Non-Destructive Moisture Content Prediction Model for Corn Starch Based on Near-Infrared Spectroscopy and Chemometrics

Stella Maria Dyah Cahyarani[1], Dhevika Aji Nugraha[2], Reza Adhitama Putra Hernanda[1], Hoonsoo Lee[1], Hanim Zuhrotul Amanah[2*]

[1]Department of Biosystems Engineering, Chungbuk National University, Cheongju, Republic of Korea.
[2]Department of Agricultural and Biosystems Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia.
Email*): hanim_za@ugm.ac.id

Received:
26 December 2025

Revised:
7 March 2026

Accepted:
9 March 2026

Published:
29 March 2026

DOI:
10.29303/jrpb.v14i1.1225

ISSN 2301-8119, e-ISSN 2443-1354

Available at:
http://jrpb.unram.ac.id/

Abstract: Moisture content is a critical quality attribute of corn starch that affects shelf life, functional performance, and commercial value. This study developed and externally validated a rapid and non-destructive method to quantify corn starch moisture using near-infrared (NIR) spectroscopy and chemometric/machine-learning regression. Commercial corn starch was conditioned at approximately 76% relative humidity (saturated NaCl) for 20 days to generate moisture variability, and spectra were acquired using a SpectraStar XT-R instrument (900-2200 nm). Three spectral pre-processing strategies (MSC, SNV, and Savitzky-Golay first derivative) were evaluated prior to model development. A total of 951 samples were split by stratified sampling into calibration (70%, n = 666) and independent prediction (30%, n = 285) sets. Three models were compared: partial least squares regression (PLSR), support vector regression optimized by particle swarm optimization (SVR-PSO), and a one-dimensional convolutional neural network (1D-CNN). The best performance was achieved by PLSR with SNV (R^2_p = 0.929, $RMSE_p$ = 0.274%, RPD = 3.755), while SVR-PSO with MSC showed comparable accuracy (R^2_p = 0.929, $RMSE_p$ = 0.273%, RPD = 3.762). The 1D-CNN yielded lower predictive performance (best R^2_p = 0.841). Overall, NIR spectroscopy combined with optimized pre-processing and conventional regression models provides an accurate alternative to gravimetric drying for quality control of corn starch.

Keywords: corn starch; moisture content; near-infrared spectroscopy; machine learning; deep learning

INTRODUCTION

Corn (Zea mays L.) is a high-starch global cereal crop (Li et al., 2024; Liu et al., 2020), and is widely used as a key energy ingredient in livestock feed (Burns et al., 2021). Additionally, corn is widely processed into various food products, such as starch, sweeteners, and corn oil (Yu & Moon, 2021). Its derivative, corn starch is one of the most economically significant biopolymers and versatile materials in modern industry. In the food sector, corn starch is commonly used as a thickening and gelling agent, as well as for bulking and retaining water (Ai & Jane, 2016). For corn starch to function effectively in these applications, its physicochemical condition must be precise. This performance does not just happen, it is highly dependent on its composition, processing history, and its interaction with water (Biliaderis, 2009; S. Wang et al., 2015).

Among all quality parameters, moisture content can be said to be the most critical determinant of the quality, functionality, and commercial value of corn starch (Li et al., 2024). From an economic point of view, starch is a commodity sold by weight. Excess water is considered a significant financial loss, as buyers effectively pay for the water rather than the biopolymer/dry matter (Brouk, 2008). Therefore, strict moisture content control is crucial for quality and safety assurance. Starch with a moisture content above the critical safety threshold (around 13-14%) becomes a fertile medium for fungal and bacterial growth (Ozbekova & Kulmyrzaev, 2019). This leads to spoilage, the potential formation of harmful mycotoxins (Abdullah et al., 2000), and a drastically reduced shelf life. Therefore, rapid, precise, and reliable determination of moisture content is necessary in processing for quality control at every stage, from raw material reception to final product delivery (Chen et al., 2017; J. Zhang et al., 2023).

Despite its importance, the measurement of moisture content in corn starch still relies on traditional laboratory-based methods. A commonly used method is thermogravimetry, which involves heating a precisely weighed sample in an oven for several hours until a constant weight is achieved (Horwitz & AOAC International, 2006). Although this method is considered as reference standard for moisture content measurement, it is very time-consuming (Chen et al., 2017) and making it unsuitable for real-time process adjustments. As an alternative to these slow and destructive methods, rapid and non-destructive analytical techniques have become increasingly essential for modern quality control (Li et al., 2024). For example, spectroscopy methods offer the ability to analyze moisture content instantly without needing to destroy or prepare the sample first.

Near-infrared (NIR) spectroscopy, a form of vibrational spectroscopy, has emerged as a superior tool for non-destructive, rapid, and without the need for sample preparation (Manley, 2014; Padhi et al., 2024). NIR allows for on-site and real-time measurements (Chen et al., 2017). Its utility for moisture analysis is based on fundamental physics, since the NIR range (approximately 780–2500 nm) is dominated by overtone and combination absorptions related to fundamental molecular vibrations (Aenugu et al., 2011; Padhi et al., 2024). The O-H bonds in water exhibit strong and distinct absorption bands, for example, the combination band around 1940 nm and the first overtone around 1450 nm (Büning-Pfaue, 2003; Nicolaï et al., 2007) make NIR spectroscopy highly sensitive to even small variations in moisture content.

However, acquiring the NIR spectrum is only the first step; to obtain the desired information, extraction from the spectral data must be performed. The resulting spectral data are inherently complex. NIR bands are characterized by being broad, weak, and highly overlapping, rather than sharp and distinct peaks. Furthermore, the desired chemical signal (from O-H bonds) is often convoluted with strong, undesired physical effects (Ducanchez et al., 2022). These physical phenomena can obscure the chemical information and may severely degrade the model's predictive power if not handled properly. Therefore, effective spectral preprocessing is required in the chemometric workflow. This process involves applying

mathematical transformations to the raw spectra to remove non-chemical variability, increase the signal to noise ratio, and linearize the relationship between the spectra and desired property (Rinnan et al., 2009). Common preprocessing methods include scatter correction techniques, such as standard normal variate (SNV) or multiplicative scatter correction (MSC) to correct for particle size effects, and derivative transformations, such as Savitzky-Golay to separate overlapping peaks and remove baseline shifts (Jiao et al., 2020; Pizarro et al., 2004; Yan, 2025).

After pre-processing, an accurate calibration model must be built to correlate the complex spectral data with the reference moisture content data. In this study, three powerful and commonly used regression methods are used: PLSR, SVR, and 1D-CNN. PLSR is one of the most widely applied methods for quantitative analysis in NIR spectroscopy (Bai et al., 2022). It is a bilinear modelling approach that handles collinearity effectively by reducing high dimensional spectral data into a small set of orthogonal latent variables (LVs) (Sawatsky et al., 2015). These LVs are constructed to maximize the covariance between the spectral data and the reference values. In contrast, SVR is a nonlinear machine-learning method grounded in statistical learning theory (Awad & Khanna, 2015). By applying kernel-based transformations, SVR can represent complex nonlinear relationships between spectral responses and moisture content that may not be captured adequately by PLSR (Y. Wang et al., 2015). Recently, deep learning approaches such as convolutional neural networks (CNNs) have gained increasing attention in spectroscopic analysis due to their ability to capture nonlinear spectral relationships and perform automated feature extraction. However, their practical feasibility under limited sample conditions in starch-based matrices remains insufficiently explored. Finally, CNN represents a deep learning approach that uses convolutional layers to automatically extract local spectral features and hierarchical patterns, offering an alternative for modelling highly nonlinear data without heavy reliance on manual feature engineering (LeCun et al., 2015). The success of these models is highly dependent on the optimization of the preprocessing steps (Walsh et al., 2024).

However, the effective application of these modelling approaches requires validation on specific sample matrices, as spectral responses are highly sensitive to both chemical composition and physical characteristics. In this context, while NIR spectroscopy has been applied to corn processing products, important gaps remain for purified corn starch. Previous studies by Chen et al. (2017), have primarily focused on moisture prediction in corn processing products, particularly in corn flour, which represents a substantially different matrix containing proteins and lipids compared with the purified carbohydrate structure of corn starch. Therefore, calibration models developed for the complex matrix of corn flour are not directly applicable to corn starch due to significant spectral differences. Furthermore, few studies have benchmarked modern deep learning (1D-CNN) against traditional chemometrics for this specific high-purity matrix.

Therefore, this study aims to develop and rigorously validate a rapid, non-destructive NIR-based method for moisture quantification in corn starch by integrating spectral preprocessing (including scatter correction, derivative transformation, and smoothing) with three calibration models (PLSR, SVR, and 1D-CNN). By systematically benchmarking multiple preprocessing–model combinations against reference moisture values, this work seeks to identify the optimal chemometric strategy for this matrix. The overall objective is to establish a robust analytical protocol suitable for quality control and real-time process monitoring, ensuring the quality, safety, and functional consistency of corn starch in demanding industrial environments.

MATERIALS AND METHODS

Sample Preparation

This study used commercially available corn starch under the brand name "Maizenaku". To obtain a range of moisture content variations, we artificially created conditions by placing 5 grams of corn starch in a 60 mm aluminium dish (3 plates per sample) within a closed polypropylene chamber (530 × 380 × 200) mm containing a solution of saturated sodium chloride (NaCl) (technical grade, CV Chem-Mix Pratama, Indonesia). The saline solution controlled the surrounding conditions at about 76% relative humidity, based on the saturated NaCl chamber method described by Amanah et al. (2024). Measurements were taken over 20 days of storage, from day 0 through day 20. We collected data at 3-hour intervals on day 1, at 6-hour intervals on days 2 and 3, and at 12-hour intervals from days 4 to 7, and then at 24-hour intervals thereafter until data collection was completed. In total we have 108 samples.

NIR Spectroscopy

Spectral data collection was carried out using NIR spectroscopy (SpectraStar™ XT-R, Westborough, USA, 680–2600 nm) equipped with a tungsten halogen lamp with an MTBF rating of 10,000 hours. The instrument was set to a wavelength resolution of 1 nm, continuous cup rotation at a speed of 20 deg/sec, and 3 scans per sample. Before spectral acquisition, the instrument was allowed to warm up for 4–5 hours. Instrument performance was verified by authorized facility technicians using certified standards (TAS wavelength certified standard, US-SIDS-0005; and TAS photometric certified standard, US-SIDS-0004) and evaluated via the UScan TAS Performance Test. Spectral measurements were initiated only after the performance checks met the manufacturer's acceptance criteria.

After confirming acceptable instrument performance, the starch sample was placed on the sample holder until it filled the spectroscopic plate. NIR spectroscopy data were collected using rotation types of data collection methods based on the movement of the sample holder. During data collection, 10 repetitions were carried out for each sample, therefore, we have 108 x 10 = 1080 spectra.

Moisture content (MC) determination

MC was measured by the thermogravimetric method following the National Standard of the People's Republic of China (GB 5009.3-2016). Three grams of corn starch were weighed on an analytical balance (Mettler Toledo ML204T; readability 0.1 mg) and dried in a hot-air oven at 105 °C for 3 hours to constant mass. MC was calculated from mass loss, the difference between the sample mass before and after drying, which was interpreted as the amount of water removed from the sample.

Outlier Removal

Before preprocessing, the raw spectral dataset was screened for outliers using MATLAB [MathWorks, Natick, MA, USA; R2024a]. This screening process adopted the outlier detection concept described by Bjerrum et al. (2017) to ensure data integrity prior to main modeling.

Specifically, spectral outliers were identified using a preliminary PLS model constructed with the SIMPLS algorithm (standard in MATLAB). The dataset underwent 10-fold cross-validation (CV), varying the number of principal components (PCs) to find the optimal complexity based on the lowest Mean Squared Error (MSE) and a cumulative explained variance threshold. Using this preliminary model, outliers were mathematically defined as samples exhibiting an absolute prediction error (APE) greater than 2.5 times the standard deviation of the prediction errors (APE > 2.5 SD).

This automated filtering process effectively removed abnormal spectra from the initial dataset, resulting in a final, high-quality dataset of 951 spectra. The resulting cleaned dataset was then utilized as the input for the subsequent spectral preprocessing analysis and calibration modeling.

Data Separation

The data resulting from the outlier removal processes were split into a calibration set (70% of the total data) and a prediction set (30% of the total data). To ensure a highly representative distribution of the target variable across both subsets, a stratified data splitting technique based on the reference values was employed. Specifically, the spectral data were grouped according to their unique moisture content values, and within each distinct group, the spectra were systematically divided at the 70:30 ratio. This stratification guarantees that both sets possess an identical variation range and a proportional distribution of moisture content. The calibration set was utilized to construct the regression models, whereas the prediction set was used for external validation to assess the generalization capability of the developed models.

Spectral preprocessing

Preprocessing is a mathematical analysis aimed at improving the quality and accuracy of spectral data to make it easier to process by a system. The preprocessing component is important to perform to create a stable, reliable, and accurate spectral model. Several preprocessing methods used in this research include multiplicative scatter correction (MSC), standard normal variate (SNV), and Savitzky-Golay first derivative (SG1). Detailed equations and comprehensive information can be found in (Rinnan et al., 2009) and (Kusumaningrum et al., 2018).

PLSR model development

This study used PLSR as the model development. PLSR is a multivariate statistical method that integrates concepts from principal component analysis and multiple linear regression (B. Nie et al., 2023). Commonly, PLSR used to spectral data for extracting LVs that capture the covariance between spectral data (x) and the response variable (y). In this study, PLSR was applied to develop a prediction model for moisture content using the spectral dataset. So, the PLSR formulation is expressed in Equation 1-4.

$$X = TP^T + E \quad (1)$$

$$Y = UQ^T + E \quad (2)$$

$$U = \beta \times T \quad (3)$$

$$Y_i = X_i\beta + \beta_0 \quad (4)$$

In these equations, X and Y denote the spectral matrix and the reference data. **T** and **U** are the score matrices, whereas **P** and **Q** are the loading vectors for **X** and **Y.** E represent residual matrices with the same dimensions as X and Y. β and β_0 are beta coefficient, a matrix containing the PLS coefficient.

SVR – PSO model development

The second model development is SVR, a machine learning method optimized using particle swarm optimization (PSO). SVR is a regression variant of the support vector machine (SVM) framework. Unlike standard linear regression which attempts to minimize error on all points, SVR is based on the structural risk minimization principle (Awad & Khanna, 2015). It is highly effective at modelling complex non-linear relationships between spectral data and moisture content using kernel functions. The optimal performance of SVR relies heavily on

the precise tuning of its hyperparameters, specifically the regularization parameter (C), the kernel parameter (γ or σ2) and the ∈-insensitive parameter (∈). Suboptimal choices of these parameters can degrade predictive performance and generalization, which is why many studies explicitly search for hyperparameter settings that improve SVR generalization capability. PSO is a population-based optimization method (Poli et al., 2007). In hybrid PSO–SVR modelling, PSO is used to seek the optimal SVR hyperparameters automatically to improve the generalization capability of the resulting SVR model. By automatically tuning these values, the hybrid PSO-SVR model aims to locate the unique and optimal solution and achieve superior generalization accuracy compared to conventional approaches (Safarzadegan Gilan et al., 2012).

Deep Learning Approach: 1D-CNN

The deep-learning framework utilized in this work was a 1D-CNN, adapted from the architecture proposed by Zhou et al. (2022). Because the spectral data had already been standardized during the chemometric preprocessing stage, no additional normalization layers were required before feeding the data into the CNN.

As illustrated in Figure 1, the architecture was designed specifically for NIR spectral regression. It consisted of a single 1D convolutional layer (with 32 filters, a kernel size of 5, a stride of 1, and 'same' padding) to extract features from the spectra. The weights for this layer were initialized using the 'HeNormal' method. A max-pooling layer (pool size = 2) was then added behind the convolution layer to reduce dimensionality and retain dominant features. The Exponential Linear Unit (ELU) was applied to introduce non-linearity. For the prediction stage, a multi-layer perceptron (MLP) comprising three dense layers was used. ELU also served as the activation function for these hidden layers to prevent the vanishing gradient problem. The dense layers contained 64, 32, and 4 neurons, respectively. The final output layer utilized a linear activation function for the regression task.

The model was optimized using the Adam optimizer, and the loss function was defined as mean squared error (MSE). Training was executed with a small batch size of 8 for a maximum of 500 epochs. The initial learning rate was set to 3.125×10^{-4}, determined by applying a linear scaling rule (scaled as 0.01 x batch size/256) to maintain training stability. Furthermore, to avoid overfitting, an early stopping technique was applied if there was no significant improvement in the validation loss by at least 10^{-4} over 50 epochs, alongside a dynamic learning rate reduction. This 1D-CNN was implemented in Visual Studio Code using Python language.

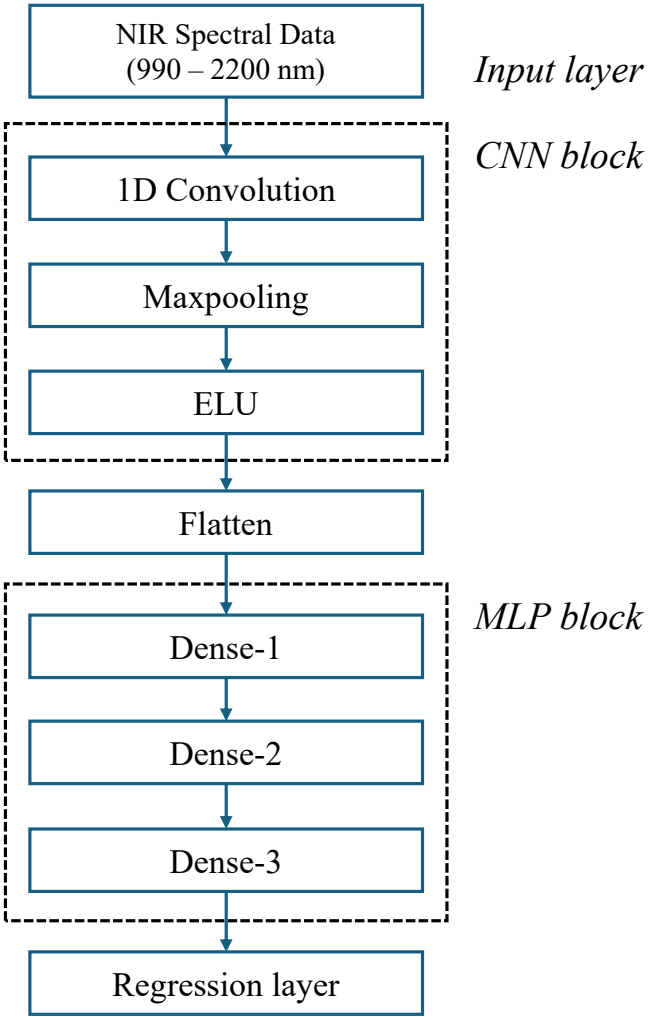

Figure 1. Deep learning model

Statistical analysis

The performance of the PLSR, SVR, and 1D-CNN models was evaluated by comparing the predicted moisture content against the reference values. The accuracy and robustness of the models were assessed using the coefficient of determination (R^2), the root mean square error (RMSE), and the ratio of prediction to deviation (RPD). A good model is characterized by high R^2 and RPD values coupled with a low RMSE. The mathematical formula of these parameters can be calculated in equation 5-6.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{i,p} - y_{i,a}\right)^2}{\sum_{i=1}^{n}\left(\bar{y} - y_{i,p}\right)^2} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{i,p} - y_{i,a}\right)^2}{n}} \quad (6)$$

Software

All analyses were implemented using MATLAB (MathWorks, Natick, MA, USA; R2024a) and Python (version 3.13.5) scripted in Visual Studio Code (version 1.107.0; Microsoft Corporation, Redmond, WA, USA) software. The overall workflow of the study is presented in Figure 2.

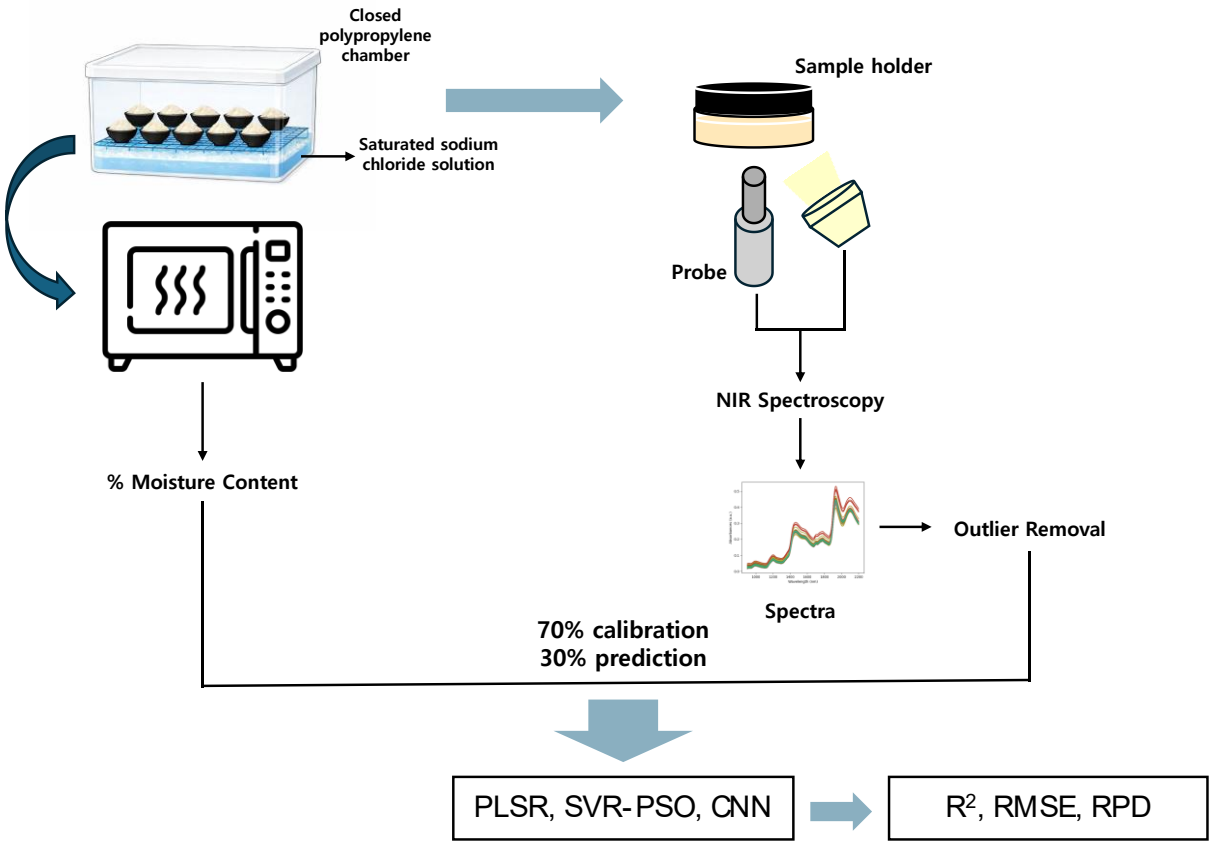

Figure 2. Research workflow

RESULT AND DISCUSSION

Statistics of Moisture Content

The descriptive statistics of the moisture content for the corn starch samples are presented in Table 1. The initial spectra consist of 1,080 data obtained from 36 samples (measured in triplicate with 10 scans each). Following the outlier removal process, 129 spectra were excluded to enhance model stability. After that, the final modeling was conducted using a clean dataset of 951 spectra, ensuring that the calibration models (PLSR, SVR, and 1D-CNN) were trained on robust and representative spectral data. Using stratified sampling, the dataset was split into a calibration set (70%, n = 666) for model training and a prediction set (30%, n = 285) for external validation. The close similarity in the mean, SD, and CV values between the two sets confirms that the prediction set is highly representative of the calibration set (Westad & Marini, 2015). The distribution of the moisture content across both sets is further illustrated in the histogram presented in Figure 3. As shown in Table 1, the moisture content ranged from a minimum of 10.380% to a maximum of 14.881%, with a mean value of 13.399%. This range provides sufficient variability for the development of a reliable NIR calibration model, effectively covering both the standard commercial specifications (typically 10–12%) and the critical levels (>13%) where spoilage risks increase (Whistler & BeMiller, 2009).

Table 1. Descriptive statistics of reference moisture content

Dataset	n	Min (%)	Max (%)	Mean (%)	SD (%)	CV (%)
Total	951	10.380	14.881	13.399	1.027	7.666
Calibration	666	10.380	14.881	13.393	1.027	7.670
Prediction	285	10.380	14.881	13.413	1.027	7.658

Note: n – number of samples; SD – standard deviation; CV – coefficient of variance

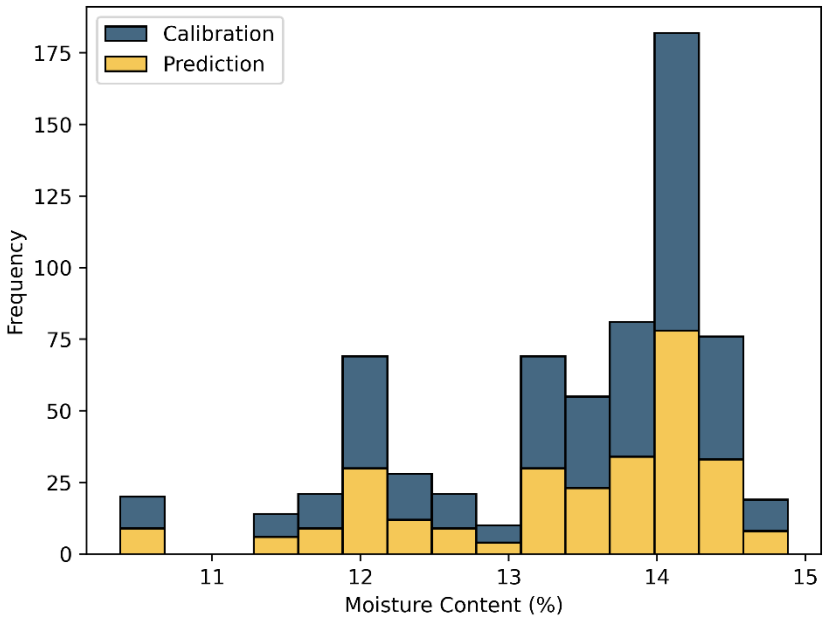

Figure 3. Histogram plot of moisture content on calibration and prediction dataset

Spectral Features

NIR spectral profiles of the corn starch samples are presented in Figure 4, which provides a visualization of raw spectral data. This figure displays the variation in electromagnetic radiation absorption by the sample molecules across the wavelength range of 990 – 2200 nm. Distinct peaks are observed at specific wavelengths, signifying the response of chemical bonds contained within the sample (Amanah et al., 2024).

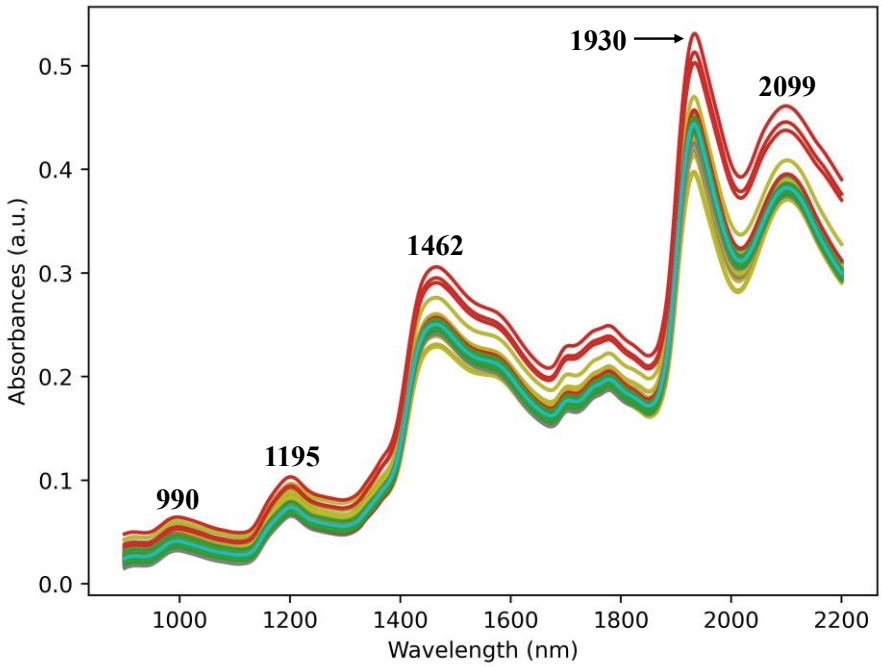

Figure 4. Spectral profiles of the corn starch sample

In this study, prominent absorption peaks were identified at 990, 1195, 1462, 1930, and 2099 nm. The absorption band around 990 nm is attributed to the second overtone of O-H

stretching, contributing to the moisture signal (Workman & Weyer, 2007). The peak at 1195 nm corresponds to the second overtone of C-H stretching vibrations, which is characteristic of the carbohydrate structure in starch (Aenugu et al., 2011). Furthermore, based on the spectral assignment guidelines by Brülls et al. (2007) the band near 1462 nm is associated with the first overtone of O-H stretching vibrations. This feature is characteristic of water molecules hydrogen-bonded to the starch matrix, consistent with the moisture content variation in the samples. The most significant feature related to moisture content is observed at 1930 nm, which is definitively assigned to the combination of O-H stretching and bending vibrations (Weyer & Lo, 2001). Finally, the peak at 2099 nm is indicative of the combination of O-H bending and C-O stretching, reflecting the interaction between water molecules and the starch granules (Manley, 2014).

Overall, the spectral response is dominated by O-H molecular vibrations, confirming the presence of water molecules, with additional contributions from carbohydrate-related bands (for example C-H and C-O combination/overtone features) inherent to the starch matrix. The predominance of these O–H features provide clear evidence of moisture-related spectral variation, which aligns directly with the target property and experimental treatments applied in this study.

Model Performance Comparison

To assess the effectiveness of the proposed methods, the performance of PLSR, SVR-PSO, and 1D-CNN models was compared across different preprocessing techniques. The quantitative results, including R^2_c, R^2_p, $RMSE_c$, $RMSE_p$, and RPD, are comprehensively summarized in Table 2.

Table 2. Performance of regression model for the prediction of moisture content in corn starch

Model	Preprocessing	R^2_c	$RMSE_c$ (%)	R^2_p	$RMSE_p$ (%)	RPD
PLSR	Raw	0.915	0.298	0.917	0.296	3.474
	MSC	0.928	0.275	0.929	0.275	3.744
	SNV	**0.928**	**0.274**	**0.929**	**0.274**	**3.755**
	SG1	0.929	0.271	0.904	0.318	3.230
SVR-PSO	Raw	0.809	0.446	0.810	0.448	2.371
	MSC	**0.929**	**0.272**	**0.929**	**0.273**	**3.762**
	SNV	0.919	0.290	0.920	0.291	3.626
	SG1	0.723	0.537	0.724	0.540	1.535
1D-CNN	Raw	0.617	0.631	0.622	0.632	1.627
	MSC	0.599	0.646	0.602	0.648	1.585
	SNV	0.832	0.418	0.833	0.420	2.447
	SG1	**0.843**	**0.405**	**0.841**	**0.409**	**2.508**

Note: R^2_c and R^2_p denote the coefficients of determination for calibration and prediction; $RMSE_c$ and $RMSE_p$ refer to the root mean squared errors for calibration and prediction; RPD indicates the ratio of prediction to deviation.

As presented in Table 2, the PLSR model demonstrated strong predictive capabilities across all preprocessing methods. Even using raw spectra, the PLSR model achieved high accuracy with an R^2_p of 0.917 and an RPD of 3.474. This aligns with previous findings that PLSR remains the industry standard for modelling linear spectral responses in agricultural products (Nicolaï et al., 2007). The application of spectral preprocessing further improved the model's performance. Specifically, the PLSR model coupled with SNV achieved the highest accuracy among linear models, with an R^2_p of 0.929 and the lowest prediction error ($RMSE_p$ = 0.274%). This improvement confirms that SNV is effective in removing multiplicative interferences caused by particle size variations in powdered samples (Rinnan et al., 2009). According to commonly used NIR calibration interpretation guidelines, models with an RPD

value greater than 3 are considered to provide excellent quantitative information, making them suitable for quantitative applications such as quality control (Z. Nie et al., 2009).

To further interpret why different preprocessing methods affected model performance, the spectral profiles after preprocessing were visually inspected. Figure 5 compares the mean spectra processed by SNV, MSC, and SG1, highlighting that scatter correction mainly stabilizes baseline and multiplicative effects, whereas derivative preprocessing emphasizes local spectral gradients and may amplify high-frequency noise. Therefore, in this study, both the PLSR-SNV and SVR-MSC models achieved RPD values of 3.755 and 3.762, respectively. These results indicate that both models provide robust quantitative performance suitable for practical quality-control applications.

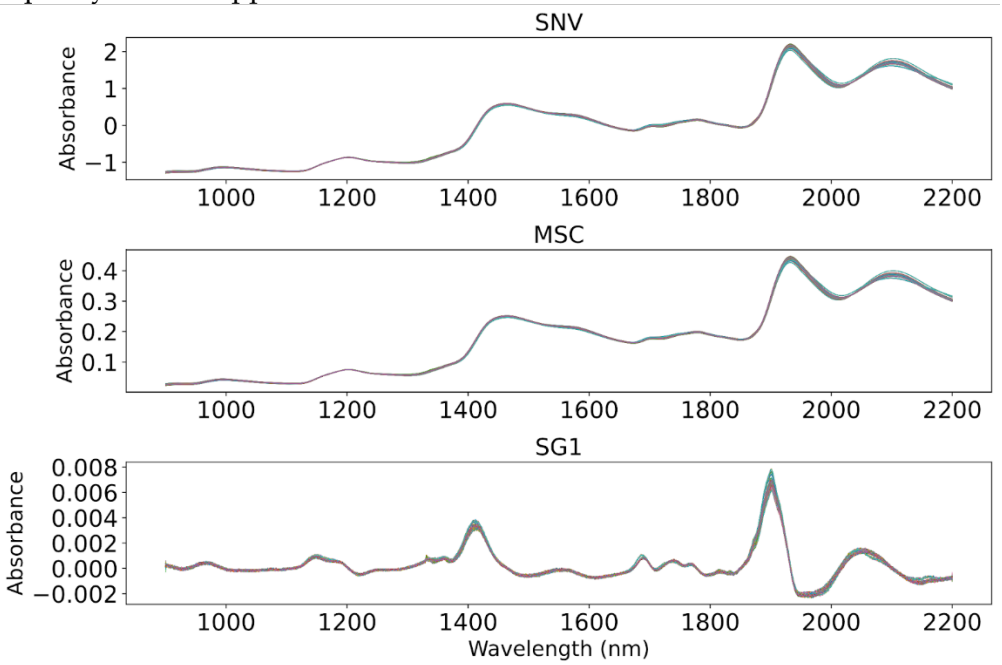

Figure 5. Showing the spectral with different preprocessing options.

Similarly, for the non-linear approach, the SVR model performance was highly dependent on preprocessing. While the SVR model built on raw spectra resulted in a lower accuracy (R^2_p = 0.809, $RMSE_p$ = 0.448%), the application of MSC improved the performance drastically. This model achieved R^2_p of 0.929 and $RMSE_p$ of 0.273%, matching the best PLSR result. In contrast, the SG1 derivative method resulted in a significant drop in performance (R^2_p = 0.724), suggesting that derivative transformation might have amplified the noise in the SVR feature space (Davies & Fearn, 2006).

In contrast, the 1D-CNN approach summarized in the bottom section of Table 2 shows a different trend. The best performance for 1D-CNN model was achieved using SG1 preprocessing, with an R^2_p of 0.841 and $RMSE_p$ of 0.409%. This performance was notably lower compared to the optimal PLSR and SVR models. This finding aligns with common challenges encountered in chemometric applications of deep learning reported in Ng et al., (2020). While 1D-CNN is highly effective in processing raw signal data and automatically extracting complex features, it typically requires a substantially larger dataset (often thousands of samples) to fully train the parameters and perform well compared to shallow models like PLSR. Since the spectral data for this study was limited, the deep architecture may not have reached its full generalization potential, making it less accurate than the simpler methods like PLSR and SVR (Mishra et al., 2022).

Comparing the three approaches, PLSR with SNV preprocessing and SVR with MSC preprocessing emerged as the optimal methods, both achieving an R^2_p of 0.929 and $RMSE_p$ of 0.27%. These findings suggest that the association between NIR spectral responses and corn starch moisture content is primarily linear, which is effectively addressed by established chemometric techniques. Although the 1D-CNN model showed potential, its lower performance suggests that applying deep learning models is not suitable for small sized spectral datasets that are common in laboratory research. Overall, the high accuracy achieved by the PLSR and SVR models confirms that NIR spectroscopy is a feasible and reliable technique for the rapid and non-destructive determination of moisture content in corn starch.

Important wavelengths identified by PLSR-VIP

To support spectral interpretation using our dataset, variable importance in projection (VIP) scores were calculated from the optimal PLSR model (same preprocessing setting as in Table 2). Wavelengths with VIP > 1 were considered influential for moisture prediction (Figure 6a). The VIP profile highlights a strong contribution in the ~1450 nm region, which is consistent with the first overtone of O–H stretching of water (Y. Zhang & Guo, 2020), and the most dominant contribution in the ~1900–2000 nm region, which corresponds to O–H combination bands associated with moisture in starch matrices. Additional influential regions were also observed toward the longer-wavelength range (>2100 nm), which may reflect combination/overtone features related to water–starch interactions. Overall, the VIP-based interpretation indicates that moisture prediction in corn starch is primarily governed by O–H-related absorptions.

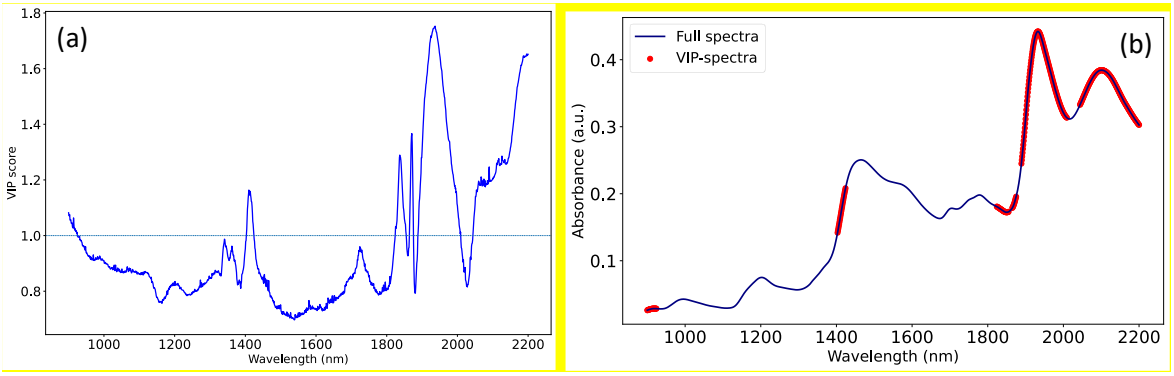

Figure 6. PLSR-VIP results: (a) VIP scores; (b) VIP-selected wavelengths overlaid on the mean spectrum

To further visualize the VIP-selected variables, the mean spectrum of the calibration set is shown with the VIP wavelengths (VIP > 1) overlaid (Figure 6b). The selected variables are mainly concentrated around the ~1450 nm region and the strong absorption region at ~1900–2000 nm, indicating that the most informative wavelengths for prediction coincide with moisture-sensitive O–H bands in the starch matrix. Compared with the shorter-wavelength region, the VIP-selected variables are more concentrated at longer wavelengths (~1450 and 1900–2000 nm), suggesting that the predictive information is dominated by strong O–H absorptions. Overall, this overlay supports the fact that the prediction is primarily driven by water-related absorptions rather than protein-associated features in purified corn starch.

LIMITATIONS

This study used an internal stratified train–test split for model evaluation. Therefore, the developed models have not yet been validated using truly independent external samples, such as different batches or brands, and calibration transfer across instruments was not assessed. Future work will include external validation and instrument-to-instrument transfer evaluation to confirm model generalizability and robustness under broader real-world variability.

CONCLUSION

This study confirmed that near-infrared (NIR) spectroscopy coupled with chemometric analysis can be used to rapidly and non-destructively quantify moisture content in corn starch. The spectral analysis identified prominent absorption bands at 990, 1195, 1462, 1930, and 2099 nm, confirming the presence of O-H and carbohydrate-related absorptions bonds associated with water and starch components. Comparison of regression models show that established chemometric methods outperformed the deep learning approach for this specific dataset. PLSR-SNV (R^2_p of 0.929, $RMSE_p$ of 0.274%) and SVR-MSC (R^2_p of 0.929, $RMSE_p$ of 0.273%) emerged as the most optimal performance. These results indicate that the relationship between spectral data and moisture content is primarily linear, which is effectively captured by traditional algorithms. Although the 1D-CNN model showed potential, its lower performance (R^2_p of 0.841) suggests that applying deep learning models is not suitable for small sized spectral datasets that are common in laboratory research.

Overall, the developed NIR calibration models offer an accurate alternative to the time-consuming gravimetric method and demonstrate strong potential for real-time quality control of corn starch.

FUNDING DETAILS

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

CONFLICT OF INTEREST

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

REFERENCES

Abdullah, N., Nawawi, A., & Othman, I. (2000). Fungal spoilage of starch-based foods in relation to its water activity (aw). Journal of Stored Products Research, 36(1), 47–54. https://doi.org/10.1016/S0022-474X(99)00026-0

Aenugu, H. P. R., Kumar, D. S., Parthiban, N., Ghosh, S. S., & Banji, D. (2011). Near Infra-Red Spectroscopy- An Overview. International Journal of ChemTech Research, 3(2), 825–836.

Ai, Y., & Jane, J. (2016). Macronutrients in Corn and Human Nutrition. Comprehensive Reviews in Food Science and Food Safety, 15(3), 581–598. https://doi.org/10.1111/1541-4337.12192

Amanah, H. Z., Rahayoe, S., Harmayani, E., Hernanda, R. A. P., Khoirunnisaa, Rohmat, A. S., & Lee, H. (2024). Construction of a sustainable model to predict the moisture content of porang powder (Amorphophallus oncophyllus) based on pointed-scan visible near-infrared spectroscopy. Open Agriculture, 9(1), 20220268. https://doi.org/10.1515/opag-2022-0268

Awad, M., & Khanna, R. (2015). Support Vector Regression. In M. Awad & R. Khanna, Efficient Learning Machines (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Bai, X., Zhang, L., Kang, C., Quan, B., Zheng, Y., Zhang, X., Song, J., Xia, T., & Wang, M. (2022). Near-infrared spectroscopy and machine learning-based technique to predict quality-related parameters in instant tea. Scientific Reports, 12(1), 3833. https://doi.org/10.1038/s41598-022-07652-z

Biliaderis, C. G. (2009). Structural Transitions and Related Physical Properties of Starch. In Starch (pp. 293–372). Elsevier. https://doi.org/10.1016/B978-0-12-746275-2.00008-2

Bjerrum, E. J., Glahder, M., & Skov, T. (2017). Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics (arXiv:1710.01927). arXiv. https://doi.org/10.48550/arXiv.1710.01927

Brouk, M. (2008). Corn Processing Co-Products. High Plains Diary Conference.

Brülls, M., Folestad, S., Sparén, A., Rasmuson, A., & Salomonsson, J. (2007). Applying spectral peak area analysis in near-infrared spectroscopy moisture assays. Journal of Pharmaceutical and Biomedical Analysis, 44(1), 127–136. https://doi.org/10.1016/j.jpba.2007.02.013

Büning-Pfaue, H. (2003). Analysis of water in food by near infrared spectroscopy. Food Chemistry, 82(1), 107–115. https://doi.org/10.1016/S0308-8146(02)00583-6

Burns, M. J., Renk, J. S., Eickholt, D. P., Gilbert, A. M., Hattery, T. J., Holmes, M., Anderson, N., Waters, A. J., Kalambur, S., Flint-Garcia, S. A., Yandeau-Nelson, M. D., Annor, G. A., & Hirsch, C. N. (2021). Predicting moisture content during maize nixtamalization using machine learning with NIR spectroscopy. Theoretical and Applied Genetics, 134(11), 3743–3757. https://doi.org/10.1007/s00122-021-03926-8

Chen, Y., Delaney, L., Johnson, S., Wendland, P., & Prata, R. (2017). Using near infrared spectroscopy to determine moisture and starch content of corn processing products. Journal of Near Infrared Spectroscopy, 25(5), 348–359. https://doi.org/10.1177/0967033517728146

Davies, A. M. C., & Fearn, T. (2006). Back to basics: Calibration statistics. Spectroscopy Europe, 18(2).

Ducanchez, A., Ryckewaert, M., Heran, D., & Bendoula, R. (2022). Discriminating between Absorption and Scattering Effects in Complex Turbid Media by Coupling Polarized Light Spectroscopy with the Mueller Matrix Concept. Sensors, 22(23), 9355. https://doi.org/10.3390/s22239355

Horwitz, W. & AOAC International (Eds.). (2006). Official methods of analysis of AOAC International (18. ed., current through rev. 1, 2006). AOAC International.

Jiao, Y., Li, Z., Chen, X., & Fei, S. (2020). Preprocessing methods for near-infrared spectrum calibration. Journal of Chemometrics, 34(11), e3306. https://doi.org/10.1002/cem.3306

Kusumaningrum, D., Lee, H., Lohumi, S., Mo, C., Kim, M. S., & Cho, B. (2018). Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy. Journal of the Science of Food and Agriculture, 98(5), 1734–1742. https://doi.org/10.1002/jsfa.8646

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Li, X., Xu, Z., Tang, L., Zhao, G., Wu, Y., Zhang, P., & Wang, Q. (2024). An effective moisture interference correction method for maize powder NIR spectra analysis. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 312, 124033. https://doi.org/10.1016/j.saa.2024.124033

Liu, C., Huang, W., Yang, G., Wang, Q., Li, J., & Chen, L. (2020). Determination of starch content in single kernel using near-infrared hyperspectral images from two sides of corn seeds. Infrared Physics & Technology, 110, 103462. https://doi.org/10.1016/j.infrared.2020.103462

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. Chem. Soc. Rev., 43(24), 8200–8214. https://doi.org/10.1039/C4CS00062E

Mishra, P., Passos, D., Marini, F., Xu, J., Amigo, J. M., Gowen, A. A., Jansen, J. J., Biancolillo, A., Roger, J. M., Rutledge, D. N., & Nordon, A. (2022). Deep learning for near-infrared spectral data modelling: Hypes and benefits. TrAC Trends in Analytical Chemistry, 157, 116804. https://doi.org/10.1016/j.trac.2022.116804

Ng, W., Minasny, B., & McBratney, A. (2020). Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. Science of The Total Environment, 702, 134723. https://doi.org/10.1016/j.scitotenv.2019.134723

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology, 46(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Nie, B., Du, Y., Du, J., Rao, Y., Zhang, Y., Zheng, X., Ye, N., & Jin, H. (2023). A novel regression method: Partial least distance square regression methodology. Chemometrics and Intelligent Laboratory Systems, 237, 104827. https://doi.org/10.1016/j.chemolab.2023.104827

Nie, Z., Tremblay, G. F., Bélanger, G., Berthiaume, R., Castonguay, Y., Bertrand, A., Michaud, R., Allard, G., & Han, J. (2009). Near-infrared reflectance spectroscopy prediction of neutral detergent-soluble carbohydrates in timothy and alfalfa. Journal of Dairy Science, 92(4), 1702–1711. https://doi.org/10.3168/jds.2008-1599

Ozbekova, Z., & Kulmyrzaev, A. (2019). Study of moisture content and water activity of rice using fluorescence spectroscopy and multivariate analysis. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 223, 117357. https://doi.org/10.1016/j.saa.2019.117357

Padhi, S. R., John, R., Tripathi, K., Wankhede, D. P., Joshi, T., Rana, J. C., Riar, A., & Bhardwaj, R. (2024). A Comparison of Spectral Preprocessing Methods and Their Effects on Nutritional Traits in Cowpea Germplasm. Legume Science, 6(2), e2977. https://doi.org/10.1002/leg3.229

Pizarro, C., Esteban-Díez, I., Nistal, A.-J., & González-Sáiz, J.-M. (2004). Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. Analytica Chimica Acta, 509(2), 217–227. https://doi.org/10.1016/j.aca.2003.11.008

Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization: An overview. Swarm Intelligence, 1(1), 33–57. https://doi.org/10.1007/s11721-007-0002-0

Rinnan, Å., Berg, F. V. D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, 28(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Safarzadegan Gilan, S., Bahrami Jovein, H., & Ramezanianpour, A. A. (2012). Hybrid support vector regression – Particle swarm optimization for prediction of compressive strength and RCPT of concretes containing metakaolin. Construction and Building Materials, 34, 321–329. https://doi.org/10.1016/j.conbuildmat.2012.02.038

Sawatsky, M. L., Clyde, M., & Meek, F. (2015). Partial least squares regression in the social sciences. The Quantitative Methods for Psychology, 11(2), 52–62. https://doi.org/10.20982/tqmp.11.2.p052

Walsh, J., Neupane, A., & Li, M. (2024). Evaluation of 1D convolutional neural network in estimation of mango dry matter content. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 311, 124003. https://doi.org/10.1016/j.saa.2024.124003

Wang, S., Li, C., Copeland, L., Niu, Q., & Wang, S. (2015). Starch Retrogradation: A Comprehensive Review. Comprehensive Reviews in Food Science and Food Safety, 14(5), 568–585. https://doi.org/10.1111/1541-4337.12143

Wang, Y., Cao, H., Zhou, Y., & Zhang, Y. (2015). Nonlinear partial least squares regressions for spectral quantitative analysis. Chemometrics and Intelligent Laboratory Systems, 148, 32–50. https://doi.org/10.1016/j.chemolab.2015.08.024

Westad, F., & Marini, F. (2015). Validation of chemometric models – A tutorial. Analytica Chimica Acta, 893, 14–24. https://doi.org/10.1016/j.aca.2015.06.056

Weyer, L. G., & Lo, S. -C. (2001). Spectra– Structure Correlations in the Near-Infrared. In J. M. Chalmers & P. R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy (1st ed.). Wiley. https://doi.org/10.1002/0470027320.s4102

Whistler, R. L., & BeMiller, J. N. (2009). Starch: Chemistry and technology (3rd ed). Academic Press.

Workman, Jr., Jerry, & Weyer, L. (2007). Practical Guide to Interpretive Near-Infrared Spectroscopy (0 ed.). CRC Press. https://doi.org/10.1201/9781420018318

Yan, C. (2025). A review on spectral data preprocessing techniques for machine learning and quantitative analysis. iScience, 28(7), 112759. https://doi.org/10.1016/j.isci.2025.112759

Yu, J.-K., & Moon, Y.-S. (2021). Corn Starch: Quality and Quantity Improvement for Industrial Uses. Plants, 11(1), 92. https://doi.org/10.3390/plants11010092

Zhang, J., Guo, Z., Ren, Z., Wang, S., Yue, M., Zhang, S., Yin, X., Gong, K., & Ma, C. (2023). Rapid determination of protein, starch and moisture content in wheat flour by near-infrared hyperspectral imaging. Journal of Food Composition and Analysis, 117, 105134. https://doi.org/10.1016/j.jfca.2023.105134

Zhang, Y., & Guo, W. (2020). Moisture content detection of maize seed based on visible/near-infrared and near-infrared hyperspectral imaging technology. International Journal of Food Science & Technology, 55(2), 631–640. https://doi.org/10.1111/ijfs.14317