., & Hirooka, E. Y. (2016). Influence of plant densities and fertilization on maize grains by near-infrared spectroscopy. *Spectroscopy Letters, 49*(2), 73–79. https://doi.org/10.1080/00387010.2015.1076005

Barbin, D. F., Valous, N. A., Dias, A. P., Camisa, J., Hirooka, E. Y., & Yamashita, F. (2015). VIS-NIR spectroscopy as a process analytical technology for compositional characterization of film biopolymers and correlation with their mechanical properties. *Materials Science and Engineering: C, 56*, 274–279. https://doi.org/10.1016/j.msec.2015.06.029

Beléia, A., Prudencio-Ferreira, S. H., Yamashita, F., Sakamoto, T. M., & Ito, L. (2004). Sensory and instrumental texture analysis of cassava (*Manihot esculenta*, Crantz) roots. *Journal of Texture Studies, 35*(5), 542–553. https://doi.org/10.1111/j.1745-4603.2004.35505.x

Cai, T., & Chang, K. C. (1999). Processing effect on soybean storage proteins and their relationship with tofu quality. *Journal of Agricultural and Food Chemistry, 47*(2), 720–727. https://doi.org/10.1021/jf980571z

Chen, C. C., Hsieh, J. F., & Kuo, M. I. (2023). Insight into the processing, gelation and functional components of tofu: A review. *Processes, 11*(1), 202. https://doi.org/10.3390/pr11010202

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment, 30*(3), 271–278. https://doi.org/10.1016/0034-4257(89)90069-2

da Silva Medeiros, M. L., Cruz-Tirado, J. P., Lima, A. F., de Souza Netto, J. M., Ribeiro, A. P. B., Bassegio, D., et al. (2022). Assessment oil composition and species discrimination of Brassicas seeds based on hyperspectral imaging and portable near infrared (NIR) spectroscopy tools and chemometrics. *Journal of Food Composition and Analysis, 107*, Article 104403. https://doi.org/10.1016/j.jfca.2022.104403

Erkinbaev, C., Henderson, K., & Paliwal, J. (2017). Discrimination of gluten-free oats from contaminants using near infrared hyperspectral imaging technique. *Food Control, 80*, 197–203. https://doi.org/10.1016/j.foodcont.2017.04.036

Feng, C. H., Makino, Y., Oshita, S., & García Martín, J. F. (2018). Hyperspectral imaging and multispectral imaging as the novel techniques for detecting defects in raw and processed meat products: Current state-of-the-art research advances. *Food Control, 84*, 165–176. https://doi.org/10.1016/j.foodcont.2017.07.013

Fukushima, D. (1991). Recent progress of soybean protein foods: Chemistry, technology, and nutrition. *Food Reviews International, 7*(3), 323–351. https://doi.org/10.1080/87559129109540915

Gao, J., Zhao, L., Li, J., Deng, L., Ni, J., & Han, Z. (2021). Aflatoxin rapid detection based on hyperspectral with 1D-convolution neural network in the pixel level. *Food Chemistry, 360*, Article 129968. https://doi.org/10.1016/j.foodchem.2021.129968

Guan, X., Zhong, X., Lu, Y., Du, X., Jia, R., Li, H., et al. (2021). Changes of soybean protein during tofu processing. *Foods, 10*(7), 1594. https://doi.org/10.3390/foods10071594

He, W., He, H., Wang, F., Wang, S., Li, R., Chang, J., et al. (2022). Rapid and uninvasive characterization of bananas by hyperspectral imaging with extreme gradient boosting (XGBoost). *Analytical Letters, 55*(4), 620–633. https://doi.org/10.1080/00032719.2021.1952214

Huang, L., Liu, Y., Huang, W., Dong, Y., Ma, H., Wu, K., et al. (2022). Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. *Agriculture, 12*(1), 74. https://doi.org/10.3390/agriculture12010074

Huang, L., Zhou, Y., Meng, L., Wu, D., & He, Y. (2017). Comparison of different CCD detectors and chemometrics for predicting total anthocyanin content and antioxidant activity of mulberry fruit using visible and near infrared hyperspectral imaging technique. *Food Chemistry, 224*, 1–10. https://doi.org/10.1016/j.foodchem.2016.12.037

Ingle, P. D., Christian, R., Purohit, P., Zarraga, V., Handley, E., Freel, K., et al. (2016). Determination of protein content by NIR spectroscopy in protein powder mix products. *Journal of AOAC International, 99*(2), 360–363. https://doi.org/10.5740/JAOACINT.15-0115

Iqbal, A., Sun, D. W., & Allen, P. (2014). An overview on principle, techniques and application of hyperspectral imaging with special reference to ham quality evaluation and control. *Food Control, 46*, 242–254. https://doi.org/10.1016/j.foodcont.2014.05.024

James, A. T., & Yang, A. (2014). Influence of globulin subunit composition of soybean proteins on silken tofu quality. 2. Absence of 11SA4 improves the effect of protein content on tofu hardness. *Crop & Pasture Science, 65*(3), 268–273. https://doi.org/10.1071/cp13399

James, A. T., & Yang, A. (2016). Interactions of protein content and globulin subunit composition of soybean proteins in relation to tofu gel properties. *Food Chemistry, 194*, 284–289. https://doi.org/10.1016/j.foodchem.2015.08.021

Kandpal, L. M., Lee, S., Kim, M. S., Bae, H., & Cho, B. K. (2015). Short wave infrared (SWIR) hyperspectral imaging technique for examination of aflatoxin B1 (AFB1) on corn kernels. *Food Control, 51*, 171–176. https://doi.org/10.1016/j.foodcont.2014.11.020

Kucha, C. T., Liu, L., Ngadi, M., & Claude, G. (2021). Hyperspectral imaging and chemometrics as a non-invasive tool to discriminate and analyze iodine value of pork fat. *Food Control, 127*, Article 108145. https://doi.org/10.1016/j.foodcont.2021.108145

Kurasch, A. K., Hahn, V., Miersch, M., Bachteler, K., & Würschum, T. (2018). Analysis of tofu-related traits by a bench-scale tofu production method and their relationship with agronomic traits in European soybean. *Plant Breeding, 137*(3), 271–282. https://doi.org/10.1111/pbr.12581

Liao, X., Cao, N., Li, M., & Kang, X. (2019). Research on short-term load forecasting using XGBoost based on similar days. *Proceedings*, 675–678. https://doi.org/10.1109/icitbs.2019.00167

Lim, B. T., Deman, J. M., Deman, L., & Buzzell, R. I. (1990). Yield and quality of tofu as affected by soybean and soymilk characteristics. Calcium sulfate coagulant. *Journal*

of Food Science, 55(4), 1088–1092. https://doi.org/10.1111/j.1365-2621.1990.tb01605.x

Loggenberg, K., & Poona, N. (2020). A feature selection approach for terrestrial hyperspectral image analysis. *South African Journal of Geology, 9*(2), 302–320. https://doi.org/10.4314/sajg.v9i2.20

Lv, X., Ming, D., Chen, Y. Y., & Wang, M. (2019). Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. *International Journal of Remote Sensing, 40*(2), 506–531. https://doi.org/10.1080/01431161.2018.1513666

Medus, L. D., Saban, M., Francés-Víllora, J. V., Bataller-Mompeán, M., & Rosado-Muñoz, A. (2021). Hyperspectral image classification using CNN: Application to industrial food packaging. *Food Control, 125*, Article 107962. https://doi.org/10.1016/j.foodcont.2021.107962

Meng, S., Chang, S., Gillen, A. M., & Zhang, Y. (2016). Protein and quality analyses of accessions from the USDA soybean germplasm collection for tofu production. *Food Chemistry, 213*, 31–39. https://doi.org/10.1016/j.foodchem.2016.06.046

Mujoo, R., Trinh, D. T., & Ng, P. K. W. (2003). Characterization of storage proteins in different soybean varieties and their relationship to tofu yield and texture. *Food Chemistry, 82*(2), 265–273. https://doi.org/10.1016/S0308-8146(02)00547-2

Pal, M., Charan, T. B., & Poriya, A. (2021). K-nearest neighbour-based feature selection using hyperspectral data. *Remote Sensing Letters, 12*(2), 128–137. https://doi.org/10.1080/2150704x.2020.1864051

Poysa, V., & Woodrow, L. (2002). Stability of soybean seed composition and its effect on soymilk and tofu yield and quality. *Food Research International, 35*(4), 337–345. https://doi.org/10.1016/S0963-9969(01)00125-9

Poysa, V., Woodrow, L., & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International, 39*(3), 309–317. https://doi.org/10.1016/j.foodres.2005.08.003

Qin, Q., Wang, Q.-G., Li, J., & Sam Ge, S. (2013). Linear and nonlinear trading models with gradient boosted random forests and application to singapore stock market. *Journal of Intelligent Learning Systems and Applications, 5*, 1–10. https://doi.org/10.4236/jilsa.2013.51001

Qiu, Z., Chen, J., Zhao, Y., Zhu, S., He, Y., & Zhang, C. (2018). Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences, 8*(2), 212. https://doi.org/10.3390/app8020212

Ribera-Fonseca, A., Noferini, M., Jorquera-Fontena, E., & Rombolà, A. D. (2016). Assessment of technological maturity parameters and anthocyanins in berries of cv. Sangiovese (*Vitis vinifera* L.) by a portable vis/NIR device. *Scientia Horticulturae, 209*, 229–235. https://doi.org/10.1016/j.scienta.2016.06.004

Santagapita, P. R., Tylewicz, U., Panarese, V., Rocculi, P., & Dalla Rosa, M. (2016). Non-destructive assessment of kiwifruit physico-chemical parameters to optimise the osmotic dehydration process: A study on FT-NIR spectroscopy. *Biosystems Engineering, 142*, 101–109. https://doi.org/10.1016/j.biosystemseng.2015.12.011

Scheres, S. H. W. (2010). Classification of structural heterogeneity by maximum-likelihood methods. *Methods in Enzymology, 482*(C), 295–320. https://doi.org/10.1016/s0076-6879(10)82012-9

Song, W., Yang, R., Wu, T., Wu, C., Sun, S., Zhang, S., et al. (2016). Analyzing the effects of climate factors on soybean protein, oil contents, and composition by extensive and high-density sampling in China. *Journal of Agricultural and Food Chemistry, 64*(20), 4121–4130. https://doi.org/10.1021/acs.jafc.6b00008

Squeo, G., De Angelis, D., Summo, C., Pasqualone, A., Caponio, F., & Amigo, J. M. (2022). Assessment of macronutrients and alpha-galactosides of texturized vegetable proteins by near infrared hyperspectral imaging. *Journal of Food Composition and Analysis, 108*, Article 104459. https://doi.org/10.1016/j.jfca.2022.104459

Stanojevic, S. P., Barac, M. B., Pesic, M. B., & Vucelic-Radovic, B. V. (2011). Assessment of soy genotype and processing method on quality of soybean tofu. *Journal of Agricultural and Food Chemistry, 59*(13), 7368–7376. https://doi.org/10.1021/jf2006672

Su, W. H., Yang, C., Dong, Y., Johnson, R., Page, R., Szinyei, T., et al. (2021). Hyperspectral imaging and improved feature variable selection for automated determination of deoxynivalenol in various genetic lines of barley kernels for resistance screening. *Food Chemistry, 343*, Article 128507. https://doi.org/10.1016/j.foodchem.2020.128507

Sun, J., Wang, G., Zhang, H., Xia, L., Zhao, W., Guo, Y., et al. (2020). Detection of fat content in peanut kernels based on chemometrics and hyperspectral imaging technology. *Infrared Physics & Technology, 105*, Article 103226. https://doi.org/10.1016/j.infrared.2020.103226

Teye, E., Anyidoho, E., Agbemafle, R., Sam-Amoah, L. K., & Elliott, C. (2020). Cocoa bean and cocoa bean products quality evaluation by NIR spectroscopy and chemometrics: A review. *Infrared Physics & Technology, 104*, Article 103127. https://doi.org/10.1016/j.infrared.2019.103127

Warner, T. A., & Shank, M. C. (1997). Spatial autocorrelation analysis of hyperspectral imagery for feature selection. *Remote Sensing of Environment, 60*(1), 58–70. https://doi.org/10.1016/S0034-4257(96)00138-1

Xu, M., Jin, Z., Lan, Y., Rao, J., & Chen, B. (2019). HS-SPME-GC-MS/olfactometry combined with chemometrics to assess the impact of germination on flavor attributes of chickpea, lentil, and yellow pea flours. *Food Chemistry, 280*, 83–95. https://doi.org/10.1016/j.foodchem.2018.12.048

Xu, P., Tan, Q., Zhang, Y., Zha, X., Yang, S., & Yang, R. (2022). Research on maize seed classification and recognition based on machine vision and deep learning. *Agriculture, 12*(2), 232. https://doi.org/10.3390/agriculture12020232

Yang, L., Gao, H., Meng, L., Fu, X., Du, X., Wu, D., et al. (2021). Nondestructive measurement of pectin polysaccharides using hyperspectral imaging in mulberry fruit. *Food Chemistry, 334*, Article 127614. https://doi.org/10.1016/j.foodchem.2020.127614

Yu, X., Tang, L., Wu, X., & Lu, H. (2018). Nondestructive freshness discriminating of shrimp using visible/near-infrared hyperspectral imaging technique and deep learning algorithm. *Food Analytical Methods, 11*(3), 768–780. https://doi.org/10.1007/s12161-017-1050-8

Zhou, L., Zhang, C., Liu, F., Qiu, Z., & He, Y. (2019). Application of deep learning in food: A review. *Comprehensive Reviews in Food Science and Food Safety, 18*(6), 1793–1811. https://doi.org/10.1111/1541-4337.12492