

Research

Artificial intelligence-driven near-infrared spectrophotometry model for rapid quantification of anti-nutritional factors in soybean (*Glycine max.*)

Norberto Jose Palange^{1,2,3}  · Tonny Obua¹  · Julius Pyton Sserumaga⁴  · Enoch Wembabazi⁵ · Mildred Ochwo-Ssemakula¹ · Ephraim Nuwamanya^{1,5} · Isaac Onziga Dramadri^{1,2} · Moses Matovu⁴ · Richard Edema^{1,2} · Phinehas Tukamuhabwa¹

Received: 5 February 2025 / Accepted: 30 May 2025

Published online: 15 June 2025

© The Author(s) 2025 

Abstract

Anti-nutritional factors can impact soybean nutrient bioavailability when consumed by *monogastric* animals. However, conventional methods available for quantifying anti-nutritional factors such as phytate and trypsin inhibitors in feeds are laboratory-intensive, time-consuming, expensive, and error-prone. This study developed near-infrared spectrophotometry (NIRS)-based models to quantify phytate and trypsin inhibitors in soybean. Thus, a set of 190 soybean genotypes assayed through conventional wet chemistry was used as a reference for model development and cross-validation. Using a benchtop NIR instrument (DS2500), spectra readings between 400 and 2500 nm were taken from each soybean sample. Mean values for phytate and total trypsin inhibitors (TTI) were 1.77 mg g^{-1} (SD = 1.23) and 0.89 mg g^{-1} (SD = 0.24), respectively. Predictive models were developed through partial least squares (PLS) and random forest (RF) regressions. The random forest models outperformed partial least squares regression with the best predictive performance of $R^2_{\text{test}} = 0.97$; RPD = 5.95 and $R^2_{\text{test}} = 0.96$; RPD = 3.62 for phytate and TTI, respectively. The high R^2 and RPD values demonstrate the model's strong predictive capability and accuracy, suggesting that the NIRS-based models can effectively quantify phytate and TTI in soybean. Thus, breeders can efficiently select for low-anti-nutritional genotypes and accelerate the development of nutritionally beneficial legumes while reducing soybean processing costs. NIRS offers a promising alternative to traditional phenotyping methods due to its speed, simplicity, environmental friendliness, and cost-effectiveness. Its integration into breeding programs can streamline the screening process, especially in early selection stages.

Keywords Soybeans · Phytate · Trypsin inhibitors · Near-infrared spectrophotometry (NIRS) · Partial least squares · Random forest

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42452-025-07235-3>.

✉ Julius Pyton Sserumaga, j.serumaga@gmail.com; Norberto Jose Palange, cheltonpalange@gmail.com | ¹Department of Crop Science and Horticulture, College of Agriculture and Environmental Science, Makerere University, P. O. Box 7062, Kampala, Uganda. ²Makerere Regional Centre for Crop Improvement (MaRCCI), College of Agriculture and Environmental Science, Makerere University, P. O. Box 7062, Kampala, Uganda. ³Faculty of Food and Agricultural Sciences, Rovuma University, P.O. Box 544, Nampula, Mozambique. ⁴National Agricultural Resources Research Organization, Livestock Resources Research Institute (NaLIRRI), P.O. Box 5704, Kampala, Uganda. ⁵National Agricultural Resources Research Organization, National Crops Resources Research Institute (NaCRRI), P. O. Box 7084, Kampala, Uganda.



Discover Applied Sciences

(2025) 7:647

| <https://doi.org/10.1007/s42452-025-07235-3>

1 Introduction

Soybean [*Glycine max*, $2n = 40$] has high protein content (40–50%), lipids (20–30%), carbohydrates (26–30%) [15, 24, 28, 38] and micronutrients (3.93–6.15%) [7]. However, the presence of anti-nutritional factors (ANFs) potentially reduces soybean nutritional value when consumed by monogastric mammals, including humans [27]. The most common ANFs in soybean include trypsin inhibitors, indigestible polysaccharides and phytate [37, 41]. To tackle the deleterious effects of ANFs, several processing methods have been developed including physical, chemical, and biochemical which can be applied singly or in combination [41, 45, 53]. The drawbacks of these methods include high costs, the risk of denaturing soybean proteins, and, sometimes, the need for special expertise. To overcome these limitations, breeding for low ANFs content remains the most cost-effective approach for developing nutritious soybean varieties. However, the success of this approach in part relies on the breeder's ability to screen germplasms for ANFs at varying selection stages. Conventionally, in the breeding pipelines, breeders primarily focus on agronomic traits, mainly selecting yield-related traits [30, 31], disease resistance [29], and at times protein content [31], often neglecting other biochemical qualities of importance in animal nutrition. Whereas the biochemical traits are desirable for inclusion in breeders' selection indexes, they are frequently overlooked due to cost constraints associated with available phenotyping methods. Blind selection of genotypes due to low phenotyping capability for ANFs restricts breeders to screening ANFs at later stages of the breeding pipeline, when few lines are left, resulting in genetic erosion [49]. Thus, rapid and cost-effective phenotyping methods are needed to prevent elimination of lines with low ANF content during selection process. Usually, profiling of ANFs such as phytate and trypsin inhibitors is based on conventional wet chemistry methods, categorized as follows:

(i) Chromatography-based methods, including high-performance liquid chromatography (HPLC) and gas chromatography-mass spectrometry (GC-MS), are primarily used for complex compound separation and quantification, although less commonly applied to phytate and trypsin inhibitors (TI) due to their polar nature and the need for derivatization [35, 37], (ii) Spectrophotometric methods, such as colorimetric assays using UV-Vis spectrophotometry, are widely used for both phytate and TI due to their simplicity, affordability, and moderate sensitivity [2], (iii) Atomic absorption spectroscopy (AAS) is used to estimate phytate indirectly through the measurement of mineral chelation (e.g., Zn^{2+} or Fe^{3+} binding), providing supporting data on phytate's bioavailability impact [44], (iv) Immunoassay-based methods, like enzyme-linked immunosorbent assay (ELISA), offer high specificity and sensitivity but are more frequently used for proteinaceous ANFs like trypsin inhibitors rather than for small molecules like phytate [47], (v) Gravimetric analysis, while labor-intensive, can be used to measure crude fiber-bound components including total phytate in some traditional protocols [10]. While effective for routine laboratory analysis, the methods are often resource-intensive, labor-intensive, and may require specialized equipment and trained personnel. Additionally, they are environmentally unfriendly due to the copious amounts of chemical waste generated in the laboratories. Therefore, rapid, cost-effective phenotyping methods for assessing the biochemical traits along with the agronomic attributes at all selection stages in breeding are of utmost importance. Dry chemistry-based methods are a reliable and efficient option, while near-infrared spectroscopy (NIRS) is emerging as a promising alternative. NIRS is capable of capturing directly, in a lapse of time, a variety of compounds based on the vibrational behavior of their chemical bonds [6]. NIRS has successfully quantified hydrogen cyanide (HCN) potential in cassava [19], nitrogen, minerals, sugars, and starch in tropical root and tuber crops [22], moisture, crude fat, and protein content of soybeans [54], protein, oil and fatty acids in soybean [51], seed germination and vigor [1] as well as nutritional and anti-nutritional factors in soybean [5].

Spectral variation of interest can generally be masked by multiplicative light scattering, background noise artifacts, and baseline shifts due to differences in atom sizes and chemical properties [33]. Therefore, appropriate data pretreatment techniques are applied to diminish the physic-chemical effects of NIRS calibration [21, 32]. The most popular spectral treatments applied prior to NIRS model development are standard normal variate (SNV) and de-trending [4], Savitzky-Golay smoothing and derivatives [8]. For NIRS prediction model development, methods such as partial least squares (PLS) regression and random forest (RF) are widely employed for complex data analysis. Among this, PLS is generally preferred to handle linear relationship between the spectral data and the traits or when interpretability of the model is crucial [50, 52]. On the other hand, the RF, a machine learning algorithm, has been regarded suitable for datasets where there are non-linear relationships or when there is a need to handle large amounts of predictors and samples with complex interactions [12]. Although NIRS has been implemented and proven accurate

for various traits, this tool has not yet been explored for phytate and trypsin inhibitors phenotyping in the Ugandan soybean breeding programs. Therefore, this study developed near-infrared spectrophotometry (NIRS)-based models to quantify phytate and trypsin inhibitors in soybean in a rapid, environmentally friendly, cost-effective, simple, accurate and reliable way using minimal sample preparation.

2 Material and methods

2.1 Sample collection

Seeds from 190 soybean genotypes were collected from the Makerere University Center for Soybean Improvement and Development (MAKCSID). They included lines collected in Uganda (80), Zimbabwe (6), Taiwan (23), the USA (70), and Japan (11). In 2023, the genotypes were planted at the Makerere University Agricultural Research Institute Kabanyolo (MUARIK) to ensure consistent seed multiplication and standardized seed conditions.

2.2 Study site

MUARIK is in Uganda's Central region at 0°28' N, 32°36' E, and 1,180 m above sea level. The average annual temperature is 21.4 °C and an average annual rainfall of 1234 mm [31]. The experiment was set up using 19 blocks in an augmented design with ten plots and two replications. Each plot consisted of three rows, each spaced 60 cm apart (inter-row spacing) and with plants spaced 5 cm apart within the row (intra-row spacing).

2.3 Sample preparation

Dried soybean seeds (oven-dried at 60 °C for 48 h) were milled using a food processor grinder (FOSS, Brook Crompton Laboratory mill, type 2-TDAB03J, England, 2014) at the Nutritional and Bioanalytical Laboratory of the National Crops Resources Research Institute (NaCRRI) in Uganda and used for quantification of phytate and trypsin inhibitors and scanning by near-infrared spectrophotometry (NIRS) to generate spectra data.

2.4 Wet chemistry

Phytate extraction was done according to Israel (2013) using an acidic method, with slight modifications on the sample initial volume and refrigeration during centrifugation. Briefly, 20 ml of 0.5 M HCl was added to 0.5 g of finely powdered soybean sample, which was then vortexed and shaken on an orbital shaker for 1 h. The homogenized mixture was centrifuged at 4,000 rpm for 30 min at 4 °C using a HERMLE 300 K centrifuge (HERMLE Labortechnik GmbH; Siemenstr. 25, 78564 Wehingen, Germany). For each sample, 5 ml of the supernatant containing the soluble fraction of the compound was transferred into three new Falcon tubes, and the pH was adjusted by adding an equal volume of 0.5 M NaOH. The reaction was terminated by adding 5 ml of NaCl. The sample was then filtered, and 2 ml of a chromogenic solution (ferric chloride III, FeCl₃) was added for spectrophotometric (Anthos Mikrosysteme GmbH; Burkamp 22, 26169 Friesoythe, Germany) reading at 492 nm. The optical densities (OD) were used to calculate the concentration of phytate in the samples as equivalents of an external sodium phytate calibration standard, using a linear regression calibration curve prepared from five serial dilutions of the standard solution. Optical control samples, prepared similarly, but with no analyte added, were subtracted from the final concentration.

Total trypsin inhibitors (TTI) were initially extracted and quantified using the method described by Anozie et al. [2], with minor modifications including adjustments to sample size, refrigeration during centrifugation, and substitution of perchloric acid with trichloroacetic acid (TCA) for protein precipitation. Briefly, 0.05 g of powdered soybean sample was homogenized with 5 ml of phosphate-buffered saline (PBS), vortexed, and shaken on an orbital shaker for 1 h. The homogenate was centrifuged at 10,000 rpm for 10 min at 4 °C (HERMLE 300 K centrifuge, Germany), and 0.1 ml of the supernatant was mixed with 0.1 ml of 1 mg/ml trypsin solution. After a 10-min incubation at 0 °C, 0.3 ml of a 2% casein substrate was added [2, 34], followed by a 20-min incubation at 37 °C in a water bath. The reaction was stopped with 0.2 ml of 10% TCA, and the mixture was centrifuged at 10,000 rpm for 5 min to remove undigested casein and protein

precipitates. The absorbance of the supernatant was read at 410 nm. All steps were performed in triplicate with appropriate blanks (inhibitor-free and casein-only). It is important to note that the Anozie et al. [2] method measures trypsin inhibitor activity (TIA) rather than absolute trypsin inhibitor content, which can limit interpretability when comparing across genotypes or integrating results into breeding pipelines. Specifically, this method presents two main limitations:

- Lack of extract dilution optimization: Trypsin inhibition may fall outside the linear range, which can lead to underestimation or overestimation of TIA.
- No clear conversion protocol: The method does not provide a validated or standardized approach for converting % inhibition into trypsin inhibitor content (mg/g), which can introduce variability depending on assumptions made during conversion.

Despite these limitations, this method was chosen for its simplicity, cost-effectiveness, and previous successful use in low-resource settings for comparative screening purposes. However, we acknowledge its shortcomings, particularly for use as a reference method in building NIR calibration models. To improve the reliability and accuracy of our findings, we are currently validating a subset of samples using the ISO 14902:2001/AOCS Ba 12–75 standard method for determining trypsin inhibitor units (TIU) via standardized enzymatic assay. Preliminary results indicate a strong positive correlation between TIA values from the Anozie et al. method and the ISO/AOCS-based method, suggesting that while the former is less precise, it may still serve as a viable screening tool when appropriately calibrated. This validation is ongoing, and we anticipate that integrating the standardized method into our NIRS training set will further enhance model accuracy and reproducibility.

2.5 Spectral data collection

A benchtop FOSS NIRS™ DS2500 (FOSS Analytical, DK-3400 Hilleroed, Denmark) equipment was used to collect spectral data from the soybean powder samples. To minimize the effects of scattering and noise artifacts in the NIRS data, a combination of methods was employed. These included fine grinding and maintaining uniform sample thickness to reduce variability, controlling temperature and humidity, and fully covering the sample cup base with powder for consistent sample distribution.

For spectral data collection, each genotype was scanned three times, and 4200 data points (0.5 nm interval each) were generated in approximately 60 s in the 400 to 2500 nm range.

2.6 Data analysis

The triplicate spectral data per genotype were averaged using the *aggregate* function of the *stats* R package. Outliers from the response variables and predictors were detected and removed using the interquartile range (IQR) method in R, a widely used statistical approach for robust outlier detection in high-throughput phenotyping datasets [42]. The distribution of training, test, and total datasets was assessed using the *ggplot2* and *gridExtra* packages for visualization, while the *dplyr* package was used to efficiently group, summarize, and manipulate data structures prior to plotting. Skewness in phytate and TTI variables was corrected using *log* and *Box-Cox* transformations, respectively. Descriptive statistics for phytate and TTI in the training and test sets were computed using *dplyr* package in R. Various pretreatment techniques, including *standard normal variate* (SNV), *de-trending* [4], and *Savitzky–Golay* (SG) filtering with a 2nd-order polynomial, window size of 11 and 1st derivative parameters ($p=2$, $w=11$, $m=1$) [8], were applied independently or in combination to the spectral data. These techniques were implemented using the *prospectr*, *viridis*, and *reshape2* packages in R to address baseline shifts, scattering, and other non-sample related signals. Among these methods, SNV alone demonstrated satisfactory performance, as it consistently enhanced the signal-to-noise ratio and improved model calibration by effectively correcting multiplicative scatter effects without introducing artifacts or overfitting. SNV's simplicity and robustness also contributed to better cross-validation accuracy and generalizability compared to more complex or combined preprocessing strategies. The spectral pattern and principal component were plotted using untreated full spectral data using *ggplot2* in R. Models were developed through partial least squares (PLS) and random forest (RF) regressions using full range untreated and treated spectra, respectively, using *pls* and *random forest* packages in R.



Fig. 1 Distribution for training, test and total sets for phytate and trypsin inhibitors

Table 1 Descriptive statistics for the reference data (phytate and TTI), training and testing sets at the 80:20 proportion

Parameters	Mean	Median	Min	Max	SD	Variance	Skewness	Kurtosis
Phytate	1.77	1.60	0.02	6.14	1.23	1.52	1.14	1.20
TTI	0.89	0.93	0.07	1.34	0.24	0.06	-1.18	1.74
DSTrS (TrS)_Phytate	1.79	1.62	0.01	6.14	1.27	1.61	1.16	1.14
DSTS (TS)_Phytate	1.68	1.59	0.19	4.10	1.07	1.15	0.68	-0.44
DSIS (TrS)_TTI	0.90	0.93	0.07	1.34	0.24	0.06	-1.26	1.96
DSIS (TS)_TTI	0.88	0.93	0.25	1.28	0.26	0.07	-0.79	-0.20

DSTrS (TrS) —Descriptive statistics for training set; DSTS (TS) —Descriptive statistics for testing set; SD—standard deviation; Min—minimum; max—maximum; TTI—total trypsin inhibitors

2.7 NIRS prediction model development and cross-validation statistics

Partial least squares (PLS) and random forest (RF) models were used to regress spectral data with the concentration of ANFs in the sample, using the *pls* and *caret* packages in R. For model development data were split into training (80%) and test (20%) sets using the *createDataPartition* function of *caret* package in R. For PLS model, the number of components was 10, and in the random forest model the number of trees was 500 (*ntree* = 500). A ten-fold repeated cross-validation (repeated 3 times) was performed to assess the predictive performance of the generated models using the *predict* function of the *caret* package in R. This resampling strategy helps reduce variance in model performance estimates by averaging results across multiple partitions of the data, thereby enhancing robustness and reliability of the evaluation. The performance metrics for the trained model including standard error of calibration (SEC) and “out-of-bag error” (OOB Error) were estimated using *randomForest* package in R. For the cross validation statistics, the Coefficient of Determination (R^2), Mean Squares Error (RMSE), Mean Absolute Error (MAE), Standard Deviation (SD), Ratio of Performance to Deviation (RPD), Concordance Correlation Coefficient (CCC) were computed using *metrics* package in R. Importance of variables in the spectral data was assessed by random forest model and visualized by the *varImpPlot* function in R. Regression curves for observed and predicted values for phytate and TTI were plotted using Microsoft excel 2016.

3 Results

3.1 Reference data, training, and test sets statistics

Phytate mean value was 1.77 mg g^{-1} , ranging from 0.02 to 6.14 mg g^{-1} and total trypsin inhibitors, mean was 0.89 mg g^{-1} , ranging from 0.07 to 1.34 mg g^{-1} . Following outlier removal and transformation, the data for both traits exhibited a normal distribution (Fig. 1). The mean of training set for phytate (TrS)_Phytate is 1.79 and test set (TS)_Phytate, 1.59. For trypsin inhibitors, the mean for training set (TrS)_TTI is 0.90 and test set (TS)_TTI, 0.88 (Table 1).

3.2 Spectral pattern and principal components

The spectra showed peaks at approximately 490, 1200, 1500, 1750, 2100, and 2450 nm. The first two principal components (PC) explained most of the variation in the spectra data where PC1 (68.77%) and PC2 (18.68%) cumulatively explained 87.44% variation (Fig. 2). The first six PCs explain up to 98.44% of total variation in the spectra (PC1 = 0.69; PC2 = 0.19; PC3 = 0.064; PC4 = 0.03; PC5 = 0.001 and PC6 = 0.001).

3.3 Model development and cross-validation

Models generated by the partial least squares method had generally low prediction accuracy irrespective of spectra pretreatment. The highest prediction accuracy for phytate was $R^2_{\text{test}} = 0.46$ (RPD = 1.33) using combined SNV + SG_d1 treatments and for total trypsin inhibitors, $R^2_{\text{test}} = 0.08$ (RPD = 1.02) with SNV treatment (Supplementary Table A.1 and A.2). Using random forest, the training set treated with SNV showed the highest coefficient of determination for the calibration

Fig. 2 Full spectra range of near-infrared spectrophotometry of soybean powder (A) and principal component analysis (PCA) of the first two principal components (B)

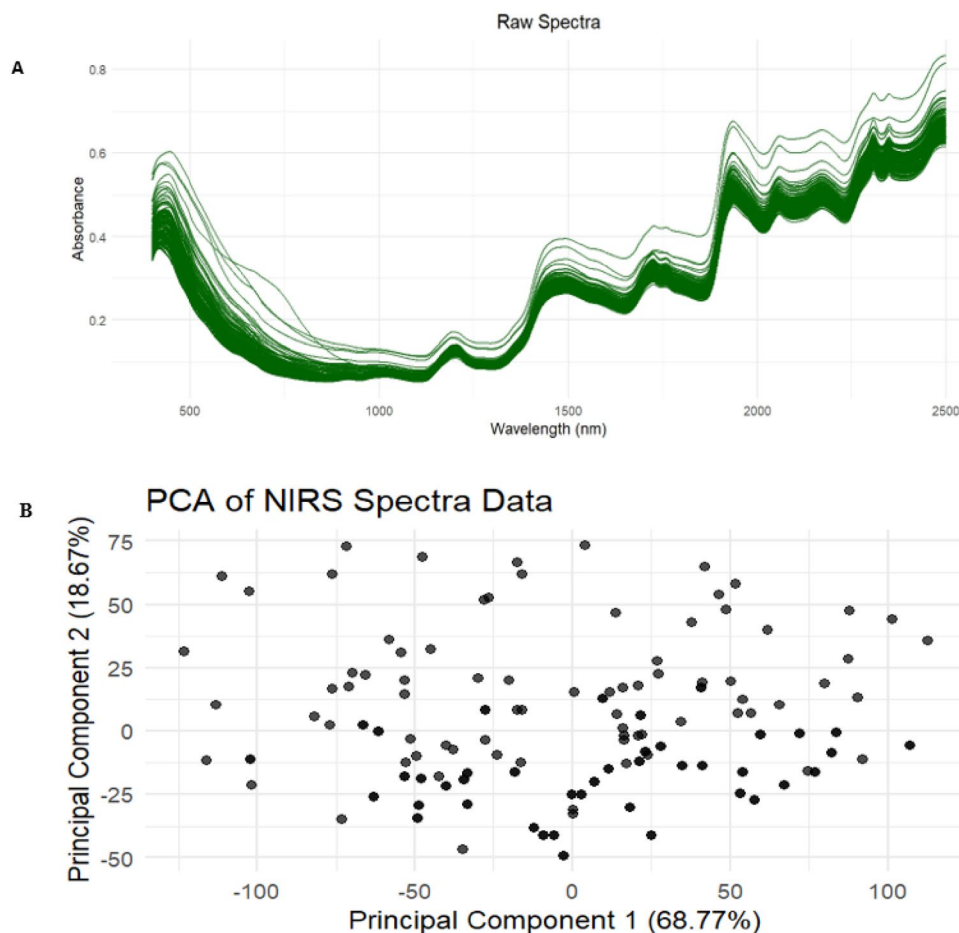


Table 2 Training set statistics used for calibrating the random forest model

Trait	Spectra	Pretreatment	<i>ntree</i>	R^2_{train}	SEC	OOB Error
Model_Phytate	400–2500	No	500	0.95	0.34	0.83
Model_TTI	400–2500	No	500	0.93	0.09	0.24
Model_Phytate	400–2500	SNV	500	0.99	0.32	0.81
Model_TTI	400–2500	SNV	500	0.99	0.04	0.23
Model_Phytate	400–2500	SG_d0	500	0.94	0.35	0.85
Model_TTI	400–2500	SG_d0	500	0.93	0.09	0.24
Model_Phytate	400–2500	SG_d1	500	0.96	0.28	0.74
Model_TTI	400–2500	SG_d1	500	0.93	0.09	0.22
Model_Phytate	400–2500	SG_d2	500	0.97	0.27	0.71
Model_TTI	400–2500	SG_d2	500	0.95	0.09	0.22
Model_Phytate	400–2500	De-trending	500	0.96	0.28	0.73
Model_TTI	400–2500	De-trending	500	0.94	0.09	0.22
Model_Phytate	400–2500	SNV, SG_d1	500	0.96	0.28	0.71
Model_TTI	400–2500	SNV, SG_d1	500	0.93	0.09	0.24

TTI—total trypsin inhibitors; *ntree*=number of trees; R^2_{train} —coefficient of determination for the training set; SEC—standard error of calibration; OOB Error—out-of-bag error; TTI—total trypsin inhibitors; SNV—standard normal variate; SG—Savitzky-Golay; SG_d0—Savitzky-Golay with no derivative; SG_d1—Savitzky-Golay with first derivative; SG_d2—Savitzky-Golay with second derivative

Table 3 Test set statistics used for cross-validating the random forest model

Trait	Spectra	Treatment	<i>ntree</i>	R^2_{test}	RMSE	MAE	SD	RPD	CCC
Model_Phytate	400–2500	No	500	0.75	0.56	0.43	1.07	1.92	0.82
Model_TTI	400–2500	No	500	0.54	0.56	0.13	0.26	1.39	0.58
Model_Phytate	400–2500	SNV	500	0.97	0.54	0.44	1.07	5.95	0.98
Model_TTI	400–2500	SNV	500	0.96	0.07	0.05	0.26	3.62	0.95
Model_Phytate	400–2500	SG_d0	500	0.74	0.57	0.44	1.07	1.88	0.81
Model_TTI	400–2500	SG_d0	500	0.52	0.19	0.13	0.26	1.36	0.56
Model_Phytate	400–2500	SG_d1	500	0.77	0.57	0.47	1.07	1.89	0.83
Model_TTI	400–2500	SG_d1	500	0.56	0.18	0.13	0.26	1.44	0.61
Model_Phytate	400–2500	SG_d2	500	0.73	0.61	0.5	1.07	1.76	0.79
Model_TTI	400–2500	SG_d2	500	0.64	0.18	0.13	0.26	1.46	0.606
Model_Phytate	400–2500	De-trending	500	0.8	0.49	0.41	1.07	2.2	0.88
Model_TTI	400–2500	De-trending	500	0.43	0.2	0.14	0.26	1.3	0.53
Model_Phytate	400–2500	SNV, SG_d1	500	0.82	0.52	0.43	1.07	2.1	0.86
Model_TTI	400–2500	SNV, SG_d1	500	0.46	0.2	0.15	0.26	1.33	0.54

ntree—number of trees; R^2 —coefficient of determination for the test set; RMSE—Root mean squared error; MAE—mean absolute error; SD—standard deviation; RPD—ratio of performance to deviation; CCC—concordance correlation coefficient; TTI—total trypsin inhibitors; SNV—standard normal variate; SG—Savitzky-Golay; SG_d0—Savitzky-Golay with no derivative; SG_d1—Savitzky-Golay with first derivative; SG_d2—Savitzky-Golay with second derivative

set ($R^2_{\text{cal}}=0.99$) compared to other pretreatments for both traits. The cross-validation accuracy in the untreated spectra showed a coefficient of determination for the test set ($R^2_{\text{test}}=0.75$) and 0.54 for phytate and total trypsin inhibitors, respectively. After spectra treatment with the SNV method, the R^2_{test} for phytate was 0.97 with a ratio of performance to deviation (RPD)=5.95, whereas that for TTI was $R^2_{\text{test}}=0.96$ with an RPD=3.62. High prediction accuracies were also found when spectra were treated with SNV + SG_d1, de-trending and SG_d1 with $R^2_{\text{test}}=0.82$, 0.80, and 0.77, respectively, for phytate. Lower prediction accuracies of $R^2_{\text{test}}=0.46$, 0.43, and 0.53 for TTI were achieved using the same treatments (Table 2 and 3). The linear regression analysis between observed and predicted values for phytate and total trypsin inhibitors produced a high coefficient of determination value, indicating a strong positive correlation and confirming the model's ability to accurately predict the targeted anti-nutritional factors (Fig. 3A and B).

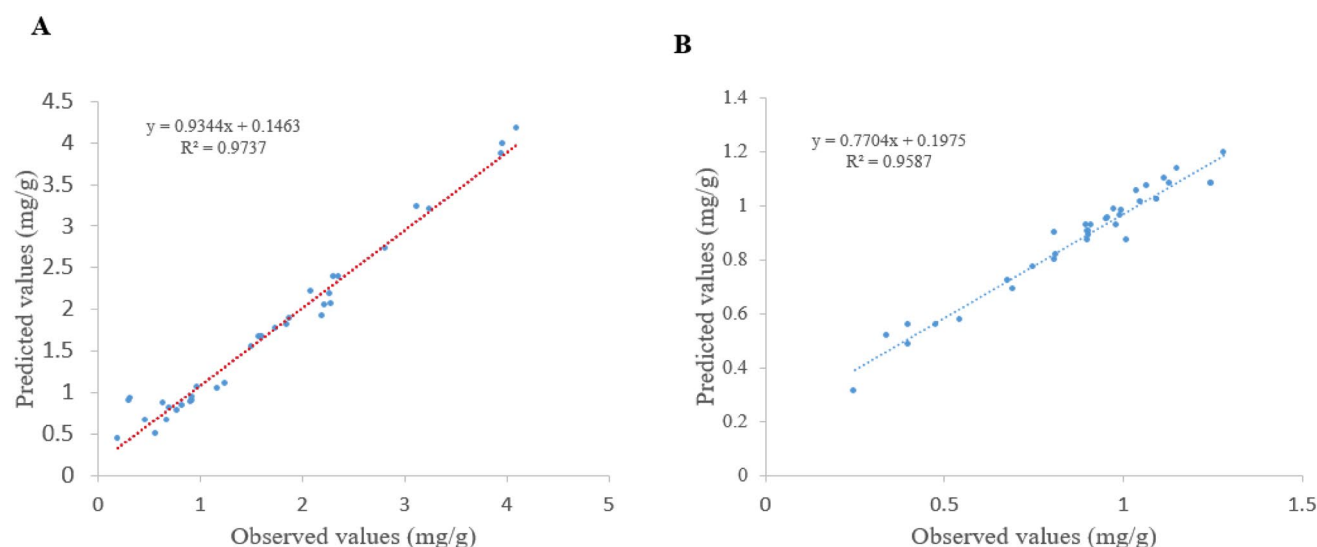


Fig. 3 Observed versus predicted values for phytate and trypsin inhibitors. **A** Depicts the observed versus predicted values for phytate and **B** for TTI

3.4 Variable importance in the spectra data

Spectral variable importance for the model's accuracy and decision-making process was assessed through the percentage increase in mean squared error ($\%IncMSE$) and the extent of contribution of each spectral variable to the purity of the nodes in the decision trees by the increase node purity ($IncNodePurity$). The parameters helped to identify most influential spectral variable for the random forest model performance. For phytate building model, the spectral variables including X419, X419.5, X418.5 and X418 both in the ($\%IncMSE$ and $IncNodePurity$) showed to be more critical for the model's accuracy and decision-making process, as they appear at the top of the graphs and have the highest importance values (Fig. 4A). For total trypsin inhibitors, X1887, X451.5, X1886, X681 and X1888 ($\%IncMSE$) and X683.5, X659, X458, X457, X461.5 and X459 ($IncNodePurity$) were the most influential, as confirmed by the higher importance values (Fig. 4B).

4 Discussion

NIRS has been reported to be capable of capturing directly, in a lapse of time, a variety of compounds based on the vibrational behavior of their chemical bonds [6] in a variety of crops including tuber crops [19, 22], soybean [5, 9, 54]. Thus, this study aimed to develop NIRS-based prediction models to aid rapid, simple but reliable, environmentally friendly and cost-effective quantification of phytate and total trypsin inhibitors (TTI) in soybean.

Soybean reference samples used to develop the model had higher concentrations of phytate than total trypsin inhibitors, similar to what has been reported in a study aimed at optimizing the reduction of phytate and trypsin inhibitors in soybean meal [39]. However, other studies have reported lower concentrations of phytate than trypsin inhibitors [14, 43].

The principle behind NIRS is the interaction of light photons with the chemical bonds present in organic compounds, such as the O–H, C–H, N–H, and C=O groups [19], and S=O [11]. Thus, the nature of the chemical bond defines the pattern of light absorption. Peaks around 1100 nm have previously been associated with aromatic compounds [11]. In this study, a peak was found around 1200 nm which could be related to phytate (1L-myo-inositol-1,2,3,4,5,6-hexakis) [26]. Whereas the peak at 1400 nm may be indicative of presence of water in the samples [11, 19].

In this study, the first and second principal components (PC1 and PC2) cumulatively accounted for 87.44% of the variation in the population. In contrast, lower principal components of 71.8% explained total variation on water-soluble protein content analysis in soybean [50]. Higher PCs reported in this study indicate that the principal component is a strong representation of the underlying patterns and trends in the data, and thus could be a valuable tool for the modeling.

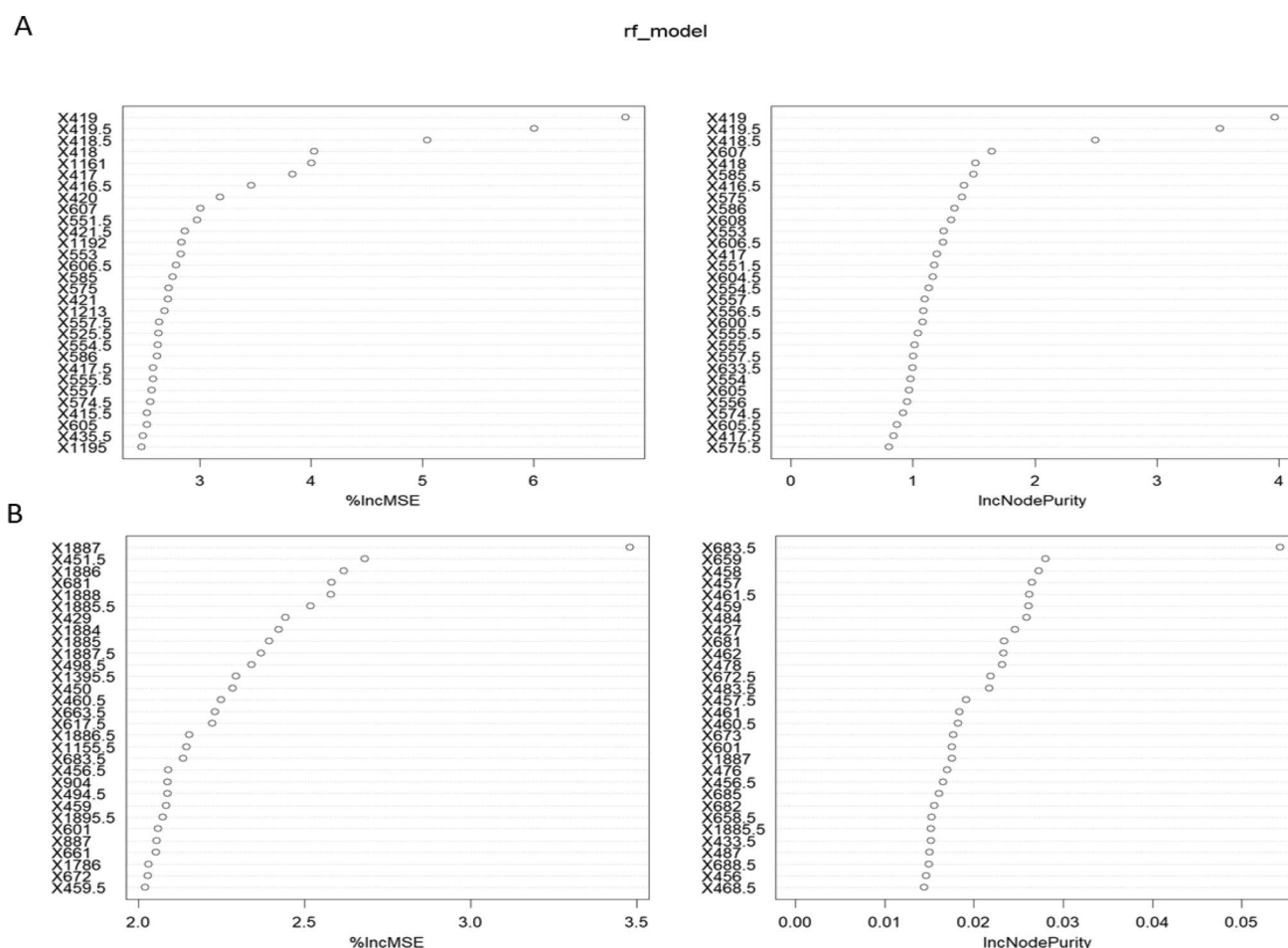


Fig. 4 Spectral variable importance for phytate (**A**) and trypsin inhibitors (**B**). Both graphs help to identify which spectral variables are most influential for the random forest model accuracy. Spectral variables with longer lines are more critical for refining the model, indicating that they have the most significant impact on the model's predictions

The training sets for both phytate and trypsin inhibitors had higher means and wider ranges compared to the testing sets, indicating a more diverse distribution of values in the training data. This diversity contributes to a more robust model that can better handle varying input conditions. Similarly, better model performance was observed when evaluating water-soluble proteins in soybean, as the training set also had a higher mean and range than the test set [50]. Thus, the wider ranges suggest that the trained model is better equipped for generalization to the test set.

Random forest statistically outperformed partial least squares for both targeted traits using full spectral data, indicating that the response variables have a non-linear relationship with spectra data. Whereas partial least squares deals with linear patterns and collinear predictors [22], random forest is a more flexible, non-linear model based on regression of multiple decision trees and averages their predictions [11].

Partial least squares model has been reported to have poor prediction accuracy for $R^2 = 0.030$ for total phosphorus and 0.012 for phytate phosphorus [46]. Although PLS model has shown high predictive capability for water soluble protein ($R^2_{cv} = 0.83$) [50], in this study the model was outperformed by random forest predictive model. Random forest has also been reported to have better predictive performance than partial least squares regression models for soil quality parameters [11]. However, partial least squares classification model, which is an extension of PLS regression, performed better in classified spectral data for hydrogen cyanide in cassava [19]. Random forest, a machine learning algorithm, often outperforms linear models like partial least squares (PLS) and linear discriminant analysis (LDA) in terms of predictive performance and interpretability, mainly for large, non-linear datasets [25]. Random forest's ability to handle high-dimensional data and capture complex patterns makes it well-suited for a variety of data analysis and revealing conditional relationships between variables [48].

Spectra pretreatment has previously been reported to have a substantial impact on model prediction performance [17, 40, 50]. For instance, spectra pretreatments were associated with significant differences in prediction accuracy for hydrogen cyanide in fresh cassava [19]. These pretreatment steps are generally crucial for ensuring that the predictive models associate the true chemical information from the sample, rather than being misled by unnecessary confounding variations which could lower prediction accuracy [36]. In this study, the standard normal variate (SNV) method, singly or in combination with SG_d1, was the most effective pretreatment method, indicating that it was able to effectively reduce noise, correct for scattering and baseline variations related to non-chemical related signals [33, 50]. SNV normalizes the spectral data, leading to more accurate and robust models [4]. Interestingly, it was found that the same pretreatment could have different effects on prediction accuracy depending on the nature of the response variable. Similarly, same NIRS spectra pretreatment (SNV-DT) had different effects on prediction accuracy for nutritional parameters such as amylose and sugar content in cowpea [32].

Using random forest model, the prediction accuracy in the data pretreated with SNV was as high as $R^2_{\text{test}} = 0.97$ (RPD = 5.95) and $R^2_{\text{test}} = 0.96$ (RPD = 4.32) for phytate and total trypsin inhibitors, respectively. This indicates that the models are capable of efficiently quantifying anti-nutritional factors in soybean powder. The superior performance of *RF* with SNV pretreatment can be attributed to its ability to model complex, non-linear relationships between the predictors and the response variables. By preprocessing the data with SNV, the underlying relationships may become more linear, making it easier for the *RF* model to capture these patterns effectively and predict more accurately [4, 33].

Differently from what has been reported on intact yellow cotyledon seeds [50], studies have shown that soybean powder is more suitable for modeling than whole grains [20, 54]. A study aiming to predict pasture nutritive values has reported R^2_{cal} of 0.89–0.99, with a RPD of 2.6 for dry matter content, but higher (RPD = 10.3) for crude protein [23]. Although studies have demonstrated that PLS shows promise in predicting various chemical properties across different fields, its overall performance is still comparatively lower than random forest [12, 16, 18, 52]. Similarly, lower prediction accuracy, $R^2_{\text{cv}} = 0.74$ has been reported previously for trypsin inhibitors activity using partial least squares method [16] and $R^2_{\text{cv}} = 0.84$ when analyzed for phosphorus concentration in intact seeds [13]. Quantification of phosphorus is often used as an indirect method of assessing phytate levels in crops, as phytate is the primary organic phosphorus storage form [3, 13].

The contribution of each spectral variable to the model performance was also assessed. Variables captured by both methods (%*IncMSE* and *IncNodePurity*) are critical for the model's accuracy, indicating that their absence can significantly reduce the prediction performance by increasing the mean square error. Increasing node purity (*IncNodePurity*) help to make splits in the data, leading to more homogeneous groups within the nodes generated by random forest model. Thus, the most influential variables can be benchmarked to reduce collinearity, thereby improving the accuracy of linear predictions. Alternatively, these variables can be used to eliminate uninformative ones, which would otherwise increase the computation time in *RF* models.

The NIRS models developed in this study will enable soybean breeders to perform rapid phenotyping and make objective, accurate genotype selections based on phytate and TTI levels. This approach allows for the evaluation of a larger number of genotypes in a shorter timeframe. It also offers a cost-effective alternative to traditional laboratory methods, saving both time and resources. Breeding soybeans with reduced levels of phytate and TTI will improve soybean nutritional value for animal and human consumption, decreasing the need for costly processing steps to reduce these anti-nutritional factors.

5 Conclusion

This research addressed the challenges breeders face in selecting genotypes based on biochemical traits. For the first time in Uganda and beyond, the developed NIRS model will enable rapid, high-throughput phenotyping of anti-nutritional factors in soybeans, significantly streamlining the selection process. It represents a major step forward in soybean breeding, offering breeders an efficient tool for evaluating nutritional quality at scale. Furthermore, the study highlights the power of NIRS for rapid, environmentally friendly and cost-effective phenotyping of crops.

Future steps will include external validation using samples from diverse locations, as well as exploring the broader applications of this NIRS model beyond soybeans, potentially benefiting other crops with similar breeding objectives. It may also be interesting to determine if there are specific wavelengths for better model performance for each trait.

Acknowledgements To the Partnership for Applied Skills in Sciences, Engineering and Technology-Regional Scholarship and Innovation Fund (PASET-RSIF) and Carnegie Corporation of New York for funding this work. To the Makerere Regional Centre for Crops Improvement (MaRCCI) for providing the facilities, and Makerere University Centre for Soybean Improvement and Development (MAKCSID) for providing the soybean material used in this study. Extensively to the National Agricultural Research Organization (NARO), and the technicians at the National Crops Resources Research Institute (NaCRRI) for their valuable support during data collection.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Norberto Jose Palange, Tonny Obua, Julius Pyton Sserumaga, Enoch Wembabazi and Ephraim Nuwamanya. Phinehas Tukamuhabwa, Mildred Ochwo-Ssemakula and Moses Matovu supervised the work. Isaac Onziga Dramadri and Richard Edema helped with data curation. The first draft of the manuscript was written by Norberto Jose Palange and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Partnership for Applied Skills in Sciences, Engineering and Technology-Regional Scholarship and Innovation Fund (PASET-RSIF) [B8501G30218], and Carnegie Corporation of New York.

Data availability All relevant data are presented in tables and figures in this manuscript. Further inquiries can be addressed to the corresponding authors.

Code availability Not applicable.

Declarations

Ethics approval and consent to participate The seeds used in the study are owned by the Makerere University Centre for Soybean Improvement and Development (MAKCSID) led by Senior Soybean Breeder in Uganda, Prof. Phinehas Tukamuhabwa. Therefore, the collection of the seeds used in the study complies with local or national guidelines with no need for further affirmation.

Consent for publication All authors have read and agreed to the published version of the manuscript.

Competing interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Al-Amery M, Geneve RL, Sanches MF, Armstrong PR, Maghirang EB, Lee C, Vieira RD, Hildebrand DF. Near-infrared spectroscopy used to predict soybean seed germination and vigour. *Seed Sci Res.* 2018;28(3):245–52. <https://doi.org/10.1017/S0960258518000119>.
2. Anozie AN, Salami OA, Babatunde DE, Babatunde OE. Comparative evaluation of processes for production of soybean meal for poultry feed in Nigeria. *Evaluación comparativa de procesos para la producción de harina.* Anim Sci. 2018;52(2):193–202.
3. Aureli R, Ueberschlag Q, Klein F, Noël C, Guggenbuhl P. Use of near infrared reflectance spectroscopy to predict phytate phosphorus, total phosphorus, and crude protein of common poultry feed ingredients. *Poult Sci.* 2017;96(1):160–8. <https://doi.org/10.3382/ps/pew214>.
4. Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc.* 1989;43(5):772–7.
5. Bueno RD, Borges LL, Good God PIV, Piovesan ND, Teixeira AI, Cruz CD, De Barros EG. Quantification of anti-nutritional factors and their correlations with protein and oil in soybeans. *Anais Da Academia Brasileira de Ciencias.* 2018;90(1):205–17. <https://doi.org/10.1590/0001-3765201820140465>.
6. Cen H, He Y. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci Technol.* 2007;18(2):72–83. <https://doi.org/10.1016/j.tifs.2006.09.003>.
7. Chauhan D, Kumar K, Ahmed N, Thakur P, Rizvi QUEH, Jan S, Yadav AN. Impact of soaking, germination, fermentation, and roasting treatments on nutritional, anti-nutritional, and bioactive composition of black soybean (*Glycine max* L.). *J Appl Biol Biotechnol.* 2022;10(5):186–92. <https://doi.org/10.7324/JABB.2022.100523>.
8. Chen H, Song Q, Tang G, Feng Q, Lin L. The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models. *ISRN Spectrosc.* 2013;2013:1–9. <https://doi.org/10.1155/2013/642190>.
9. Choung M-G, Baek I-Y, Kang S-T, Han W-Y, Shin D-C, Moon H-P, Kang K-H. Determination of protein and oil contents in soybean seed by near infrared reflectance spectroscopy. *Korean J Crop Sci.* 2001;46(2):106–11.

10. Chuwa C, Dhiman AK, Saidia P, Issa-Zacharia A. Physico-chemical characteristics and the effects of processing methods on the nutritional and anti-nutritional quality of soybean (*Glycine max* (L.) Merrill). *Asian Food Sci J*. 2023;22(10):60–9. <https://doi.org/10.9734/afsj/2023/v22i10673>.
11. de Santana FB, de Souza AM, Poppi RJ. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta Part A Mol Biomol Spectrosc*. 2018;191:454–62. <https://doi.org/10.1016/j.saa.2017.10.052>.
12. Douglas RK, Nawar S, Alamar MC, Mouazen AM, Coulon F. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Sci Total Environ*. 2018;616–617:147–55. <https://doi.org/10.1016/j.scitotenv.2017.10.323>.
13. Eifler J, Wick JE, Steingrobe B, Möllers C. Genetic variation of seed phosphorus concentration in winter oilseed rape and development of a NIRS calibration. *Euphytica*. 2021. <https://doi.org/10.1007/s10681-021-02782-3>.
14. Galão OF, Carrão-panizzi MC, Marcos J, Mandarin G, Leite RS, Claus T, Visentainer JV. Kunitz trypsin Inhibitor and phytic acid levels in conventional and genetically modified soybean seeds from Londrina and Ponta Grossa, South Brazil. *Acta Scientiarum Technol*. 2014. <https://doi.org/10.4025/actascitechnol.v36i4.17711>.
15. Guo B, Sun L, Jiang S, Ren H, Sun R, Wei Z, Hong H, Luan X, Wang J, Wang X, Xu D, Li W, Guo C, Qiu LJ. Soybean genetic resources contributing to sustainable protein production. *Theor Appl Genet*. 2022;135(11):4095–121. <https://doi.org/10.1007/s00122-022-04222-9>.
16. Hoffmann D, Brugger D, Windisch W, Thurner S. Calibration model for a near infrared spectroscopy (NIRS) system to control feed quality of soy cake based on feed value assessments in-vitro. *Chem Eng Trans*. 2017;58(August):379–84. <https://doi.org/10.3303/CET1758064>.
17. Jin JW, Chen ZP, Li LM, Stepanavicius R, Thennadil SN, Yang J, Yu RQ. Quantitative spectroscopic analysis of heterogeneous mixtures: the correction of multiplicative effects caused by variations in physical properties of samples. *Anal Chem*. 2012;84(1):320–6. <https://doi.org/10.1021/ac202598f>.
18. Kamal M, Munawar AA, Sulaiman MI. Comparison of principal component and partial least square regression method in NIRS data analysis for cocoa bean quality assessment. *IOP Conf Ser Earth Environ Sci*. 2021. <https://doi.org/10.1088/1755-1315/667/1/012058>.
19. Kanaabi M, Namakula FB, Nuwamanya E, Kayondo IS, Muhumuza N, Wembabazi E, Iragaba P, Nandudu L, Nanyonjo AR, Baguma J, Esuma W, Ozimati A, Settumba M, Alicai T, Ibanda A, Kawuki RS. Rapid analysis of hydrogen cyanide in fresh cassava roots using NIR-Sand machine learning algorithms: meeting end user demand for low cyanogenic cassava. *Plant Genome*. 2024;17(2):1–14. <https://doi.org/10.1002/tpg2.20403>.
20. Ho-Sun L, Jung-Bong K, Lee Y-Y, Lee S-Y, Gwang J-G, Baek HJ, Kim C-K, Yoon MS. Estimating the important components in three different sample types of soybean by near infrared reflectance spectroscopy. *Korean J Crop Sci*. 2011;56(1):88–93.
21. Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom Intell Lab Syst*. 2012;117:100–14. <https://doi.org/10.1016/j.chemolab.2012.03.011>.
22. Lebot V, Malapa R, Jung M. Use of NIRS for the rapid prediction of total N, minerals, sugars and starch in tropical root and tuber crops. *N Z J Crop Hortic Sci*. 2013;41(3):144–53. <https://doi.org/10.1080/01140671.2013.798335>.
23. Lobos I, Gou P, Hube S, Saldaña R, Alfaro M. Evaluation of potential nirs to predict pastures nutritive value. *J Soil Sci Plant Nutr*. 2013;13(2):463–8. <https://doi.org/10.4067/S0718-95162013005000036>.
24. Malle S, Morrison M, Belzile F. Identification of loci controlling mineral element concentration in soybean seeds. *BMC Plant Biol*. 2020;20(419):1–14.
25. Marchese Robinson RL, Palczewska A, Palczewski J, Kidley N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J Chem Inf Modeling*. 2017. <https://doi.org/10.1021/acs.jcim.6b00753>.
26. Martins V, Ferrari F, White PJ. Plant Physiology and Biochemistry Phytic acid accumulation in plants: biosynthesis pathway regulation and role in human diet. *Plant Physiol Biochem*. 2021;164:132–46. <https://doi.org/10.1016/j.plaphy.2021.04.035>.
27. Mohapatra D, Patel AS, Kar A, Deshpande SS, Tripathi MK. Effect of different processing conditions on proximate composition, antioxidants, anti-nutrients and amino acid profile of grain sorghum. *Food Chem*. 2019;271:129–35. <https://doi.org/10.1016/j.foodchem.2018.07.196>.
28. Murithi HM, Beed F, Tukamuhabwa P, Thomma BPHJ, Joosten MHAJ. Soybean production in eastern and southern Africa and threat of yield loss due to soybean rust caused by *Phakopsora pachyrhizi*. *Plant Pathol*. 2016;65(2):176–88. <https://doi.org/10.1111/ppa.12457>.
29. Murithi HM, Soares RM, Mahuku G, van Esse HP, Joosten MHAJ. Diversity and distribution of pathotypes of the soybean rust fungus *Phakopsora pachyrhizi* in East Africa. *Plant Pathol*. 2021;70(3):655–66. <https://doi.org/10.1111/ppa.13324>.
30. Nadeem MA, Karaköy T, Yeken MZ, Habyarimana E, Hatipoglu R, Çiftçi V, Nawaz MA, Sönmez F, Shahid MQ, Yang SH, Chung G, Baloch FS. Phenotypic characterization of 183 Turkish common bean accessions for agronomic, trading, and consumer-preferred plant characteristics for breeding purposes. *Agronomy*. 2020. <https://doi.org/10.3390/agronomy10020272>.
31. Obua T, Nabasiyire M, Namara M, Tusiime G, Maphosa M, Tukamuhabwa P. Yield stability of tropical soybean genotypes in selected agro-ecologies in Uganda. *S Afr J Plant Soil*. 2020;37(2):168–73. <https://doi.org/10.1080/02571862.2019.1678687>.
32. Padhi SR, John R, Tripathi K, Pandhari D, Tanay W. A comparison of spectral preprocessing methods and their effects on nutritional traits in cowpea germplasm. *Legume Sci*. 2024;6:1–13. <https://doi.org/10.1002/leg3.229>.
33. Pizarro C, Esteban-Díez I, Nistal AJ, González-Sáiz JM. Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Anal Chim Acta*. 2004;509(2):217–27. <https://doi.org/10.1016/j.aca.2003.11.008>.
34. Kakade ML, Rackis JJ, McGhee JE, Puski G. (1974). Determination of trypsin inhibitor activity of soy products: a collaborative analysis of an improved procedure. *Cereal Chem*. 1974;51(3): 376–382. <http://www.aaccnet.org/cerealchemistry/abstracts/1974/cc1974a45.asp>.
35. Rahmawati D, Astawan M, Putri SP, Fukusaki E. Gas chromatography-mass spectrometry-based metabolite profiling and sensory profile of Indonesian fermented food (tempe) from various legumes. *J Biosci Bioeng*. 2021;132(5):487–95. <https://doi.org/10.1016/j.jbiosc.2021.07.001>.
36. Roger J, Supagro M. Pre-processing Methods. In *Comprehensive chemometrics: chemical and biochemical data analysis*. (Second Edition, Vol. 3, pp. 1–75). Elsevier, 2020. <https://doi.org/10.1016/B978-0-12-409547-2.14878-4>.

37. Rosso ML, Shang C, Song Q, Escamilla D, Gillenwater J, Zhang B. Development of breeder-friendly KASP markers for low concentration of kunitz trypsin inhibitor in soybean seeds. *Int J Mol Sci*. 2021;2675(22):1–16.
38. Rui X, Wang M, Zhang Y, Chen X, Li L, Liu Y. Optimization of soy solid-state fermentation with selected lactic acid bacteria and the effect on the anti-nutritional components. *J Food Process Preserv*. 2017. <https://doi.org/10.1111/jfpp.13290>.
39. Saah BF, Tene ST, Boungo GT, Klang JM. Optimization of the reduction of phytates and trypsin inhibitors of soybeans (*Glycine max* L): effect of soak-ing and cooking. *Int J Food Nutr Sci*. 2020;7(1):70–6. <https://doi.org/10.15436/2377-0619.20.2900>.
40. Sampaio PS, Soares A, Castanho A, Almeida AS, Oliveira J, Brites C. Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chem*. 2018;242:196–204. <https://doi.org/10.1016/j.foodchem.2017.09.058>.
41. Samtiya M, Aluko RE, Dhewa T. Plant food anti-nutritional factors and their reduction strategies: an overview. *Food Prod Proc Nutr*. 2020;5(2):1–14.
42. Sharma A, Singh P, Singh R, Yadav V. Outlier detection techniques for high-dimensional biological data: a review. *J Comput Biol*. 2022;29(3):295–310. <https://doi.org/10.1089/cmb.2021.0215>.
43. Sharma S, Goyal R, Barwal S. Domestic processing effects on physicochemical, nutritional and anti-nutritional attributes in soybean (*Glycine max* L. Merrill). *Int Food Res J*. 2013;20(6):3203–9.
44. Siulapwa N, Mwambungu A. Nutritional value of differently processed soybean seeds. *Indian J Nutr Dietetics*. 2015; 52(1), 45–56. <http://www.ijndonline.org/index.php/ijnd/article/view/56476>.
45. Suhag R, Dhiman A, Deswal G, Thakur D, Sharanagat VS, Kumar K, Kumar V. Microwave processing: a way to reduce the anti-nutritional factors (ANFs) in food grains. *LWT*. 2021;150:111960. <https://doi.org/10.1016/j.lwt.2021.111960>.
46. Tahir M, Shim MY, Ward NE, Westerhaus MO, Pesti GM. Evaluation of near-infrared reflectance spectroscopy (NIRS) techniques for total and phytate phosphorus of common poultry feed ingredients. *Poult Sci*. 2012;91(10):2540–7. <https://doi.org/10.3382/ps.2012-02211>.
47. Tan GY, Nan TG, Gao W, Li QX, Cui JJ, Wang BM. Development of monoclonal antibody-based sensitive sandwich ELISA for the detection of antinutritional factor cowpea trypsin inhibitor. *Food Anal Methods*. 2013;6(2):614–20. <https://doi.org/10.1007/s12161-012-9459-6>.
48. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, Sacha van Hijum AFT. Data mining in the life science swith random forest: a walk in the park or lost in the jungle? *Brief Bioinform*. 2013;14(3):315–26. <https://doi.org/10.1093/bib/bbs034>.
49. Van de Wouw M, Kik C, van Hintum T, van Treuren R, Visser B. Genetic erosion in crops : concept, research results and challenges. *Plant Genetic Resour*. 2009;8(1):1–15. <https://doi.org/10.1017/S1479262109990062>.
50. Xu R, Hu W, Zhou Y, Zhang X, Xu S, Guo Q, Qi P, Chen L, Yang X, Zhang F, Liu L, Qiu L, Wang J. Use of near-infrared spectroscopy for the rapid evaluation of soybean [*Glycine max* (L) Merril] water soluble protein content. *Spectrochimica Acta Part A: Mol Biomol Spectrosc*. 2020;224:117400. <https://doi.org/10.1016/j.saa.2019.117400>.
51. Yakubu AB, Shaibu AS, Mohammed SG, Ibrahim H, Mohammed IB. NIR-based prediction for protein, oil, and fatty acids in soybean (*Glycine max* (L.) Merrill) seeds. *Food Anal Methods*. 2024;17:1592–600. <https://doi.org/10.1007/s12161-024-02678-7>.
52. Zhao N, Wu ZS, Zhang Q, Shi XY, Ma Q, Qiao YJ. Optimization of parameter selection for partial least squares model development. *Sci Rep*. 2015;5:1–10. <https://doi.org/10.1038/srep11647>.
53. Zhong Y, Wang Z, Zhao Y. Impact of radio frequency, microwaving, and high hydrostatic pressure at elevated temperature on the nutritional and antinutritional components in black soybeans. *J Food Sci*. 2015;80(12):C2732–9. <https://doi.org/10.1111/1750-3841.13131>.
54. Zhu Z, Chen S, Wu X, Xing C, Yuan J. Determination of soybean routine quality parameters using near-infrared spectroscopy. *Food Sci Nutr*. 2018;6(4):1109–18. <https://doi.org/10.1002/fsn3.652>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com