Journal of Food Composition and Analysis 135 (2024) 106692

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

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

Evaluating the dry matter content of raw yams using hyperspectral imaging spectroscopy and machine learning

Michael Adesokan [a,b], Bolanle Otegbayo [b,*], Emmanuel Oladeji Alamu [a,c,**], Michael Afolabi Olutoyin [b], Busie Maziya-Dixon [a]

[a] Food and Nutrition Sciences Laboratory, International Institute of Tropical Agriculture (IITA), Oyo Road, Ibadan, Oyo State 200284, Nigeria
[b] Department of Food Science and Technology, Bowen University, Iwo, Osun State 232102, Nigeria
[c] Food and Nutrition Sciences Laboratory, International Institute of Tropical Agriculture, Southern Africa Hub, P.O. Box 310142, Chelstone, Lusaka, Zambia

ARTICLE INFO

Keywords:
Fresh intact yam
Dry matter content
Hyperspectral imaging
Machine learning

ABSTRACT

Yams (Dioscorea spp.) are important food and commercial crops in West African countries. They contribute significantly to global food production and provide dietary energy. The quality of yam food products depends on specific internal and external parameters, such as the DMC and other biochemical traits. However, measuring these traits can be challenging, particularly when analyzing many genotypes. This study aimed to evaluate the feasibility of using near-infrared (NIR) hyperspectral imaging (932–1721 nm) along with machine learning to rapidly measure the dry matter content (DMC) of fresh, intact yam tubers. Hyperspectral images were acquired across the yam tuber's cross-sections, and the resulting spectra from the images were averaged and preprocessed. Partial least square regression (PLSR) combined with successive progressions algorithms (SPA), Competitive Adaptive Reweighted Sampling (CARS), Artificial Neural network (ANN) and Boruta algorithms (BA) were used to select the important wavelengths for developing a prediction model for DMC (g/100 g). The PLSR-SPA-CARS model showed the most accurate prediction performances with a coefficient of determinations in calibration (R^2cal) and prediction (R^2pred) of **0.974** and **0.958**, respectively, and low root mean square error (RMSEP) of 0.898 g/100 g. The distribution of DMC was visually represented by projecting the developed model to generate color chemical maps. This study resolves that NIR hyperspectral imaging can rapidly assess the DMC of fresh, intact yam tubers.

1. Introduction

Yam (Dioscorea spp.) is a staple food and economic crop in West African countries, particularly Benin, Ghana, Côte d'Ivoire, Cameroon, and Nigeria. These countries account for 98 % of global crop production (FAOSTAT, 2021), with Nigeria accounting for 73 % of total African production (FAOSTAT, 2021). As a critical root and tuber (RTB) crop, it contributes significantly to world food production and offers dietary energy in the form of carbohydrates in human and animal diets. Yam is a food security crop because it may be stored and used during the lean months or dry seasons, particularly in the tropics, where farmers rely on rainfed agricultural production. Rising consumer demand for high-quality food products derived from RTB crops necessitates detailed characterization of their external and internal quality traits, influencing their end products' cooking behaviour and acceptability. There is a high demand for precise and cost-effective techniques for determining the quality of food products to ensure they meet customer's quality expectations (Sun, 2009). Recent advances in quality characterization of RTB crops have employed near-infrared hyperspectral imaging techniques (NIR-HSI) as a non-destructive, rapid, and environmentally friendly alternative (Peng et al., 2022; Adesokan et al., 2023) to conventional approaches, which could be costly, time-consuming, and may involve exposure to harmful chemical materials, which in most cases are not responsibly disposed of, thus, negatively impacting the environment.

NIR-HSI is a high-throughput, non-destructive technique that uses spectroscopy and imaging to determine food products' quality

* Corresponding author.
** Corresponding author at: Food and Nutrition Sciences Laboratory, International Institute of Tropical Agriculture (IITA), Oyo Road, Ibadan, Oyo State 200284, Nigeria.
E-mail addresses: m.adesokan@cgiar.org (M. Adesokan), bolanle.otegbayo@bowen.edu.ng (B. Otegbayo), o.alamu@cgiar.org, dejialamu2001@gmail.com (E.O. Alamu), afolabi.michael@bowen.edu.ng (M.A. Olutoyin), b.maziya-dixon@cgiar.org (B. Maziya-Dixon).

https://doi.org/10.1016/j.jfca.2024.106692
Received 26 June 2024; Received in revised form 26 August 2024; Accepted 27 August 2024
Available online 30 August 2024
0889-1575/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

characteristics and spatial distribution. It has demonstrated the ability to characterize biochemical and biophysical components simultaneously, including their spatial distribution within the target sample (ElMasry and Nakauchi, 2016; Adesokan et al., 2023). Several authors have described the use of imaging techniques for the reliable quantification of multiple qualitative features of roots and tuber crops such as sweet potatoes, potato, cassava, yam, taro and sugar beet (Wang et al. 2022; Pan et al., 2015; Su et al. 2019; Su et al., 2020, Heo et al., 2021 and Luo et al., 2020). Meghar et al. 2023 recently acquired hyperspectral images of 31 boiled cassava genotypes to quantify the dry matter, water absorption and textural attributes. A coefficient of multiple determination of 0.94 was reported, with root mean square error in prediction (RMSEP) of 0.96 g/100 g, which shows the potential of NIR-HSI for quality characterization of contrasting cassava genotypes.

However, large data size, multidimensional spectral information, noisy signals and collinearity of spectral bands are some of the significant drawbacks of the technique. Therefore, appropriate image processing algorithms must be implemented to reduce the dimension of the spectra data while retaining important information and minimising the data analysis time (Ray et al., 2021). Food materials have a complex matrix, which may result in variations in acquired data; therefore, sample presentation must be carefully chosen to minimise noisy signals and ensure a representative image of the target is acquired (de Araújo *et al.*, 2023). Within the RTBfoods project, water distribution in fresh tubers was monitored across the longitudinal section of yam slices (Meghar et al., 2022). Further study was recommended by Adesokan et al. (2023) to investigate the distribution of traits across the cross-section of RTB crops and to examine the variations in quality parameters from different sections of the same tuber. The conventional DMC analysis methods for RTB crops, including yam, are expensive, take much time, and are destructive to the samples (Alamu et al., 2021). It sometimes involves extensive sample preparation steps, such as peeling, cutting, and drying in the oven for 72 hours; these create limitations in improving the breeding cycles for the crops.

The dry matter content of RTB crops and other essential quality parameters could be obtained on the intact yam tubers with minimal sample preparation, making on-farm phenotyping possible and minimising the time required for the analysis workflow. Prediction models for determining the critical quality traits in RTB crops using hyperspectral imaging will provide cost-effective, high-throughput phenotyping tools to breeding programs and food processors. In addition, visualization of the spatial distribution of key quality parameters in RTB crops is essential in helping to understand the biochemical changes that occur during storage and post-harvest activities. Shao et al. (2020) reported that the spatial distribution of the soluble sugar content of sweet potatoes was possible using hyperspectral imaging techniques. Anthocyanin content prediction in purple sweet potatoes using hyperspectral imaging was also reported by Tian et al. (2021) in their study. Presently, there is scanty information on the applications of NIR-HSI on yam quality analysis; therefore, our hypothesis is to explore the potential of hyperspectral imaging systems to accurately predict the DMC of fresh yam without destroying the samples. The current study developed prediction models for rapidly determining the DMC of intact fresh yams using PLSR and other machine learning algorithms. The optimum prediction model was used to establish a spatial distribution map for the DMC (g/100 g) of fresh yam tubers.

2. Materials and methods

2.1. Sample preparations

Sixteen yam genotypes with contrasting food quality, comprising landraces and improved varieties, were collected from the experimental field plots at the IITA stations in Ibadan, Ubiaja and Abuja (Supplementary Table S1). Healthy and large yam tubers from each genotype were washed, peeled, and allowed to dry at room temperature. After preparation, an approximately 2 cm thick slice was cut off from both ends of the yam tubers to eliminate the stalk (Adinsi et al., 2021). Then, each tuber was cut into three cross-sections, each about 3 cm thick, using a stainless-steel knife and a measuring tape (Adinsi et al. 2021). These sections were labelled as proximal (P), central (C) and distal (D). 288 sampling points comprised 144 samples (16 genotypes x 3 cross-sections x 3 locations) x 2 replicate measurements was generated during the analysis. After hyperspectral imaging analysis, each sample was placed into a well-labelled paper bag and prepared for wet laboratory analysis. See Fig. 1 for the sampling and analysis workflow.

2.2. Hyperspectral imaging system and image acquisition

This study used a push broom near-infrared hyperspectral imaging system (Fig. 2) consisting of a 12-bit (CCD camera (V-light; Lowel Light Inc., New York, NY, USA) with a two-dimensional (2D) light detector. It has a spectrograph with high sensitivity from 932 to 1721 nm and a spectral resolution of 8 nm. A translation Lab Scanner with a dimension of 40 cm×20 cm (LxW) was attached for image acquisition, and an illumination component made up of 150 W tungsten halogen lamps (Fibre-Lite DC950 Illuminator; Dolan Jenner Industries Inc., Boxborough, MA, USA). The distance of 500 mm was between the halogen lamps and the sample and at an angle of 45 ^{0}C to the horizontal plane. The exposure time of the camera and the moving speed in the translation Lab scanner stage was set at 2 ms and 40 mm/s, respectively. A Lumo control software (SPECIM) was installed on a computer for systems operations. Due to the non-uniform distribution of light and the effect of dark current in the camera, it is important to calibrate the raw hyperspectral image with white and dark calibration references (Equ.1). This is done using a white Teflon board with diffuse reflectance of 99 % (Spectralon, SRT-99–100, Labsphere Inc, North Sutton, NH, USA) to obtain white calibration references (R_w). Also, a black calibration reference (R_d) with a reference of 0 % was obtained by closing the camera shutter. Images of fresh, intact yam tubers were collected using optimum acquisition parameters such as exposure time of 4 m/s, binning of 2 and scanning speed of 16.5 mm/s. The acquired image is a three-dimensional (3D) hypercube with (x, y, and z) dimensions of the image pixel, where (x, y) represents spatial dimension and (z) represents spectral wavelength (224 nm).

$$Rs = \frac{Ro - Rd}{Rw - Rd} \quad (1)$$

where Rs = calibrated reflectance image; Ro = raw hyperspectral image; Rw = white reference image and Rd = dark reference image.

2.3. Image segmentation and region of interest (ROI)

Breeze (version 2024.1.0) and Evince Professional (version 2.7.21) are commercial software packages provided by Prediktera AB (Tvistevagen 48 A SE-907 36 Umea, Sweden) used for processing 3-dimensional hyperspectral images. These images are multidimensional and

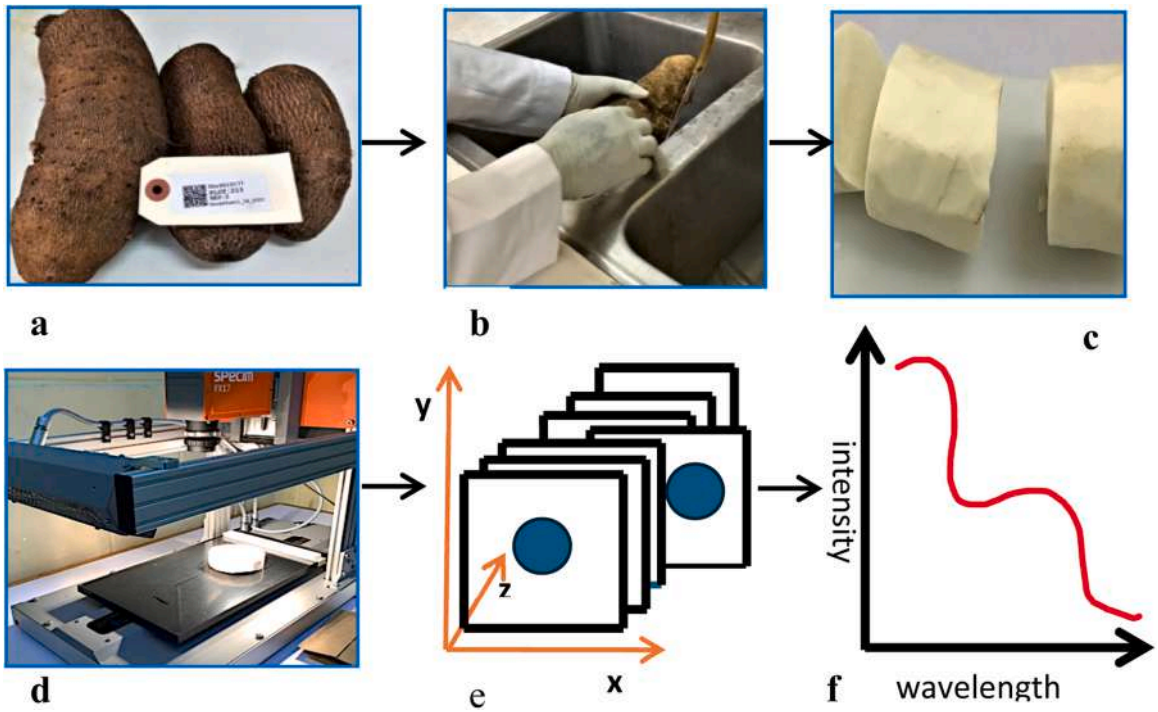

Fig. 1. Workflow for sample preparations and image acquisitions (a)Fresh yam tubers, (b) washing, (c) peeled tubers cut into cross-sections, (d) hyperspectral image system (e) 3-dimensional hypercube (f) average spectrum.

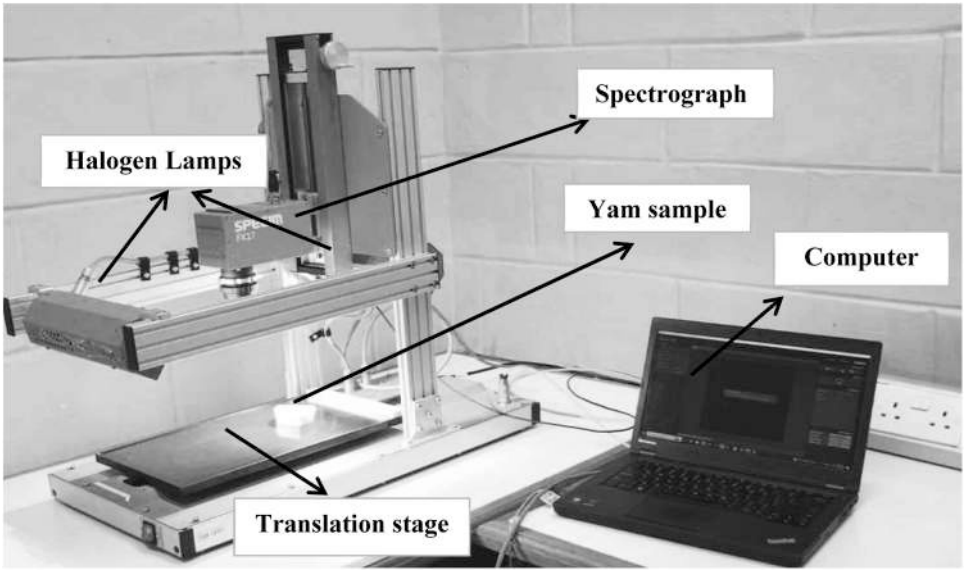

Fig. 2. Near-infrared hyperspectral imaging systems.

often contain redundant information and spectral noise, which can impact the predictive ability of models and complicate spectral data manipulation. To address this, preprocessing steps such as image segmentation and pretreatments were implemented using principal component analysis (PCA), which removed the background spectral intensities and reduced the spectral dimensions while retaining important features (Shao et al., 2022). The calibrated reflectance image (R_s) for each sample was imported to the Breeze software, and a PCA model was created to analyze the spectral variations in the sample images based on all the pixels in the image to generate the variance scatter plot (Fig. 3a). Each point in the scatter plot corresponds to pixels in the image or the background. The PCA model clustered pixels based on spectral similarities to separate the sample pixels from the background. The model showed the spectral variance between the sample and the background pixels, as shown in the maximum variance image of the samples (Fig. 3b). PC 1 explains the maximum variance (89.7 %) and visualizes the most important spectral variations in the sample image (x-axis of the scatter plot). The pixels of the sample in the scatter plot were selected to mark the region of interest, while the background pixels were eliminated (Fig. 3c).

Also, the raw image spectra contain some non-important and noisy signals, which may be caused by strayed light from the instruments or external interferences from the environment. These signals could affect the identification of relevant spectra information during the modelling

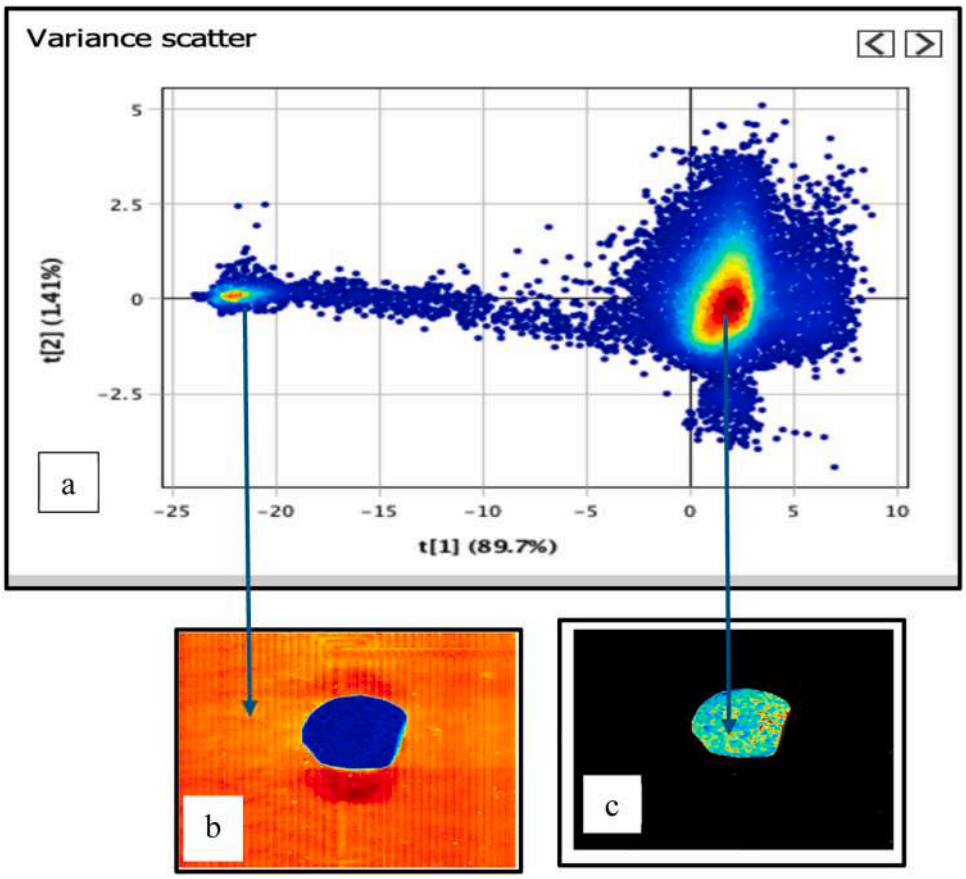

Fig. 3. PCA showing the pixels maximum variance scatter plot (a) maximum variance scatter plot of sample and background pixels (b) background and sample pixels, (c) sample pixels after removing the background pixels.

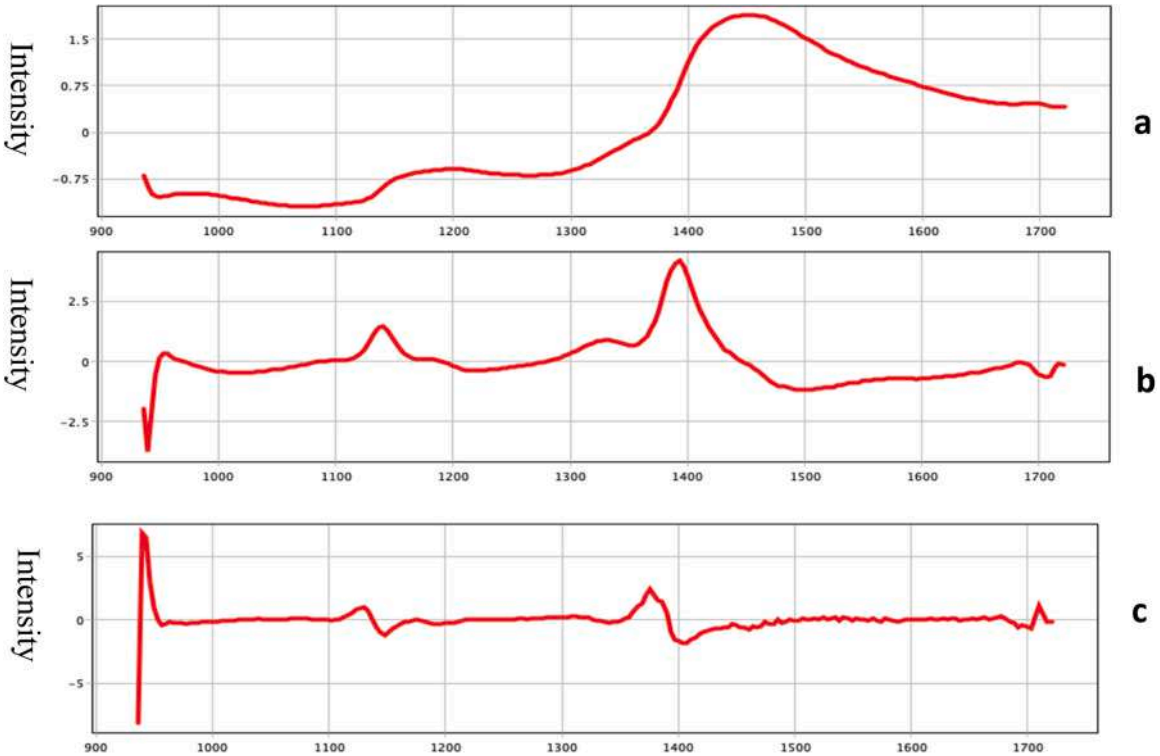

Fig. 4. Spectra pretreatments using (a) SNV, (b) first derivative and (c) second derivative.

process (Osco et al. 2022). Therefore, several preprocessing algorithms, described in the next sections, were applied to eliminate random noise and improve the signal-to-noise ratio (Xu et al., 2023).

2.4. Spectra preprocessing

The raw spectra were pre-processed to eliminate noise and correct the baseline distortions. Preprocessing the spectra before model building helped to enhance the signal-to-noise ratio, remove atmospheric interferences and improve the prediction accuracy. Hyperspectral images were subjected to different pretreatment algorithms, including Standard Normal Variates (SNV), Savitzky-Golay (SG), Multiplicative scatter corrections (MSC) and Derivatives. The Standard Normal Variance (SNV) method normalizes the spectra by subtracting each spectrum by its mean and dividing it by its standard deviation. After SNV, each spectrum will have a mean of 0 and a standard deviation of 1 (Jiao et al., 2020). SNV is useful in removing scattering effects caused by physical variations of surface imperfections (Fig. 4a). MSC created a reference spectrum by averaging multiple spectra to form a spectrum with minimal scattering effects. The reference spectrum was divided by their average spectrum to represent the chemical information in the samples accurately. SG uses a polynomial least-squares fitting approach to approximate the data points within a moving window, effectively removing high-frequency noise while preserving the underlying trends and features in the data set (Dombi et al., 2020). Important parameters for SG, such as the window size and the polynomial order, were considered during the SG pretreatments. The window size, which determines the number of neighbouring data points that fit the local polynomial, was set at 2, and the polynomial order, which determines its complexity (Zimmermann et al., 2013), was set at 4 to obtain the optimum model. The first and second derivatives (Fig. 4b-c) were used to reduce the baseline shifts, enhance peak detection and resolve superimposed peaks (Fellows et al., 2020).

2.5. Effective wavelength selection

Wavelength selection in hyperspectral data analysis is crucial to removing redundant information and noisy signals, potentially weakening a calibration model's predictive capacity. This study used Successive Progressive Algorithm (SPA), Competitive Adaptive Weighted Sampling (CARS), Boruta Algorithm (BA) and Artificial Neural Network (ANN) to select effective wavelengths from the full spectral wavelength bands using the R statistical software version (R.4.3.3). The SPA algorithms selected the wavelengths with the minor redundant information using vector projections analysis to find the variable group with minimum reducing information; this minimizes the collinearity between variables and then reduces the number of variables required for the model development. The multilinear regression model of different subsets of the wavelengths was established, and the subgroup with the minor root means square error (RMSE) was selected during SPA operations (Hu et al., 2023). The effective wavelength selection was also done using other algorithms like BA, similar to the Random Forest algorithm; it evaluates the relevance of each wavelength feature using statistical testing and shadow characteristics. Boruta created a randomized shadow feature for each original wavelength band, which was used as a baseline to compare the importance of the original wavelength bands. The BA developed a random forest model using the original wavelengths and the shadow feature as training datasets and generated importance scores for all the features. The original wavelength scores are compared with the maximum importance score of the shadow features, and the ordinal wavelength bands with a consistent importance score over its shadow feature are selected as important wavelengths. The algorithm iterates in several cycles until the selection is completed (Manikandan et al., 2024). Other feature selection approaches used were CARS and ANN algorithms. CARS aimed to enhance the efficiency of selecting important wavelengths by adaptively reweighting and competing among the different wavelengths, thus helping to select the most informative bands and improving the prediction models.

2.6. Model development and validation

The PLSR model was developed to determine yam samples' dry matter content (DMC) using full wavebands and effective wavelength selection. The calibration model was developed using the quantitative relationship between the extracted spectra and the DMC of fresh yam samples. The calibration and validation sets were chosen from the entire spectral data using the Kennard-stone algorithm, with a training-to-test set ratio of 3–1 to ensure data representativeness for modelling (Wei et al., 2024). During the development of the PLS model, the spectral data were the X variables, and the reference results for yam DMC (g/100 g) were input as the Y variables. The model's performance was evaluated using the correlation coefficient during calibration (R^2_{cal}) and validation (R^2_{pred}). Other performance criteria for the models are the root-mean-square error of prediction (RMSEP) and the ratio of standard deviation to standard error in prediction (RPD) (Maraphum et al., 2022). See Eq. 2 to 4 below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \acute{y}i)^2}{\sum_{i=1}^{n} (y_i - \ddot{y}i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\acute{y}i - \ddot{y}i)^2}{n}} \quad (3)$$

$$RPD = \frac{SD}{RMSEP} \quad (4)$$

where yi, ýi, ÿ and SD represent the measured response variable, the predicted mean, and the standard deviation, respectively.

2.7. Visualization of the distribution of DMC

The optimal PLSR model created a spatial distribution map for DMC. Each hyperspectral image contains a spectrum for every pixel, and the prediction model can determine the response attributes of each pixel in the image. The hyperspectral images were unfolded at specified wavelengths into a two-dimensional vector and then multiplied by the regression coefficient for the PLSR model. The resulting matrix was refolded into coloured images with dimensions similar to the initial image, generating a distribution map using a pseudo colour bar to represent the spatial distribution of the DMC in each yam piece.

2.8. Reference analysis

The dry matter content of fresh yam tubers was determined by the method of (Adesokan et al. 2020). Aluminium cups were washed and dried in an Oven (Memmert UN 55, GmbH) for 16 h at 103 °C until a consistent weight was reached. The cups were removed and allowed to cool for 10 minutes in a desiccator. About 10 g of the fresh homogenous yam samples were weighed (using a weighing balance calibrated to 4 decimal places) into the pre-ignited aluminium cups and placed in the oven maintained at 103 ^{0}C for 16 hrs. The cups were removed and

Table 1
Summary of dry matter content of fresh yam (g/100 g) N = 288.

		Cross-sections			
Location		Proximal	Central	Distal	
Abuja	**MIN**	28.48	28.42	22.06	(26.32)*
	MEAN	37.61	36.31	33.79	(35.90)*
	MAX	43.79	43.8	40.15	(42.58)*
	SD	4.72	4.62	4.23	(4.52)*
	CoV	12.55	12.72	12.52	
Ibadan	**MIN**	22.29	19.44	25.79	(22.51)*
	MEAN	32.99	40.14	30.70	(34.01)*
	MAX	39.41	31.91	37.25	(36.79)*
	SD	5.03	4.85	3.05	(4.31)*
	CoV	15.25	12.08	8.19	
Ubiaja	**MIN**	29.42	29.81	25.74	(28.32)*
	MEAN	38.04	36.98	37.32	(37.45)*
	MAX	46.73	44.73	46.85	(46.10)*
	SD	4.48	4.16	4.8	(4.48)*
	CoV	11.78	11.25	12.86	

SD: standard deviation, CoV: coefficient of variation; *: indicate the overall average DMC for prox,central and distal for each tuber

transferred immediately into a desiccator to cool. The loss in weight after drying was estimated as the sample's moisture content, and the DM (g/100 g) was computed as (100 - Moisture content).

$$Moisture\ content = \frac{weight\ of\ sample\ before\ drying - weight\ of\ sample\ after\ drying}{weight\ of\ sample\ beforedrying} \quad (5)$$

$$DM(g/100g) = (100 - Moisture\ content)$$

3. Results and discussion

3.1. Descriptive analysis of dry matter content (g/100 g)

The dry matter content (DMC) of the fresh yam genotypes used for the study is shown in Table 1. Overall, DMC ranged from 26.30 to 42.60 g/100 g, with a mean±SD of 35.90±4.52 for the Abuja location, while for the Ibadan location, DMC ranged from 22.51 to 36.79 g/100 g with a mean±SD of 34.01±4.31. Ubiaja location had DMC ranging from 28.30 to 46.10 g/100 g and mean±SD of 37.45± 4.48, respectively. The average DMC of samples from the Ubiaja location is consistently higher than the two other locations, which could be due to differences in the agroecological conditions of the three locations, such as annual rainfall patterns, temperature, and soil nutrients. A dry matter content of 36.2 ±2.4 g/100 g was reported for Dioscorea rotundata by Matsumoto et al. (2021), which was within the range of values obtained in this study. The results also showed that the proximal part of the yam cross-sections has the highest DMC than central and distal regions, except at Ubiaja, where the distal section has an overall average of more than the proximal and central (Table 2). These results were consistent with the findings of Hamadina and Asiedu (2015), who discovered that the yam head area (28.6 g/100 g) had a higher dry matter content than the middle section (26.9 g/100 g), while the tail had the lowest value (22.30 g/100 g). However, the values reported in their study were lower than the current work, which could be related to different yam genotypes, as this study used advanced breeding lines from IITA germplasm. Mestres et al. (2023) also reported that dry matter and starch content vary significantly between various sections of boiled yam tubers. The dry matter content of RTB crops can affect product yield and influence the storage stability of tuber crops such as yam.

Table 2
Analysis of variance of DM (g/100 g) in yam genotypes across locations.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Genotypes	13	1382.90	106.38	16.05	<.0001
Location	2	1718.61	859.30	129.64	<.0001
Cross-sections	2	192.89	96.45	14.55	<.0001
Genotypes*cross-section	26	293.29	11.28	1.70	0.02

P<0.001 shows significant at 95 % confidence interval.

The analysis of variance (Table 2) indicates that the location, cross-section, and genotype significantly impact the dry matter content of yam tubers. This suggests diversity in the dry matter content among the selected yam genotypes. Furthermore, the significant effect ($p<0.001$) of the cross-section (proximal, central, and distal) suggests variations in the distribution of dry matter content across the sections of a yam tuber, which agrees with the findings of Cheng et al. (2023).

3.2. Spectral characteristics

The raw spectra for all the samples (n = 288) were extracted in the 932.61–1720.20 nm wavelength range, as shown in Fig. 5. Notably, a slight increase in reflectance values was observed between 980 and 1001.40 nm and a prominent reflectance increase at 1400.20 and 1470.00 nm. The bands at 980 and 1470 nm are related to the first and second stretching overtones of O-H bonds, which may be attributed to the moisture in the samples (Yang et al., 2021; Jiyang et al., 2022; Ma and Sun, 2020). Also, the absorption peak at 1213.90 nm corresponds to the second overtone of the C-H bond and is attributed to the carbohydrate (Yang et al., 2021; Ding and Xu, 2000; Zhang et al., 2020). All the raw spectra for the yam samples exhibited similar patterns based on the genotypes and planting location; some noise was observed between 932.61 and 977.11 and 1673.70–1720.20 nm, respectively. These spectrum regions were eliminated to improve the quality of the spectra data.

3.3. Model development and validation

3.3.1. PLSR models for DMC (g/100 g)

A PLSR model combined with various machine learning algorithms for spectra pre-processing was implemented on the full spectra data and selected wavelength bands to establish quantitative models for DMC by establishing a correlation between the spectra reflectance from the image and the laboratory reference from each section of the yam samples. PLSR is one of the most commonly used linear regression algorithms and an optimal option for developing a prediction model because it could incorporate both the spectra data matrix (x) and the variable matrix (y). PLSR can also resolve the collinearity problem in large data sets (Ismy et al., 2023). A total of 288 spectra data points were generated from the samples, and during the modelling, Kennard-stone's algorithm was used to divide the samples into 192 training and 96 test sets, where the test set was one-third of the total spectra. Pre-processing methods include standard normal variate, Satvizsky-Golay smoothing and derivative. After removing the wavelengths with noisy signals, the PLSR model was developed in the wavelength range of 949.43–1695.20 nm. Cross-validation was performed using the leave-one-out approach and implemented in 7 groups (Figs. 6 and 7).

Table 3 shows the PLSR models for quantification of DMC (g/100 g)

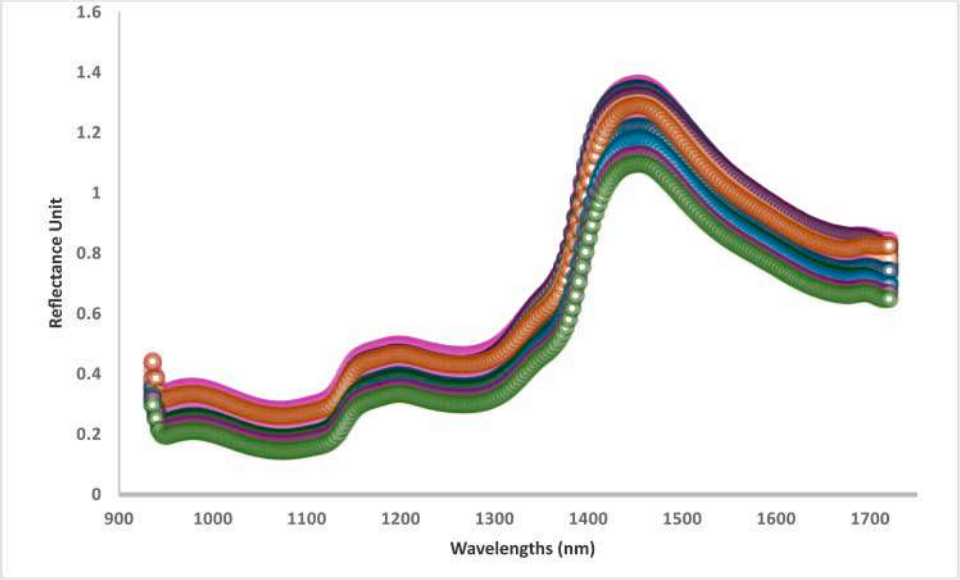

Fig. 5. Raw spectra of yam samples.

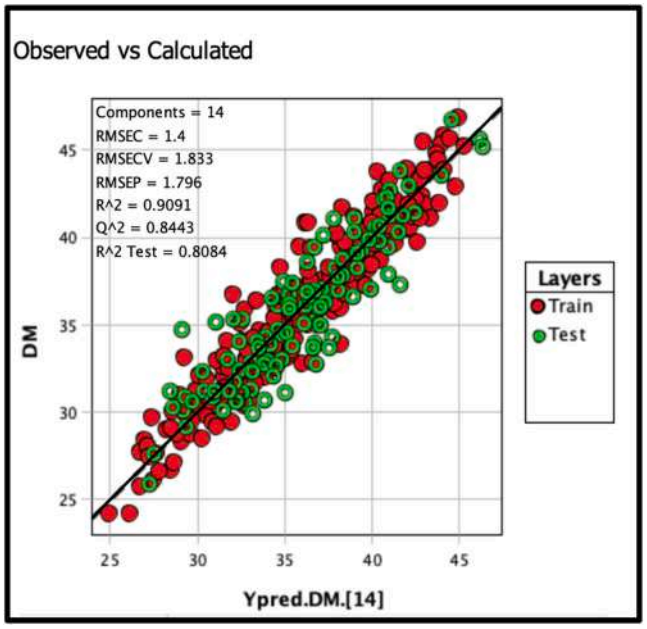

Fig. 6. PLSR model for dry matter content of fresh yam with SNV pretreatment.

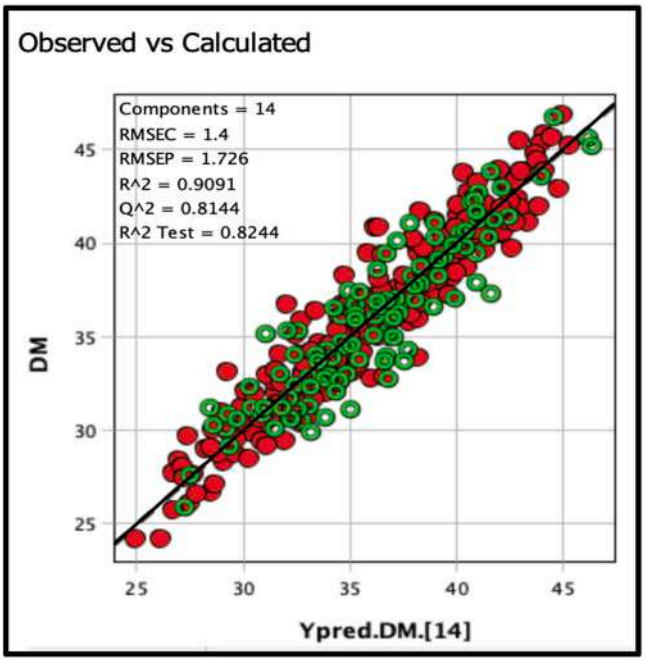

Fig. 7. PLSR model for dry matter content of fresh yam with SNV+SG + first derivative pretreatments.

in fresh yam samples using different combinations of preprocessing algorithms. The model developed with PLSR combined with SNV + Savitzky Golay (polynomial (4), left window (2), and right window (2) and first derivatives demonstrated the best prediction performances with a coefficient of determination in calibration (R^2 cal) and prediction (R^2 pred) of 0.90 and 0.83 respectively. The model processed with SNV + SG + 2nd derivatives had a higher R^2 cal of 0.92 but slightly lower R^2 pre (0.82), and the standard error of prediction is greater than the model processed using the 1st derivative by 0.04 Several authors have routinely used PLSR to develop models from hyperspectral data (Jiang, 2017; Yu et al. 2023; Zhang et al. 2023a). In their study to predict the moisture content of steamed and dried purple potatoes, Heo et al. (2021) also reported the highest accuracy from a model developed using PLSR combined with SG and the first derivative preprocessing with R^2 pre of 0.96 for quantification of moisture content of steamed and dried potatoes. The R^2 cal is higher than the value obtained in this current study but has a similar mean square prediction error.

Hyperspectral image spectra data always contain reductant information and collinearity problems, which affect the model's accuracy and robustness (He et al., 2023). Therefore, SPA, CARS, BA, and ANN were implemented to extract important wavelengths from the raw spectra information. Figs. 8, 9, 10 and 11 indicate the selected SPA, BA, CARS and ANN wavelengths, respectively. The green region in Fig. 9 indicates the selected wavelengths for the BA, while the red region shows the non-informative or rejected wavelengths. Table 3 shows the optimized PLSR models using the selected wavelengths. The prediction accuracy for DMC was improved for all the wavelength selection methods; PLSR combined individually with SPA, CARS, BA and ANN gave R^2 pre of 0.930, 0.973, 0.910, 0.940 and 0.943, respectively. Furthermore, PLSR combined with multiple algorithms such as

Table 3
Multivariate Linear Model for DMC (g/100 g).

Model	Spectra Pretreatments	(Training = 193, Test =96)					
		SEL: 0.250					
PLSR	**Full wavelengths**	**R^2 cal**	**R^2 CV**	**R^2 pred**	**RMSEC**	**RMSEP**	**RPD**
	No treatments	0.890	0.850	0.820	1.480	1.740	3.015
	Standard Normal Variate only	0.900	0.840	0.800	1.400	1.790	3.162
	Standard Normal Variate + SG	0.900	0.840	0.820	1.400	1.830	3.160
	Standard Normal Variate + $SG^{p:4,lw:2,rw:2}$ + 1st derivative	**0.900**	**0.810**	**0.830**	**1.400**	**1.690**	**3.162**
	Standard Normal Variate + $SG^{p:4,lw:2,rw:2}$ + 2nd derivative	0.920	0.800	0.820	1.290	1.730	3.120
PLSR	**Number of wavelengths**						
SPA	30	0.952	0.680	0.930	1.057	1.330	3.780
BA	36	0.960	0.850	0.940	1.144	1.830	4.080
CARS	20	0.934	0.750	0.900	1.264	1.640	3.162
ANN	20	0.932	0.685	0.943	1.224	1.730	4.189
SPA-BA	66	0.940	0.880	0.850	1.070	1.400	2.580
SPA-ANN	66	0.958	0.775	0.949	0.904	1.021	4.428
SPA-CARS	50	**0.974**	**0.610**	**0.958**	**0.773**	**0.898**	**4.880**

R^2_{cal}: coefficient of determination in calibration **R^2_{CV}**: coefficient of determination in prediction; RMSEC: roots mean square error in calibration; RMSEP: root mean square error in prediction; **RPD**: ratio of prediction to deviation; **p**: polynomial, lw: left window, **rw**: right window; PLSR = partial least square regression; SPA: successive wavelength selection,.Boruta machine learning algorithms, ANN: artificial neural network; CARS: competitive average progressive reweighed sampling, **SEL**: standard error of laboratory.

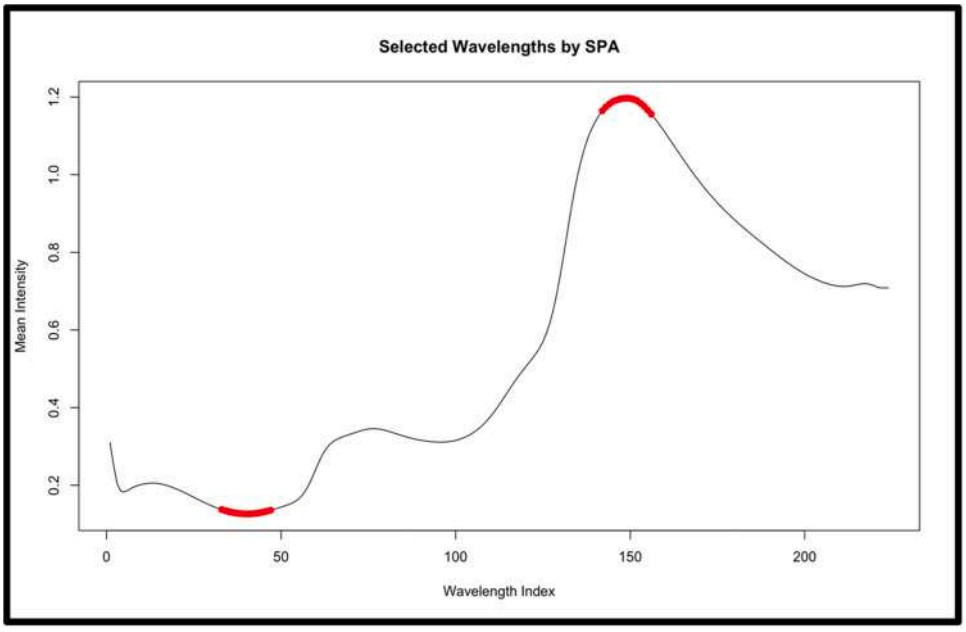

Fig. 8. Selected wavelength using Successive Progressive Algorithm (SPA).

PLSR-SPA-BA, PLSR-SPA-ANN and PLSR- SPA-CARS had R^2 pre ranged from 0.850 to 0.966 with the PLSR-SPA-CARS giving the best model where R^2cal, R^2 pred, and RMSEP were 0.974, 0.958, and 0.898 g/100 g, respectively (Table 3, Fig. 12). The optimum model was above the R^2_{pred} of 0.94 obtained by Meghar et al. (2023) for DMC on fresh, intact cassava roots using CovSel multiple linear regression (Cov Sel_MLR).

Similarly, Wang et al. (2021) investigated the prediction performance of the PLSR model for potato starch quantification in three regions, namely the middle head and umbilicus, using the hyperspectral camera. These authors also combined competitive adaptive reweighted sampling (CARS) and support vector regression (SVR) to develop a model for each region. At the same time, the umbilicus gave the best performance, with correlations R^2 cal and R^2 pre of 0.94 and 0.93, respectively, and RMSEP of 17.4 g/kg. The correlation coefficient in calibration for the PLSR-SPA-CARS model in this current study performed better than the values reported in their findings.

The PLSR model has been combined with hyperspectral data for other cereals and legumes; for instance, Qiao et al. (2022) utilized hyperspectral imaging spectroscopy to determine maize hardness. They developed a PLSR model after selecting characteristic wavelengths using the successive projection algorithm (SPA), achieving R^2 cal and RMSE values of 0.912 and 17.76 MPa, respectively. Additionally, the moisture content (MC) of maize seeds was predicted using long-wave near-infrared hyperspectral imaging technology. Wang et al. (2021) achieved optimal prediction performance of MC with the CARS-SPA-LS-SVM model, reporting R^2 pre and RMSEP values of 0.93 and 0.00094 %, respectively. NIR-HSI was investigated to quantify protein, starch, and moisture content using 77 wheat varieties converted into flour. Effective wavelength algorithms were combined with the hyperspectral data to develop a prediction model, the coefficient of determination and root mean square error for prediction of 0.98 and 1.15 g/100 g for protein, 0.9243 and 0.2068 g/100 g for starch, and 0.8646 and 2.1669 g/100 g for moisture was obtained. The performance of a prediction model is appraised using some important metrics, including the coefficient of determination in prediction and the root mean square error, which is the

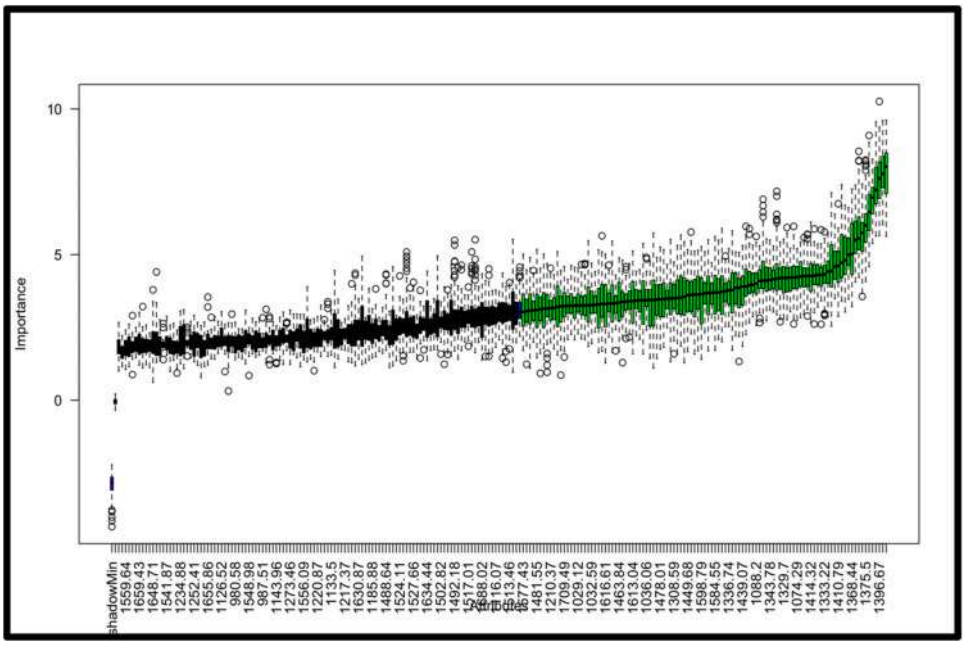

Fig. 9. Selected wavelength using Boruta Algorithm (BA).

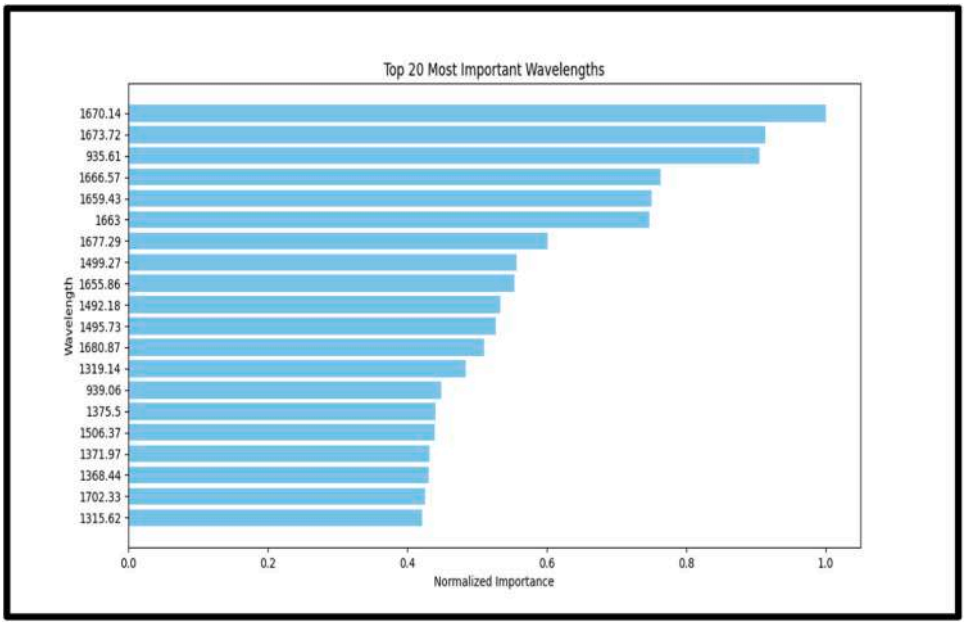

Fig. 10. Selected wavelength using Competitive Adaptive Reweighted Sampling (CARS).

measure of the difference between the predicted values and the observed values from the standard method. Generally, an accurate model should have R^2 cal and R^2 pre values closer to 1 and lower RMSEP and RMSEC. Also, RPD values must be > 2 for a model to be considered suitable for quantitative purposes (Cortes et al. 2019). The PLSR model developed in this study showed robust and reliable prediction performances and could be useful for routine analysis, especially when the PLSR model is built on the important wavelengths in the spectra bands, giving a better prediction capability.

3.3.2. *Prediction using machine learning models for DMC (g/100 g)*

Table 4 shows the results of prediction models built using other machine learning algorithms, including linear regression (LR), random forest (RF), decision tree (DT), and support vector machine (SVM). The best model for DMC using a machine learning algorithm was obtained with linear regression with an R^2 of 0.808, a mean absolute error of 1.590 and an RMSE of 2.326. Decision trees and Random Forests have a similar mean absolute error, but R^2 of 0.733 and 0.719, respectively, which are also close. SVM also had an R^2 of 0.745 and a very low MAE of 0.231. Generally, most algorithms demonstrated a good predictive capacity for DMC with a relatively low mean absolute error, indicating the prediction error. However, the PLS models built on the selected important wavelength had better prediction performance than those developed using the other machine learning algorithms. This could be due to the nature of the data, as most machine learning algorithms work best with non-linear data. Nantongio et al. (2024) also reported that iPLS performed better than RF and XGBOOST in the prediction of sensory attributes of sweet potatoes. CARS was also combined with PLSR to

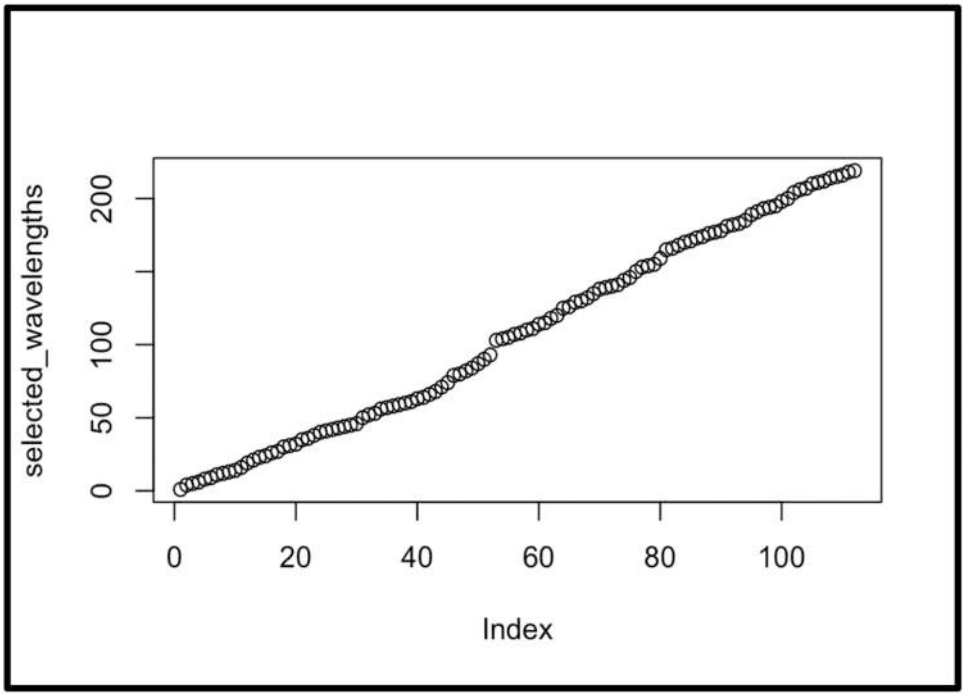

Fig. 11. Selected wavelength using Artificial Neural Network (ANN).

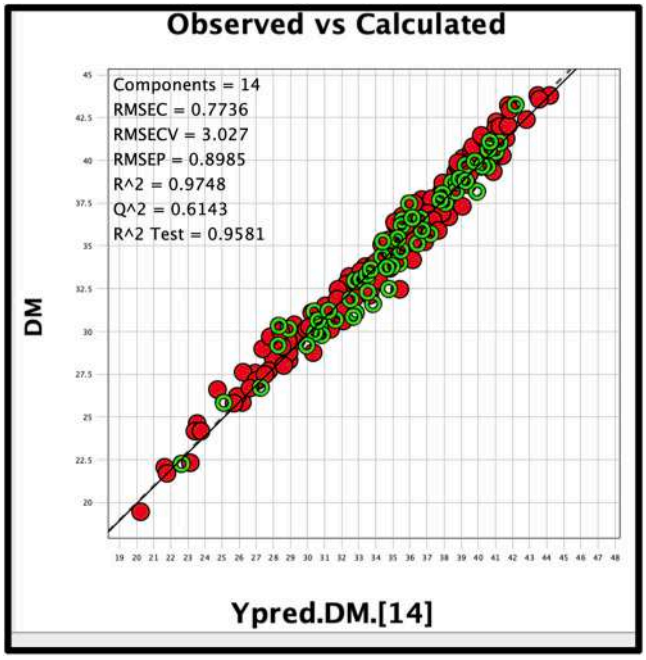

Fig. 12. PLSR model for dry matter content of fresh yam Based on wavelength selection using SPA-CARS algorithms.

predict DMC in potatoes, and the authors reported an R^2 of 0.968 and an RMSE of 0.413 %. (Wang et al., 2022). PLSR combined with SPA-CARS in this current study also demonstrated the best prediction model for DMC in fresh, intact yam tubers.

3.4. Visualization of spatial distribution of DMC(g/100 g) in fresh yam tubers

The distribution of dry matter content in fresh, intact yam was analyzed using a PLSR-SPA-CARS model. Each point on the yam samples corresponds to a pixel in the hyperspectral image and has a unique spectral intensity. Essential wavelengths were determined from the pseudo-RGB image of the sample and used to predict the dry matter content, resulting in a dry matter distribution map (Fig. 13). The distribution map clearly shows the variation in dry matter content across different parts of the fresh yam tuber. The colours on the map indicate the variance in dry matter content; blue represents low dry matter, while red represents high dry matter content. The chemical imaging map displays the variation in dry matter distribution in the yam tuber's proximal, middle, and distal regions. The proximal region exhibits higher dry matter content than the middle and distal regions of the yam cross-section.

Table 4
Results of prediction models developed using machine learning models.

Target traits	Machine learning model	R^2	MAE	RMSE
DMC	**LR**	**0.808**	**1.59**	**2.326**
	RF	0.733	1.976	2.746
	DT	0.719	1.91	2.818
	SVM Linear 2	0.745	2.234	2.368

LR: Linear Regression, RF: Random Forest, DT: Decision Tree, SVM Linear 2: Support Vector Machine Learning, R^2: coefficient of determination; MAE: Mean absolute error, RMSE: Root mean square error.

4. Conclusion

The study used the PLSR model and various machine learning algorithms to process spectra data and develop a model to quantify fresh yam tubers' dry matter content (DMC) for routine analysis. The PLSR-SPA-CARS model showed the most accurate prediction performance by effectively selecting wavelengths compared to the model using the full spectra wavebands. The PLSR-SPA-CARS model was enhanced with SNV-SG to develop an optimal model with R^2 cal and R^2 pre values of 0.974 and 0.958, respectively. The root means square error of prediction was low at 0.898 g/100 g, indicating the model's robustness and a high RPD of 4.88. The high coefficient of prediction suggests that hyperspectral imaging is a valuable tool for rapid analysis of the dry matter

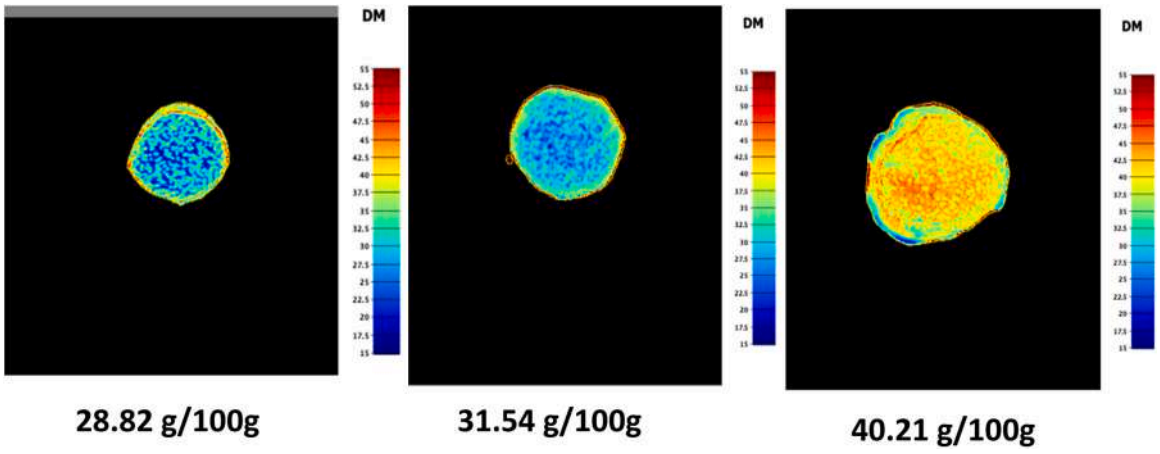

Fig. 13. Spatial distribution of DMC (g/100 g) within the cross-section of yam tuber (a) distal, (b) central and (c) proximal.

content of fresh, intact yam tubers, which is useful for yam breeders and food scientists. The study also found that the proximal section of yam tubers has the highest DMC, with an overall average of 36.21 g/100 g, followed by the central section, with an average DMC of 35.07 g/100 g. The distal region has the lowest value (33.90 g/100 g), indicating that evaluating traits across the different cross-sections of the yam is important for accurate prediction. A chemical distribution map was developed in the study to visualize the spatial distribution at various points and sections of the fresh yam tubers, which will be helpful for quality control and monitoring of nutrient transfer during storage. The results of this show that NIR hyperspectral imaging can accurately assess the DMC of fresh yam tubers and has provided low-cost and rapid tools for phenotyping dry matter content without destroying the samples, allowing for prompt decisions on the quality characteristics of large yam germplasm for breeding programs.

Funding

The authors acknowledge funding from the Bill and Melinda Gates Foundation (BMGF) for the International Institute of Tropical Agriculture's Africa Yam Project (INV-003446) and also the funding supports from the RTBfoods project.

CRediT authorship contribution statement

Bolanle Otegbayo: Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Emmanuel Oladeji Alamu:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Michael Olutoyin Afolabi:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Busie Maziya-Dixon:** Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Michael Adesokan:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors hereby declare no conflicts of interest in the publication of this manuscript.

Data Availability

Data will be made available on request.

Acknowledgements

The authors acknowledge the support of Africa Yam Project leaders, especially Dr Asrat Amele and Dr Patrick Adebola, the RTBfoods project and the project team, especially Dr Dominique Duffour and Karimah Megahr, and the staff of the Food and Nutrition Sciences Laboratory and Yam Breeding Unit of IITA (especially Mr. Alex Edemodu).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jfca.2024.106692.

References

Adesokan M., Alamu E. and Maziya-Dixon B., SOP for Determination of Dry Matter Content. RTBfoods Project Report, Ibadan, Nigeria, p. 7 (2020) ⟨https://mel.cgiar.org/reporting/download/report_file_id/17813⟩. (Accessed on February 16, 2024).

Adesokan, M., Alamu, E.O., Otegbayo, B., Maziya-Dixon, B., 2023. A review of the use of Near-Infrared Hyperspectral Imaging (NIR-HSI) techniques for the non-destructive quality assessment of root and tuber crops. Appl. Sci. *13* (9), 5226. https://doi.org/10.3390/app13095226.

Adinsi, L., Honfozo, F.L., & Akissoé, N. (2021). Sample preparation and cooking time for texture analysis of boiled yam. Biophysical characterization of quality traits, ⟨https://mel.cgiar.org/reporting/download/report_file_id/25510⟩ (Accessed on February 16 2024).

Alamu, E.O., Nuwamanya, E., Cornet, D., Meghar, K., Adesokan, M., Tran, T., Davrieux, F., 2021. Near-infrared spectroscopy applications for high-throughput phenotyping for cassava and yam: a review. Int. J. Food Sci. Technol. 56 (3), 1491–1501. https://doi.org/10.1111/ijfs.14773.

Cheng, S., Liu, Y., Su, L., Liu, X., Chu, Q., He, Z., Zheng, W., 2023. Physiological, anatomical and quality indexes of root tuber formation and development in chayote (Sechium edule. BMC Plant Biol. *23* (1), 413. https://doi.org/10.1186/s12870-023-04427-0.

ElMasry, G.M., Nakauchi, S., 2016. Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality–a comprehensive review. Biosyst. Eng. *142*, 53–82. https://doi.org/10.1016/j.biosystemseng.2015.11.009.

FAOSTAT. (2021). Food and Agriculture Organization of the United Nations. FAOSTAT Statistical Database. Rome. Available at: ⟨https://fao.org/faostat/en/#compare⟩ (assessed March 6, 2024).

Fellows, A.P., Casford, M.T., Davies, P.B., 2020. Spectral analysis and deconvolution of the amide I band of proteins presenting with high-frequency noise and baseline shifts. Appl. Spectrosc. *74* (5), 597–615. https://doi.org/10.1177/0003702819898536.

Hamadina, E.I., Asiedu, R., 2015. Dry matter, free sugar and starch changes in tuber regions of white yam (Dioscorea rotundata Poir.), and the effect of storage environment. J. Adv. Agric. *4*, 303-09.

He, H.J., Wang, Y., Wang, Y., Liu, H., Zhang, M., Ou, X., 2023. Simultaneous quantifying and visualizing moisture, ash and protein distribution in sweet potato [Ipomoea batatas (L.) Lam] by NIR hyperspectral imaging. Food Chemistry:X *18*, 100631. https://doi.org/10.1016/j.fochx.2023.100631.

Heo, S., Choi, J.Y., Kim, J., Moon, K.D., 2021. Prediction of moisture content in steamed and dried purple sweet potato using hyperspectral imaging analysis. Food Sci. Biotechnol. *30*, 783–791. https://doi.org/10.1007/s10068-021-00921-z.

Hu, H., Wang, T., Wei, Y., Xu, Z., Cao, S., Fu, L., Huang, L., 2023. Non-destructive prediction of isoflavone and starch by hyperspectral imaging and deep learning in

Puerariae Thomsonii Radix. Front. Plant Sci. *14*, 1271320 https://doi.org/10.3389/fpls.2023.1271320.

Ismy, A.S., Siahaan, H.H., Sitorus, A., 2023. A novel strategy of NIR spectra multivariate calibration in the presence both of small dataset and non-linearity: a comparative study. Case Stud. Chem. Environ. Eng. *8*, 100384 https://doi.org/10.1016/j.cscee.2023.100384.

Jiang, W., 2017. Application of Hyperspectral Technology in Potato Variety Identification and Quality Non-destructive Testing. Northeast Agricultural University, Ph.D. thesis, China.

Luo, S., He, Y., Li, Q., Jiao, W., Zhu, Y., Zhao, X., 2020. Non-destructive estimation of potato yield using relative variables derived from multi-period LAI and hyperspectral data based on weighted growth stage. Plant Methods *16*, 150. https://doi.org/10.1186/s13007-020-00693-3.

Ma, J., Sun, D.W., 2020. Prediction of monounsaturated and polyunsaturated fatty acids of various processed pork meats using improved hyperspectral imaging technique. Food Chem. *321*, 126695 https://doi.org/10.1016/j.foodchem.2020.126695.

Manikandan, G., Pragadeesh, B., Manojkumar, V., Karthikeyan, A.L., Manikandan, R., Gandomi, A.H., 2024. Classification models combined with Boruta feature selection for heart disease prediction. Inform. Med. Unlocked *44*, 101442. https://doi.org/10.1016/j.imu.2023.101442.

Maraphum, K., Saengprachatanarug, K., Wongpichet, S., Phuphuphud, A., Posom, J., 2022. Achieving robustness across different ages and cultivars for an NIRS-PLSR model of fresh cassava root starch and dry matter content. Comput. Electron. Agric. *196*, 106872 https://doi.org/10.1016/j.compag.2022.106872.

Matsumoto, R., Asfaw, A., De Koeyer, D., Muranaka, S., Yoshihashi, T., Ishikawa, H., Asiedu, R., 2021. Variation in tuber dry matter content and starch pasting properties of white Guinea yam (Dioscorea rotundata) genotypes grown in three agroecologies of NIGERIA. Agronomy *11* (10), 1944. https://doi.org/10.3390/agronomy11101944.

Meghar, K., Boyer, J., Davrieux, F., 2022. Prediction of yam cooking behaviour using hyperspectral imaging. Rep. HSI calibrations Dry. Matter, pectin, starch Texture raw fresh yam slices CIRAD Fr. https://doi.org/10.18167/agritrop/00707.

Meghar, K., Tran, T., Delgado, L.F., Ospina, M.A., Moreno, J.L., Luna, J., Davrieux, F., 2023. Hyperspectral imaging for the determination of relevant cooking quality traits of boiled cassava. J. Sci. Food Agric. https://doi.org/10.1002/jsfa.12654.

Mestres, C., Taylor, M., McDougall, G., Arufe, S., Tran, T., Nuwamanya, E., Rolland-Sabate, A., 2023. Contrasting effects of polysaccharide components on the cooking properties of roots, tubers and bananas. J. Sci. Food Agric. https://doi.org/10.1002/jsfa.12914.

Osco, L.P., Furuya, D.E.G., Furuya, M.T.G., Corrêa, D.V., Gonçalvez, W.N., Junior, J.M., de Castro Jorge, L.A., 2022. An impact analysis of pre-processing techniques in spectroscopy data to classify insect-damaged in soybean plants with machine and deep learning methods. Infrared Phys. Technol. *123*, 104203 https://doi.org/10.1016/j.infrared.2022.104203.

Pan, L., Lu, R., Zhu, Q., McGrath, J.M., Tu, K., 2015. Measurement of moisture, soluble solids, sucrose content and mechanical properties in sugar beet using portable visible and near-infrared spectroscopy. Postharvest Biol. Technol. *102*, 42–50. https://doi.org/10.1016/j.postharvbio.2015.02.005.

Peng, W., Beggio, G., Pivato, A., Zhang, H., Lü, F., He, P., 2022. Applications of near infrared spectroscopy and hyperspectral imaging techniques in anaerobic digestion of bio-wastes: A review. Renew. Sustain. Energy *Rev. 165*, 112608 https://doi.org/10.1016/j.rser.2022.112608.

Qiao, M., Xu, Y., Xia, G., Su, Y., Lu, B., Gao, X., Fan, H., 2022. Determination of hardness for maize kernels based on hyperspectral imaging. Food Chem. *366*, 130559 https://doi.org/10.1016/j.foodchem.2021.130559.

Ray, P., Reddy, S.S., Banerjee, T., 2021. Various dimension reduction techniques for high dimensional data analysis: a review. Artif. Intell. *Rev. 54* (5), 3473–3515. https://doi.org/10.1007/s10462-020-09928-0.

Shao, Y., Liu, Y., Xuan, G., Wang, Y., Gao, Z., Hu, Z., Wang, K., 2020. Application of hyperspectral imaging for spatial prediction of soluble solid content in sweet potato. RSC Adv. *10* (55), 33148–33154. https://doi.org/10.1039/C9RA10630H.

Su, W.H., Bakalis, S., Sun, D.W., 2020. Chemometric determination of time series moisture in both potato and sweet potato tubers during hot air and microwave drying using near/mid-infrared (NIR/MIR) hyperspectral techniques. Dry. Technol. *38*, 806–823. https://doi.org/10.1080/07373937.2019.1593192.

Tian, X.Y., Aheto, J.H., Bai, J.W., Dai, C., Ren, Y., Chang, X., 2021. Quantitative analysis and visualization of moisture and anthocyanins content in purple sweet potato by Vis–NIR hyperspectral imaging. J. Food Process. Preserv. *45* (2), e15128 https://doi.org/10.1111/jfpp.15128.

Wang, F., Wang, C., Song, S., Xie, S., Kang, F., 2021. Study on starch content detection and visualization of potato based on hyperspectral imaging. Food Sci. Nutr. *9* (8), 4420–4430. https://doi.org/10.1002/fsn3.2415.

Wei, X., Deng, C., Fang, W., Xie, C., Liu, S., Lu, M., Wang, Y., 2024. Classification method for folded flue-cured tobacco based on hyperspectral imaging and conventional neural networks. Ind. Crops Prod. *212*, 118279 https://doi.org/10.1016/j.indcrop.2024.118279.

Xu, S., Lu, H., Liang, X., Ference, C., Qiu, G., Fan, C., 2023. Modelling and De-Noising for Nondestructive Detection of Total Soluble Solid Content of Pomelo by Using Visible/Near Infrared Spectroscopy. Foods *12* (15), 2966. https://doi.org/10.3390/foods12152966.

Zhang, J., Guo, Z., Ren, Z., Wang, S., Yin, X., Zhang, D., Ma, C., 2023a. A non-destructive determination of protein content in potato flour noodles using near-infrared hyperspectral imaging technology. Infrared Phys. Technol. *130*, 104595 https://doi.org/10.1016/j.infrared.2023.104595.

Zhang, H., Zhang, S., Chen, Y., Luo, W., Huang, Y., Tao, D., Liu, X., 2020. Non-destructive determination of fat and moisture contents in Salmon (Salmo salar) fillets using near-infrared hyperspectral imaging coupled with spectral and textural features. J. Food Compos. Anal. *92*, 103567 https://doi.org/10.1016/j.jfca.2020.103567.