Article

Machine learning-powered models for near-infrared spectrometers: prediction of protein in multiple grain cereals

Keerthi Chadalavada[1,6], Krithika Anbazhagan[1], Adama Ndour[2], Sunita Choudhary[1], William M. Palmer[3], Jamie R. Flynn[3], Srikanth Mallayee[1], P. Sharada[4], K.V.S.V. Prasad[4], V. Padmakumar[4], Chris Jones[5], Jana Kholová[1,7*]

[1] International Crops Research Institute for Semi-Arid Tropics, Patancheru 502 324, Telangana, India; keerthichadalwada@gmail.com (K.C.); a.krithika@cgiar.org (K.A); s.choudhary@cgiar.org (S.C.); srikanthmallayee@gmail.com (S.M.); j.kholova@cgiar.org (J.K.)
[2] International Crops Research Institute for Semi-Arid Tropics, Bamako BP 320, Mali; a.ndour@cgiar.org (A.N.)
[3] Hone, Suite 65 Level 1, 113-145 Hunter St, Newcastle, 2300, Australia; william@honeag.com (W.P.); jamie@honeag.com (J.F.)
[4] International Livestock Research Institute, Patancheru 502 324, Telangana, India; p.sharada@cgiar.org (P.S.); k.v.prasad@cgiar.org (V.P.); v.padmakumar@cgiar.org (P.K.)
[5] International Livestock Research Institute, Addis Ababa P.O. Box 5689, Ethiopia; c.s.jones@cgiar.org (C.J.)
[6] Bharathidasan University, Palkalaiperur, Tiruchirappalli, 620 024, Tamil Nadu, India
[7] Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, Prague 165 00, Czech Republic; j.kholova@cgiar.org (J.K.)

* Correspondence: j.kholova@cgiar.org (J.K.)

Abstract: Achieving global goals on sustainable nutrition, health, and wellbeing will depend on delivering enhanced diets to humankind. This will require, among others, instantaneous access to information on food quality at key points within agri-food systems. Although stationary methods are usually used to quantify grain quality (wet-lab chemistry, benchtop NIR spectrometer); these do not suit many required user-cases, such as stakeholders in decentralized agri-food-chains that are typical for emerging economies. Therefore, we explored new technologies and models that might aid these particular user-cases. For this purpose, we generated the NIR spectra of 328 grain samples from multiple cereals (finger millet, foxtail millet, maize, pearl millet, sorghum) with a standard benchtop NIR Spectrometer (DS2500, FOSS) and a novel mobile NIR-based sensor (HL-EVT5, Hone). We explored a range of classical deterministic and novel machine learning (ML)-driven models to build calibrations out of the NIR spectra. We were able to build relevant calibrations out of both types of spectra. At the same time, ML-based methods enhanced the prediction capacity of calibration models compared to classical deterministic methods. We also documented that the prediction of grain protein content based on NIR spectra generated by a mobile sensor (HL-EVT5, Hone) was highly relevant for quantitative protein predictions ($R^2 = 0.91$, RMSE = 0.97, RPD = 3.48). Thus, the findings of this study lay the foundations on which to expand the utilization of NIR spectroscopy applications for agricultural research and development.

Keywords: Cereals; Grain protein; Near Infrared Spectroscopy (NIRS)-based sensors; Prediction algorithms; FOSS; Hone Lab

1. Introduction

Near-infrared spectroscopy (NIRS) is a non-destructive method widely used to predict the organic compounds of grain materials based on electromagnetic wave interactions. The technology offers time- and cost-effective access to grain quality parameters [1–3]. While several companies offer the standard benchtop NIR spectrometers (such as the FOSS-DS2500 flour analyzer [4], Bruker's Tango FT-NIR spectrometer [5], Perten-IM9520 [6]); in the last decade, the market has offered several options for handheld portable NIR instruments as well [7,8]. For example, MicroNIR OnSite-W from VIAVI Solutions [9], DLP NIRScan™ Nano EVM spectrophotometer from Texas Instruments' DMD™ [10], MEMS spectrometer from Fraunhofer [11], and Hone Lab Red from Hone [12]. While many benchtop NIRs are already used for some applications across the agri-food sector, handheld instruments are not regularly used [13–15]. This is possibly because accurate NIRS-based predictions not only require quality instrumentation to generate spectra but also good quality predictive algorithms, which can be problematic for some of the portable NIRs instruments [16–21].

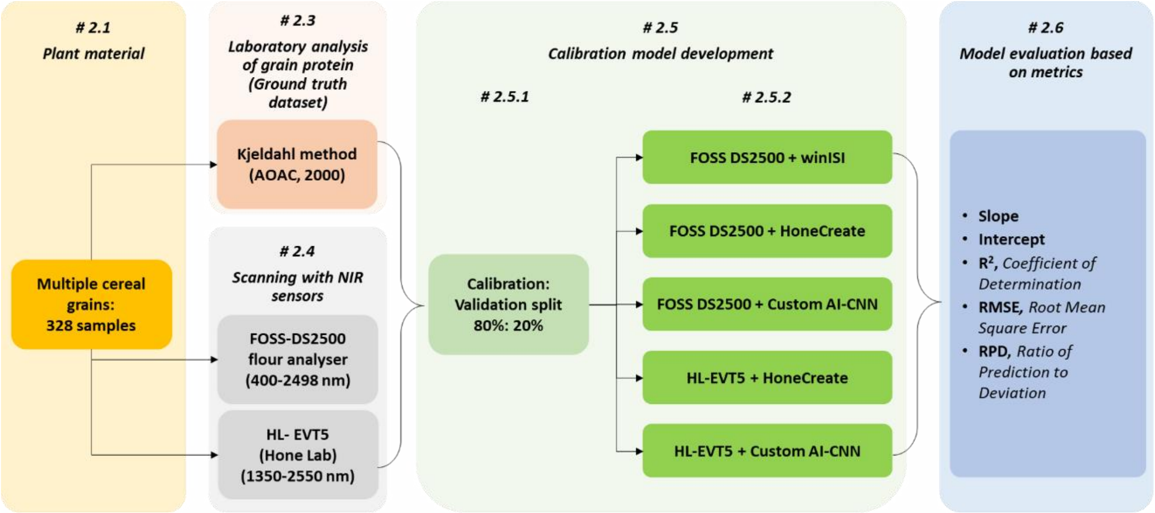
There are a range of statistical methods for the treatment of acquired NIR spectra and predicting the material composition such as principal component analysis (PCA), partial least squares regression (PLSR), and multiple linear regression (MLR) [22–30]. More recently, machine learning (ML)-based methods—such as random forest, support vector machines regression, artificial neural networks, and convolutional neural networks have been successfully used to build robust calibration models [31–38]. The ML-based methods, in particular, are gaining a lot of attention as these may offer specific advantages for applications where the feature prediction is required from imperfect spectra or spectra derived from a range of materials. This is because ML-algorithms have the enhanced capacity to identify common patterns in diverse data information from which the required algorithms are built [35].

For instance, in the cereal-based food and feed industry [23], many single species stationary NIRS systems with calibrations based on the major cereals are routinely utilized—rice [24,26,37,39,40], sorghum [41], wheat [25,42], corn [42], barley [42]. These typically use a large number of single specie samples from different environments, representing a relevant range of target trait variability [3,43–45]. With the rising global attention on food quality, minor cereals (sorghum, fonio, teff, and pearl-, foxtail-, finger-, banyard-, little-, brown top-, kodo-, and proso-millet), are being explored [46–49]. However, the industrial use of these cereals requires rapid access to their grain components, which is currently limited. At the same time, the trait variability within the species may be another significant bottleneck constraining the development of robust calibrations. Even if large variable datasets were available, the development and maintenance of separate calibrations for each of these minor-cereal species may prove inefficient in terms of time and cost. A multi-cereal species calibration could be more convenient.

Most of the reported multi-crop calibrations focus on forage and feed analysis. Very few reports have documented successful calibration models for grain across multiple species using classical statistical modelling methods. A few authors have argued the use of ML-based approaches to bridge existing gaps in the application of NIRs technology for multi-species calibrations [72].

We aimed to evaluate the potential of emerging NIRS-sensor technology and advanced computational methods. Specifically, we studied the capacity of benchtop- and mobile-NIR sensors (FOSS-DS2500 flour analyzer and Hone Lab-EVT5) combined with a range of model-building software and methodologies (WinISI, Hone platform; statistical methods and ML-driven methods) that predict protein content in multiple cereal grain samples.

2. Materials and Methods



# 2.1
Plant material
Multiple cereal grains: 328 samples
# 2.3
Laboratory analysis of grain protein (Ground truth dataset)
Kjeldahl method (AOAC, 2000)
# 2.4
Scanning with NIR sensors
FOSS-DS2500 flour analyser (400-2498 nm)
HL- EVT5 (Hone Lab) (1350-2550 nm)
# 2.5
Calibration model development
# 2.5.1
# 2.5.2
Calibration: Validation split 80%: 20%
FOSS DS2500 + winISI
FOSS DS2500 + HoneCreate
FOSS DS2500 + Custom AI-CNN
HL-EVT5 + HoneCreate
HL-EVT5 + Custom AI-CNN
# 2.6
Model evaluation based on metrics
• Slope
• Intercept
• R², Coefficient of Determination
• RMSE, Root Mean Square Error
• RPD, Ratio of Prediction to Deviation

Figure 1: Graphical overview of the methodology followed for developing robust calibration models for protein assessment in multiple cereal grains using benchtop FOSS-DS2500 and handheld HL-EVT5 NIR sensors.

2.1. Plant material

A total of 328 grain samples from five cereal species—154 genotypes of sorghum [*Sorghum bicolor* (L.) Moench], 125 genotypes of pearl millet [*Cenchrus americanus* (L.) Morrone], 20 genotypes of finger millet (*Eleusine coracana* Gaertn.), 19 genotypes of foxtail millet [*Setaria italica* (L.) P. Beauvois], and 10 genotypes of maize samples (*Zea mays* L.; details in Appendix A). The maize cultivars were obtained from the maize improvement program of International Maize and Wheat Improvement Center (CIMMYT) and the remaining material from gene bank repository of ICRISAT [52] (genebank) and ICRISAT crop improvement programs. The subset of 154 sorghum samples included four races (bicolor, caudatum, durra, and guinea) originating from Burkina Faso, Cameroon, Ethiopia, India, Lesotho, Mali, Nigeria, and USA [53–56]. The subset of 125 pearl millet samples used in the study comprised of 100 lines from the pearl millet inbred germplasm association panel (PMiGAP) [57] and 25 elite cultivars from Asian and African origin. Samples of 20 finger millet genotypes originating from India, Kenya, Malawi, Senegal, Uganda, and Zimbabwe, and 19 foxtail millet genotypes originating from China, India, Iran, Pakistan, Russia, and the USA were used in the study [52].

2.2. Sample collection and preparation

The crops were raised on alfisol soil with recommended management practices [58] under irrigated conditions at the ICRISAT campus (Patancheru, India, 17.53°N latitude and 78.27°E longitude, 545 m.a.s.l) during the postrainy season (October 2018–January 2019). The panicles from physiologically mature plants were harvested and manually threshed. Separately for each of the genotypes, grains were pooled, cleaned, and ground to flour of a particle size of <1 mm using a CM 290 Cemotech™ laboratory grinder (FOSS, Hillerød, Denmark). The flour samples were then stored in 50-ml conical polypropylene Falcon tubes at 4°C until analysis (see section 2.3) and scanning (see section 2.4).

2.3. Laboratory analysis of grain protein content ("Ground truth" dataset)

The flour samples were dried at 130°C for 2 hours in an oven and cooled to room temperature prior to chemical analysis. Standard AOAC (2000) protocols [59] were followed to estimate moisture (AOAC 925.10) and total nitrogen content (N%; Kjeldahl method, AOAC 2001.11) in each sample (i.e., for each genotype separately; see section 2.1). Total protein content was then calculated by multiplying the nitrogen content with the protein conversion factor of 6.25 [60–62].

$$\text{Protein} = \text{N\%} \times 6.25$$

All the values were reported on a dry matter basis i.e., weight of the component per total dry weight of the sample (%, [g.100 g^{-1}]) (Supplementary Table S1).

2.4. Scanning samples with NIR-sensor devices

Prior to scanning, the samples were dried at 50°C for 16 hours and cooled to room temperature. The samples were then scanned using a benchtop NIR spectrometer DS2500 flour analyzer from FOSS (FOSS-DS2500; FOSS Electric A/S, Hillerød, Denmark) [4] and Hone Lab's handheld NIR sensor-based device HL-EVT5 (Hone Lab-Engineering Validation Test Model 5/ HL-EVT5; Hone, Newcastle, Australia) [63].

Benchtop NIR Sensor: For obtaining the spectral sample signature from the FOSS-DS2500, each flour sample was transferred to the standard circular ring cup (inside diameter ~6 cm, FOSS sample cup) and scanned three times at room temperature (~26°C). The sample was mixed before each scan. The NIR spectral absorbance, ranging from 400–2498 nm, was recorded as the logarithm of reciprocal reflectance (1/R) with 2 nm intervals using the WinISI spectral analytical software (v4.4, InfraSoft International LLC, PA, USA).

Handheld NIR sensor: To obtain the sample spectral signature from a handheld HL-EVT5, the dried flour sample was spread on a glass petri plate with a minimum 5 mm thickness. The device was then placed on the layer of flour and scanned at a room temperature of ~26°C. The device was operated via a Bluetooth-connected Hone Create mobile application (v25.2.2 Hone, Newcastle, Australia; retrieved from play.google.com). Each sample was scanned at three different points of the sample spread on the petri plate. The mobile application was programmed to record two scans at each position resulting in six scans per sample. NIR reflectance spectra ranging from 1350–2550 nm with a resolution of 16 nm at a wavelength of 1550 nm (NeoSpectra-Micro optical engine, Si-Ware Systems, CA, USA) was extracted from the Hone Create platform [64].

2.5. Calibration model development

2.5.1.Definition of calibration and validation datasets

The spectral data of 328 samples extracted from the FOSS-DS2500 and HL-EVT5 were associated with the respective laboratory protein estimates (see section 2.3). The spectral data from HL-EVT5 was then randomly split into calibration and validation datasets (80%:20%, respectively) using the Hone Create Platform. Several iterations of the split were performed and compared using histograms to ensure the random split contained all species and that they were equally represented in both the calibration and validation set (Fig 3 a, b). The calibration dataset with 262 samples (80% of total dataset) was then used to develop the calibration model and the validation dataset (20% of total dataset) with 66 samples, was used to evaluate the prediction potential of the model (details in section 2.6). The exact same split of calibration and validation samples was made for the FOSS-DS2500 spectral data. That way, we could compare the sensors and the methods of building calibration models— i.e., WinISI spectral analytical software (v4.4, InfraSoft International LLC, PA, USA), cloud-based Hone Create software (v25.2.2 Hone, Newcastle, Australia; retrieved from play.google.com), and customized convolution neural network algorithms (TensorFlow/Keras API) [65, 66].

2.5.2.Spectra pre-treatment and model development:

The WinISI analytical software is designed to assess FOSS-sensor generated data in the proprietary data format. The HL-EVT5 sensor data could therefore not be evaluated using the WinISI software. However, the Hone Create Platform and customized CNN algorithm allows users the option to import spectral data from any instrument, and was therefore used to treat FOSS-DS250 and HL-EVT5 generated spectral signatures.

2.5.2.1. FOSS-DS2500 NIR spectral data processed by WinISI spectral analytical software:

The WinISI software (v4.4) offers several mathematical spectra pre-processing steps: standard normal variate (SNV, range tested), baseline shift, NIR trajectory derivative and smoothing. After spectral pre-processing, calibrations can be built using several deterministic methods: principal component regression (PCR), partial least square regression (PLSR), and modified partial least square (MPLS) — in combination with pre-treatment methods [67–69]. Iterations between the methods can be performed manually and the prediction potential of the models built can be tested using the validation dataset (described in 2.5.1). Accordingly, we performed several manual iterations between the available methods. Calibration models achieving the best metrics, i.e., slope and intercept of linear regression, coefficient of determination (R^2), root mean squared error (RMSE), and relative prediction deviation (RPD), for the calibration and validation datasets were then manually selected for further comparisons (sections 2.5.2.2. and 2.5.2.3).

In this case, the spectral data of the FOSS-DS2500 calibration set was scatter-corrected using SNV&D that reduced the interference due to physical characteristics (such as particle size and path length of sample to the spectra). The corrected spectra were then subjected to a derivative algorithm i.e., second derivative treatment, an eight nm spacing over which the derivative was calculated (a gap of four wavelength points multiplied by two nm), first smoothing using Savitzky–Golay polynomial at four data points, and no second smoothing (i.e., mathematical pre-treatment setting 2,4,4,1).

2.5.2.2. FOSS-DS2500 and HL-EVT5 NIR spectral data processed by Hone Create software: The automated Hone Create software [64] applies a matrix of pre-processing options specific to each chemical variable; these include baseline correction, area normalization, smoothing, derivative, standard normal variate (SNV), or combinations of these techniques. Hone Create automatically iterates and selects the best performing pre-processing method(s) based on regression (PLS) or classification (C4.5) models. Once processed, a range of machine learning techniques are automatically tested and compared, including Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), eXtreme Gradient Boosting (XGBoost), Deep learning, and Stacked Ensembles. The best performing calibration model is then selected based on the root mean squared error (RMSE) and coefficient of determination (R^2) of the cross-validation set. Holdover validation metrics (independent validation set, described in 2.5.1.) are automatically processed for the user, with interactive results displayed to allow the user to interrogate the dataset and trigger further model iterations as needed.

In this study, to compare the prediction potential of sensor–model combinations, the raw spectra of calibration datasets from FOSS-DS2500 and HL-EVT5 sensors (described in 2.5.1) were treated separately. The spectral data from FOSS-DS2500 and HL-EVT5 in excel format (.xlsx) was fed into the Hone Create platform. For our dataset, the best method for both spectra pre-processing included transformation via “area normalization” a followed by spectrum merging steps “smoothing and derivative”. For model development, the software automatically selected the “stacked ensemble” algorithm for both the FOSS-DS2500 and HL-EVT5 datasets. In addition, Hone Create software automatically applied the method on the validation set and displayed results with required metrics that was further compared with other tested methods. (Details in Section 2.6).

2.5.2.3. FOSS-DS2500 and HL-EVT5 NIR spectral data processed by customized CNN algorithm: Convolutional neural networks (CNNs) methodology [65] was also explored for building multivariate regression calibration models using the publicly available open-source TensorFlow/Keras API [66]. This CNN is composed of three convolutional layers, three pooling layers, and three fully connected layers. Each convolution layer had 24, 48, and 96 filters with kernel sizes set to 10, 15, and 25, respectively. All stride parameters were attributed to two. The network was then organized between convolutional layer and pooling to realize the extraction and mapping of local features from the input NIR dataset. Several fully connected layers were then consecutively arranged, and the regression of targets was performed by sigmoid. Batch normalization was added after every convolutional layer to prevent an internal covariate shift and to speed up convergence. The ADAM50 function, a gradient descent algorithm, was set to minimize the loss function with an initial learning rate of 3 × 10-4, which enabled the reverse adjustment of weights from the network using a backpropagation algorithm and reducing the mean squared error of the model after each training iteration [70]. A max-pooling layer, with kernel filter size set to two, was connected to each layer of activation function. A dropout of 0.02 was then used to deactivate 2% of the network neurons. Finally, the output of the last dropout layer was flattened to represent the high-dimension features of the input dataset. The extracted high-dimensional features were fed into a multi-layer perceptron (MLP) to execute the final regression task. There were hidden layers in the MLP with 512 and 128 neurons, successively. A regularization term (index = 10-7) was added to every hidden layer to minimize the overfitting followed by batch normalization. The model is trained with a training batch size of 64 using Google Colaboratory (NVIDIA K80s GPU, 12.72 of RAM and 358.27 GB of hard disk for one runtime), an open-source service provided by Google.

The spectral datasets from the FOSS-DS2500 and HL-EVT5 were pre-treated using the Savitzky-Golay smoothing filter [67] with a window size of 15 and a polynomial order set to 2; as done in similar studies [71,72]. The transformed data, before being fed to the CNN, was then normalized using min/max normalization of the 1st derivatives so that values ranged from between 0 and 1 [34]. Subsequently, the CNN training structure was constructed to predict protein quantity (%, [g/100g]) from spectral data. For this, the calibration set model was trained using a five-fold cross validation approach to determine the good number of epochs and the effectiveness of certain hyperparameters, such as activation functions, neuron counts and layer counts. The model (with the same architecture as that of the cross-validation) was retrained with the selected hyperparameters on the entire training dataset and tested with the validation dataset (described in 2.5.1). Iterations between the pre-treatment and normalization methods were performed and the best performing model was selected based on the common metrics of both calibration and validation datasets (details in Section 2.6).

2.6. Prediction model evaluation

To compare the predictive potential of the different sensor–model combinations, we used the statistical metrics describing the linear inter-dependency between the ground-truth (grain protein content estimated by laboratory method, see 2.3) and the best model for predicting protein content from NIRS spectra (separately for calibration and validation sets, see section 2.5.1). To assess the quality of the NIRS models developed, five parameters were used: slope and intercept of the linear regression, coefficient of determination (R^2; Equation1), root mean squared error (RMSE; Equation 2), and ratio of prediction to deviation (RPD; Equation 3) [73–76].

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(\hat{y}_i - \bar{y}_i)^2} \qquad (1)$$

$$RMSE = \sqrt{\frac{\Sigma(\hat{y}_i - y_i)^2}{n}} \qquad (2)$$

Where, n is the number of samples; y_i is the ground-truth (see section 2.3) value of sample i; $\hat{y}_i$ is the model-predicted value of sample i; $\bar{y}$ is the mean of the ground truth values; SD is the standard deviation of ground truth values.

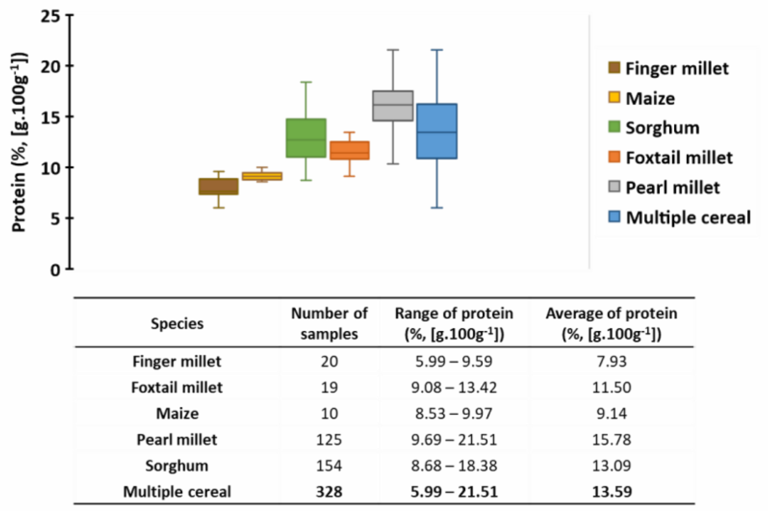
$$RPD = \frac{SD}{RMSE} \qquad (3)$$

We also adopted the previously reported classification based on RPD values [75], wherein a value for the RPD < 1.5 indicates that the calibration is not reliable; a value between 1.5 and 2.0 indicates the capacity of a model to distinguish between high and low values; a value between 2.0 and 2.5 signifies the model capacity to "approximate" quantitative prediction; a value between 2.5 and 3.0 suggests "good" quantitative prediction; and a value > 3.0 indicates "excellent" quantitative prediction.

3. Results

3.1. Diversity in grain samples

The laboratory analysis of protein content obtained from 328 grain samples across five cereals species ranged from 5.99% to 21.51% (Supplementary Table S1). The range of protein content in multiple cereals was considerably larger compared to protein content variability within any of the individual species tested (Figure 2). Among the five cereals tested, the mean protein content in pearl millet grains was the highest (21.51%) while the mean protein content was the lowest in finger millet grains (5.99%; Figure 2).



Species	Number of samples	Range of protein (%, [g.100g^{-1}])	Average of protein (%, [g.100g^{-1}])
Finger millet	20	5.99 – 9.59	7.93
Foxtail millet	19	9.08 – 13.42	11.50
Maize	10	8.53 – 9.97	9.14
Pearl millet	125	9.69 – 21.51	15.78
Sorghum	154	8.68 – 18.38	13.09
Multiple cereal	**328**	**5.99 – 21.51**	**13.59**

Figure 2. Box plots depicting variation and distribution of protein [%, g$\cdot$100 g^{-1}] content in five cereal grains estimated through laboratory analyses. [Legend: Each box represents one crop group and mean of each crop group is represented by a solid line (–). Different crop groups are distinguished by color (Finger millet = Brown; Maize = Yellow; Sorghum = Green; Foxtail millet = Orange; Pearl millet = Grey; and Multi-cereals = Blue). The adjoining table shows the number of samples of each of the five cereal species used in the study along with the range and average grain protein content [%, g. 100g^{-1}] estimated through laboratory-based analysis.

The range, average, standard deviation, and distribution of protein content across the calibration and validation datasets were comparable (Figure 3; Supplementary Table S2, S3).

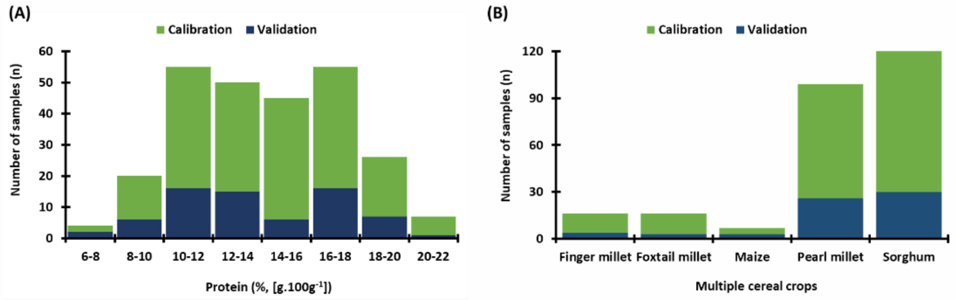


Figure 3: Histograms depicting the distribution of samples in the calibration (80%) and validation (20%) datasets across the (A) protein ranges and (B) crop species used for the development of protein prediction models with different sensor–model combinations. The details are shown in Supplementary Table S2 and S3.

3.2. NIR spectrum characters obtained from benchtop FOSS-DS2500 and handheld HL-EVT5

The NIR reflectance spectra of 328 samples were recorded using two NIR-sensor devices: the stationary benchtop FOSS-DS2500 (400–2498 nm) and the handheld HL-EVT5 (1350–2550 nm) device. The spectrum profile generated by each sensor was diametrically different (Figure 4). Nevertheless, within each of the sensors the detected NIRS absorbance (log1/R) trajectories were similar for all tested crops species with the major peaks and troughs at similar positions. This indicated that the technology used to generate the NIRS spectral signatures is different but the biological samples used here had very similar bio-chemical signatures.

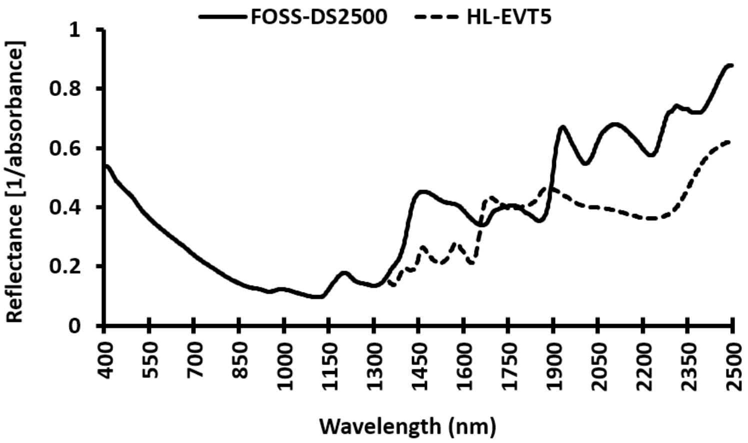


Figure 4: Mean of the near infrared (NIR) raw spectra of all grain samples extracted from the Benchtop FOSS-DS2500 (400–2498 nm; solid line (–) and handheld HL-EVT5 (1350–2550 nm; dashed line (---) devices.

Overall, the FOSS-DS2500 signal was dominated by thirteen groups of prominent peaks (Figure 5A) and five peaks for the HL-EVT5 (Figure 5B). The protein content is known to be linked to several spectral bands: (i) between 950–1050 nm as N–H second stretch overtone, (ii) around 1500 as N–H stretching first overtone, (iii) the N–H bend second overtone, and C=O stretch–N–H in-plane bending–C–N stretch combination bands are further associated with the range between 2150–2200 nm [11]. Therefore, both of the devices should be sufficient to capture some, if not all, of these critical NIR spectral ranges to predict protein content.

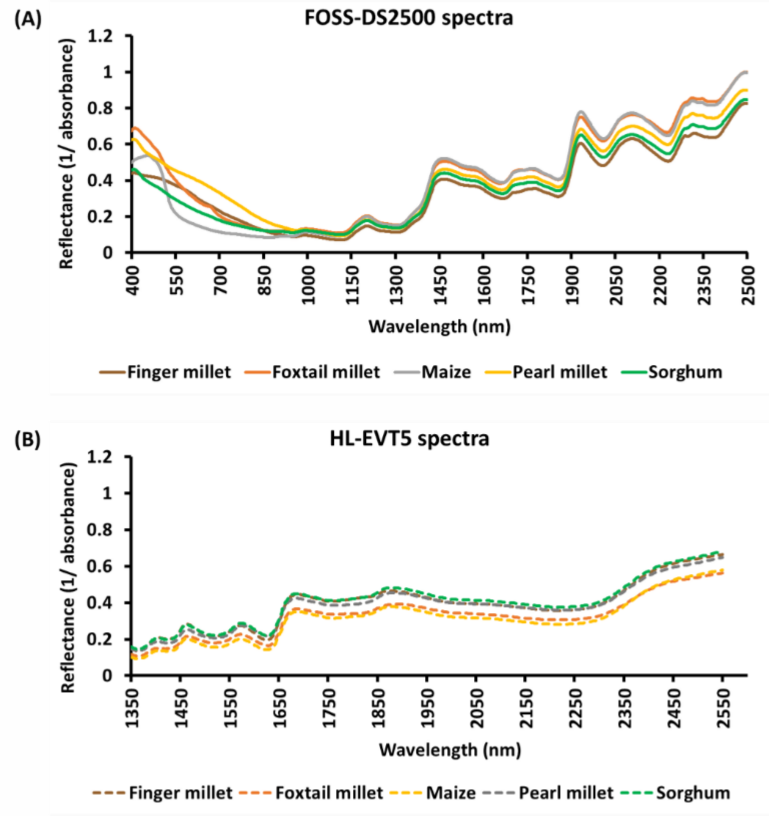


Figure 5: Mean of the near infrared (NIR) raw spectra of grain samples of five cereal species produced by (A) FOSS-DS2500 (400–2498 nm) and (B) HL-EVT5 (1350–2550 nm) devices. Different crop groups are distinguished by color (Finger millet = Brown; Foxtail millet = Orange; Maize = Yellow; Pearl millet = Grey; Sorghum = Green).

3.3 NIR spectrum generated by FOSS-DS2500 processed by WinISI software

WinISI software enabled manual iterations between spectra pre-processing steps (SNV, Detrend, NIR trajectory derivative, and smoothing) and several deterministic calibration model algorithms (PLS, MPLS, and PCR). Out of the manually iterated options, the best calibration model which attained the highest accuracy metrics was obtained using the modified partial least square (MPLS) combined with scatter correction (SNV&D) and mathematical pre-treatment "2,4,4,1" (i.e., mathematical treatment "Default 2" setting in WinISI). This calibration model achieved good RMSE values of 0.91 and R^2 of 0.90. The validation dataset had RMSE values of 1.09 and R^2 of 0.86 (Table 1; Figure 6). The RPD values for the calibration and validation datasets were 3.56 and 3.08, respectively (Table 1).

Table 1: Comparative metrics of NIR spectroscopy calibration (80%) and validation (20%) models using combinations of different sensors and deterministic algorithms developed for evaluating the prediction capacities of the models to estimate protein (%, [g.100 g^{-1}]) in multi-cereal grains [R^2=coefficient of determination; RMSE=Root Mean Squared Errors, RPD=ratio of prediction to deviation, CNN=Convolutional Neural Networks].

Sensor	Model	Set	Slope	Intercept	R^2	RMSE	RPD
FOSS-DS2500	**WinISI**	Calibration	0.87	1.74	0.90	0.91	3.56
		Validation	0.82	2.38	0.86	1.09	**3.08**
	Hone	Calibration	0.95	0.64	0.96	0.66	4.93
		Validation	0.89	1.44	0.90	1.00	**3.38**
	CNN	Calibration	0.98	0.29	0.99	0.33	9.85
		Validation	0.88	1.61	0.89	1.03	**3.26**

HL-EVT5	**Hone**	Calibration	0.97	0.43	0.98	0.42	7.79
		Validation	0.90	1.35	0.91	0.97	**3.48**
	CNN	Calibration	0.98	0.28	0.98	0.46	7.00
		Validation	0.87	1.70	0.87	1.10	**3.06**

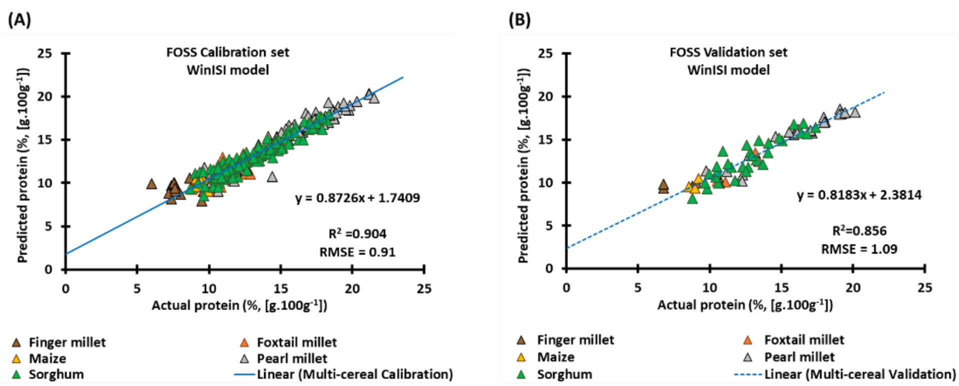


Figure 6: Scatter plot showing protein predicted for the (A) calibration and (B) validation sets of FOSS-DS2500 via WinISI analytical software. The best model was built using spectral data pre-processed by SNV&D and a deterministic MPLS method. Detailed metrics are shown in Table 1.

3.4. NIR spectrum generated by FOSS-DS2500 and HL-EVT5 processed by Hone Create software

Hone Create software automatically selected pre-processing methods specific for the generated data and iterated these with the range of deterministic and ML-based calibration methods. The pipeline enabled the performance of all possible combinations of methods, returning a 'Top 10' calibration model leaderboard, with the best performing calibration (based on R^2 and RMSE of calibration set) presented for further analysis and iterations.

For both NIR-sensor devices (FOSS-DS2500 and HL-EVT5), we observed the best prediction models included area normalization, smoothing and derivative (pre-processing steps) combined with stacked ensemble models. In the case of the longer spectrum signatures generated by the benchtop FOSS-DS2500 instrument (Figure 7A&B), Hone Create's best model determined the protein content with RMSE values of 0.66 and 1.00, and R^2 values of 0.96 and 0.90 for calibration and validation datasets, respectively. This model had an RPD value of 3.38 for the validation dataset (Table 1). With the HL-EVT5 spectra, the calibration model with an R^2 of 0.98 and RMSE of 0.42 predicted protein with an R^2 of 0.91 and RMSE of 0.97 (Figure 7C&D). The model showed RPD values of 7.79 and 3.48 for calibration and validation datasets, respectively (Table 1).

doi:10.20944/preprints202202.0335.v1

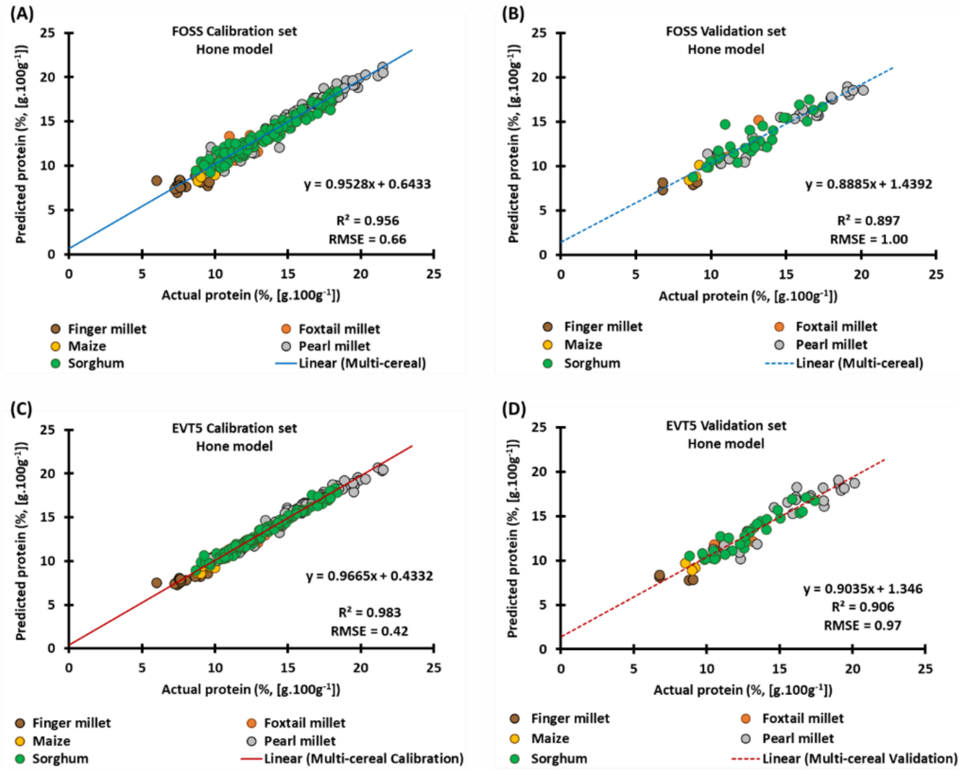


Figure 7: Scatter plots showing protein predicted for the (A&C) calibration and (B&D) validation datasets of (A&B) FOSS-DS2500 and (C&D) HL-EVT5 by Hone Create Software. The best model was built using spectral data pre-processed by an area normalization algorithm, followed by smoothing and derivative for the spectrum merging step, and the ML-driven stacked ensemble method for model building. Detailed metrics shown in Table 1.

3.5. NIR spectrum generated by FOSS-DS2500 and HL-EVT5 processed by deep learning CNN

Deep learning CNN models were also experimented with to predict the protein content in cereals grains. For the FOSS-DS2500 samples, the CNN model with R^2 of 0.99 and an RMSE of 0.33 predicted protein content in the validation dataset with prediction accuracies of R^2 of 0.89 and an RMSE of 1.03 (Figure 8A & B; Table 1). However, for the HL-EVT5 sensor samples, an RMSE of 0.46 and 1.10 and R^2 of 0.98 and 0.87 were obtained for calibration and validation datasets, respectively (Figure 8C & D; Table 1). The RPD values for FOSS-DS2500 and HL-EVT5 sensor-derived validation datasets were 3.26 and 3.06, respectively (Table 2).

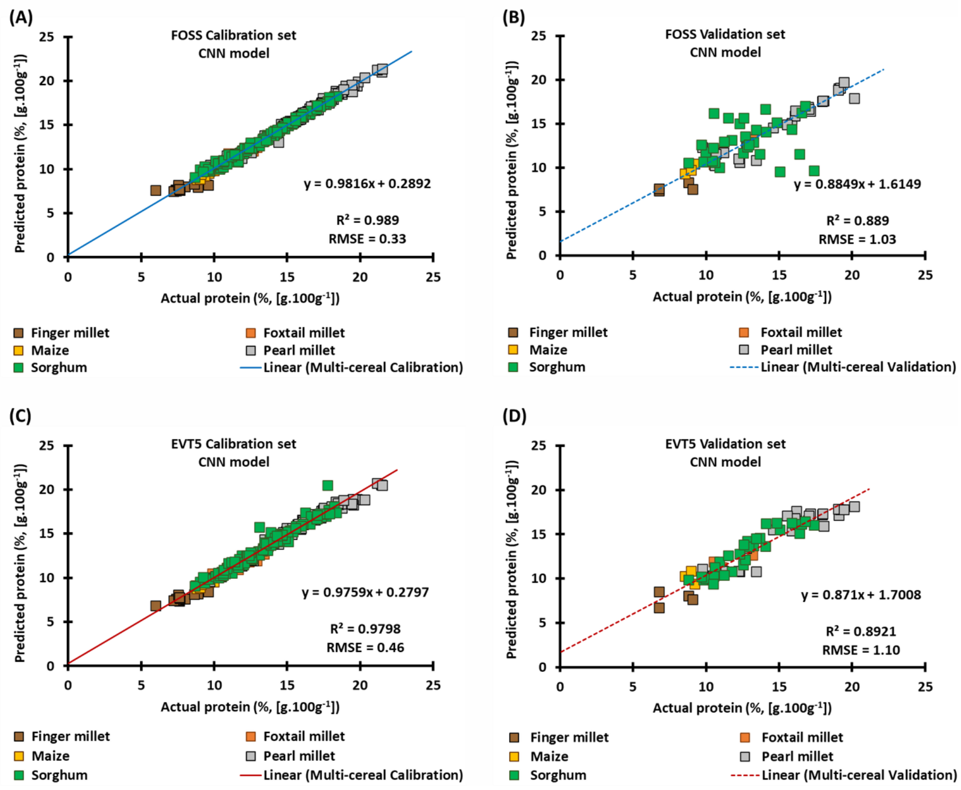


Figure 8: Scatter plots showing protein predicted for the (A&C) calibration and (B&D) validation datasets of (A & B) FOSS-DS2500 and (C & D) HL-EVT5 by a customized CNN algorithm. The best model was built using spectral data pre-processed by the Savitzky-Golay smoothing filter, followed by min/max normalization of the 1st derivatives and customized deep learning CNN algorithm for model building. Detailed metrics shown in Table 1.

3.4. *Evaluation of sensor-model combinations for protein predictions*

Two sensor technologies (FOSS-DS2500 and HL-EVT5) were used to generate NIR spectra (Figure 4) of ground grain samples in order to predict their protein content. Three analytical processes (WinISI software, Hone Create software and custom-made CNN algorithm) were tested with each dataset to build the prediction model for grain protein content ([g.100 g^{-1}], %). Five statistical metrics (slope and intercept of the linear regression, R^2, RMSE, and RPD; section 2.6) were used to assess and compare the performance of the different chemometric models.

Overall, as show in Table 1, NIR spectral signal generated by both sensors (FOSS-DS2500 and HL-EVT5) yielded reliable models when combined with relevant statistical and ML-driven methods (R^2 ≥ 0.86, RMSE ≤ 1.09). The RPD of all of the sensor–model combinations for estimating protein were greater than 3.1 (Table 1). This suggests that all generated models are well suited to provide quantitative protein estimates.

Nevertheless, for both types of NIR signal, the calibration models built using ML-methods (Hone Create software and CNN) achieved notably higher RPD (≥ 3.26), R^2 (≥ 0.89) and lower RMSE (≤ 1.03) values compared to deterministic models created through WinISI software (RPD = 3.08, R^2 = 0.86, and RMSE = 1.09) (Table 1). However, the difference between metrics of the calibration and validation datasets (Table 1) was much higher for the ML-based models compared to the deterministic methods, which could signify the tendency of ML-methods to produce the models specific to the given dataset with less generic prediction power; i.e., “over-fitted” models.

4. Discussion

4.1. Utilization of NIR spectroscopy for grain quality assessment in cereals

Cereals are important sources of human and livestock diets that significantly influence their nutritional status and health [77,78]. In the subsistence farming systems of developing countries, the staple food base frequently consists of minor cereals (such as millets, sorghum, fonio, teff) that significantly contribute to the region's food and nutritional security [78–80]. Protein, assessed in the study, is one of the key parameters not only determining nutritional grain values in the human diets, but also its suitability for food industries [13–15,23]. The protein content can be analyzed by standard laboratory analytical methods [59] but these are time-, resource- and cost-intensive and produce hazardous chemical waste in the process. To overcome such constraints in the estimation of grain qualities, NIR spectrometric systems are being deployed [13,14,16,22,81]. NIR spectroscopy has been successfully used in agricultural research since the 1960s, initially to analyze seed moisture content [82]. Over the years, significant developments of NIR-based sensor devices have been achieved. However, these are standardized and commonly used for a few major crops and mainly in well-developed centralized value-chains [14,16,83]. So far, there are not many examples where NIR-based systems have been used for the quality assessment of minor cereal grains (like millets in presented study 84,85). The rapid assessment of the minor cereal qualities could be an entry point to integrate these cereals in mainstream food value-chains. Extended utilization of these minor cereals could become a powerful approach to some of the global goals that aim to improve human nutrition [46–48], as many have high dietary values compared to the major cereals like wheat, rice or maize [80,86–88]. Therefore, in this work NIR-based calibration models were developed for multiple cereals species as a technology to facilitate the integration of the nutrition-rich cereals in human diets.

4.2. Expected data properties as pre-requisites to building reliable models

Important considerations for building reliable prediction models from NIR spectra were comprehensively summarized in [22,73,74]. In our study, we encompassed critical aspects, including: (i) range of variation in the protein content; (ii) sample number and quality of laboratory protein analysis; (iii) quality and properties of the spectra generated by two sensors (DS-2500 and HL-EVT5); and, (iv) methods for building the algorithms (discussed in section 4.3).

The range of trait variability (protein, in this case) is one of the critical prerequisites for the development of reliable calibration models from NIR-based signals [73]. Inclusion of five cereal species– sorghum, pearl millet, foxtail millet, finger millet, and maize – enabled us to extend the ranges of protein variability compared to those found in any of the individual species. In this study, the range of protein found across multiple cereals (i.e., 15.5%) was several-fold higher (1–11 fold) than the range of protein in any of the individual cereals tested in this study (Supplementary Table S3).

In addition to trait variation, the success of the model also depends on the number of samples used for laboratory analysis as well as laboratory analysis precision ("groundtruth"). Generally, a higher number of high-quality data points increases the probability of building reliable calibrations. Number of samples is particularly important for ML-based algorithms [89]. In our study, we used 328 groundtruth samples. Chemometric analysis was done in a certified facility. The literature reveals attempts of calibration building starting with ~80 samples [90] while industrial calibrations typically use > 300 samples [91]. Considering these aspects, our sample size was relevant and appropriate for the intended use.

Clearly, trait prediction from NIR signals depends on the spectra properties and quality, determined by hardware used to generate the spectra. In our case, we deployed the standard benchtop FOSS-DS2500 flour analyzer system and novel handheld HL-EVT5 sensor. The stationary NIR-spectrophotometric systems are now routinely used and certified for the assessment of major cereal grain quality [e.g., wheat in Australia; 92–94]. Recent technological advancements have enabled miniaturization, compaction, integration, and mobilization to expand the use of this technology [15]. However, rigorous testing and the relevance of the upcoming generation of NIR instruments is not commonly standardized and documented [95]. Thus, in this study the performance of the standard NIR benchtop spectrometer (FOSS-DS2500 with a spectral range of 400–2498 nm) was compared with a handheld NIR device (HL-EVT5 with a spectral range of 1350–2550 nm) in combination with different model-building methods (see section 2.5). Both sensors covered NIR spectrum ranges relevant to protein determination in grain samples (section 3.2) [2,96–98]. Through this, we could effectively compare the spectral data from both sensors as well as develop and compare a range of classical deterministic and ML-based calibration methods to capture and predict the variability of protein content across five cereal species. Considering the above characteristics of the data, we built the algorithms to infer the protein content from the NIR-based scans of cereals grain samples.

4.3. Prediction algorithms and their relevance for estimation of protein content in cereal grains

For predictions of organic grain composition from NIR spectral reflectance, deterministic methods have been widely used (e.g., MLR, PCR, PLS regression [13,16,18]. These methods were mostly specific to single species [23]. To date, only a few studies have demonstrated that calibrations built on data from multiple species could be as accurate as single species calibrations [43,45,50]. More frequently, it has been pointed out that deterministic calibrations based on material from multiple species might compromise prediction accuracy [72]. Some authors suggested that alternative modeling methodologies, such as ML-based techniques, could be a solution to building reliable calibrations across species [72] as they have the capacity to efficiently process collinear spectral data, learn latent data structures and identify latent patterns and representations. The adoption of ML methods (notably deep convolutional neural networks) in NIR spectroscopy research have begun quite recently along with the increasing availability of computing power and efficient learning algorithms (e.g., backpropagation).

The novelty of this work is that we systematically compared the new generation of handheld NIR sensor technology (HL-EVT5) with the standard benchtop NIR device (FOSS-SD2500) along with newly emerging methods (machine learning and deep learning architectures) and software (FOSS-made WinISI, Hone-made Hone Create) to build prediction algorithms.

While the FOSS software (WinISI) leaves the user to manually test the combinations of several preprocessing methods, the Hone Create software enables automatic evaluation of an enormous range of preprocessing methods and models in a relatively short amount of time. The one potential constraint of the current Hone Create pipeline is that it selects the "best" models based on the metrics of the calibration dataset. This process might prefer models that are more specific for the dataset presented with less generic prediction capacity (i.e., "over-fitting models"; when the model has good metrics for calibration set but predicts poorly for the validation dataset [72]). This was also one of the reasons why we tested the alternative process, i.e., the custom-designed CNN pipeline, where we integrated the element where the goodness of the model was evaluated based on the metrics of the validation set. The performance of CNN is largely influenced by the quality and size of the datasets used for training the model. To minimize the "over-fitting" error, the large dataset was used and split carefully to include the different multi-cereal species in both the calibration and validation dataset (as described in section 2.5.1). The code is available on the GitHub platform (https://github.com/adamavip/nirs-protein-prediction) and its particular parts can be now utilized to enhance and develop other pipelines and products.

For our dataset, the algorithms built using ML-based methods (particularly, the stacked ensemble model via Hone Create and custom-designed CNN; section 3.4) achieved the better comparative metrics for both spectra types (i.e., generated by the benchtop FOSS-DS2500 and the handheld HL-EVT5). For the same dataset, the prediction model obtained with deterministic methods (i.e., modified partial least squares method through WinISI software) was also relevant for high quality quantitative predictions (R^2=0.86, RMSE=1.09, RPD=3.08). Interestingly, the models built using the shorter spectra derived by the HL-EVT5 sensor attained a similar level of prediction accuracy (RPD=3.5) compared to the longer spectra derived from FOSS-DS2500 (RPD=3.4) (Table 1). Overall, all five of the presented calibration models developed for protein assessment can be used for high quality quantitative estimation of protein in a range of cereal grains. In addition, the study suggests that the robustness of these calibrations, especially the ML-based ones, can be further improved by including additional and more diverse samples to further expand and represent the wide range of trait and spectral variability.

Similar studies have been carried out to evaluate the grain protein on individual major cereals [97,99,100]. These studies worked with narrower ranges of trait variability and achieved accuracy metrics of calibration models comparable to those achieved in the presented study (typically $R^2 \geq 0.86$, RPD > 3.0*).* This exercise highlighted that the calibration models based on ML-algorithms can surpass the accuracy of deterministic methods. These approaches also overcome the previously reported gaps in prediction accuracies in global calibrations [72]. There is therefore great potential for these methods to build prediction algorithms for heterogeneous material.

4.4 Mobilized NIR spectrometers and need for outdoor applications

Recently, the Food and Agriculture Organization (FAO) announced 2023 as the International year of millets recognizing the superior nutritional value of these minor cereals and promoting them as a means to improve diets and globally support sustainable development goals (particularly SDG2; End hunger, achieve food security, improve nutrition, and promote sustainable agriculture [46–48,80,84,85]). However, the inability to rapidly quantify grain composition in minor cereals could limit their broader use (e.g., for food industries) but also further genetic improvement through breeding and research in general. While in centralized systems such gaps have been largely overcome by integrating stationary NIR systems in these processes (like, FOSS-DS2500 used here), the same might not suit the systems typical for e.g., sub-Saharan Africa and South Asia. In these systems, the majority of farming communities are smallholders and their postharvest logistics mostly decentralized. In such systems, bringing a machine to samples rather than bringing samples to a machine would be a paradigm shift that is required to support the needs of end-users. We have demonstrated that the technology means are already available to facilitate such a transition.

5. Conclusions

The transition of NIR spectroscopy from stationary to the handheld form will be critical for its effective utilization in decentralized systems, especially where produce-handling logistics might be problematic (e.g., monitoring of produce-quality in smallholder agri-systems and value-chains, crop improvement programs). The motivation of this study was to assess whether emerging technological approaches (handheld NIR sensors and machine learning (ML)-based algorithms) would enable rapid and accurate assessment of grain composition (e.g., protein content) in multiple cereal staples typical for such agri-food systems. We demonstrated that the NIR spectra, generated by a novel handheld NIR-based sensor (HL-EVT5), were sufficient for reliable for the quantification of grain protein content in multiple cereal species. We found that the integration of ML-based methods in modeling processes could enhance the model predictive accuracy compared to classical deterministic modeling methods. Additionally, we highlight that these advanced data modeling methods are now available for non-experts through several new software packages (e.g., Hone Create software), which was a major limitation in the past where more complex methodologies were relied upon. These novel technologies and methods could well become a power engine to drive transformation towards agri-food system prosperity.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1:
Table S1: List of 328 grain samples used in the study along with the crop species, genotype ID and protein content (%, [g.100g-1]) obtained from laboratory analysis.
Table S2: Descriptive statistics presenting the variability and range of protein content in the calibration and validation sets used in the study. [Legend: SD = standard deviation; CV% = coefficient of variation].
Table S3: Descriptive statistics [minimum (min), maximum (max), average (avg), standard deviation (SD), and standard error (SE)] of the protein content in each of the cereal species of the calibration and validation sets used in the study.
Video S1: Operating HoneTest app using HL-EVT5: https://fernlab.herokuapp.com/evt5.html.

Author Contributions: Conceptualization, J.K., S.C., K.A.; Methodology, K.C., S.C., J.K., K.A., A.N.; Software, A.N., J.F., W.P., V.P.; Validation, K.A., S.C., A.N., J.F., P.K.; formal analysis, K.C., K.A., A.N., S.M., P.S.; investigation, K.A., S.C., A.N., J.K., J.F; resources, J.K., P.K.; data curation, K.C., K.A., S.M., P.S.; writing—original draft preparation, K.C., K.A., A.N.; writing—review and editing, K.A., J.K., C.J., W.P., J.F., V.P.; visualization, K.A.; supervision, J.K.; project administration, J.K., P.K.; funding acquisition, J.K., S.C.
All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the CGIAR Research Program on Grain Legumes and Dryland Cereals, CGIAR Initiative–Crops to End Hunger, and Global Challenge Research Fund (GCRF)-funded project Transforming India's Green Revolution by Research and Empowerment for Sustainable food Supplies (TIGR²ESS).

Acknowledgments: The authors would like to thank Dr B.D. Ranjitha Kumari and Dr T. Senthil Kumar from Bharathidasan University for supporting the PhD student Ms. Keerthi Chadalavada. The authors thank the team at ICRISAT and CIMMYT for sharing the genetic material with us, Intertek and ILRI for the laboratory analyses and the Hone team for their expert inputs. Special acknowledgment to Dr. Sivasakthi Kaliamoorthy, Ms. Rekha Baddama, and Mr. Srikanth Mallayee for their support with experimentation; Ms. Premalatha Teegalnagaram and Mr. Mallesh Rahini for their support in laboratory activities; and Dr William Nelson for his inputs on improving the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Agelet, L.E.; Hurburgh, C.R., Jr. A tutorial on near infrared spectroscopy and its calibration. *Crit. Rev. Anal. Chem*. **2010**, 40, 246–260. https://doi.org/10.1080/10408347.2010.515468
2. Workman, J.; Weyer, L. Practical guide and spectral atlas for interpretive near-infrared spectroscopy, 2nd Ed.; CRC Press: Boca Raton, USA, **2012**. https://doi.org/10.1201/9781420018318
3. Villamuelas, M.; Serrano, E.; Espunyes, J.; Fernández, N.; et al. Predicting herbivore faecal nitrogen using a multispecies near-infrared reflectance spectroscopy calibration. *PLoS One*. **2017**, 12, e0176635.
4. FOSS-DS2500 flour analyzer. Available online: https://www.dksh.com/global-en/products/ins/foss-flour-analyzer-nirs-ds2500 (Accessed on 7 Jan 2021).
5. Bruker-Tango FT-NIR spectrometer from Bruker. Available online: https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-nir-spectrometers/tango-ft-nir-spectrometer.html (Accessed on 7 Jan 2021).
6. Perten-IM9520 flour analyzer. Available online: https://www.calibrecontrol.com/main-product-list/perten-im9520-flour-analyser (Accessed on 7 Jan 2021).
7. Sorak, D.; Herberholz, L.; Iwascek, S.; Altinpinar, S.; Pfeifer, F.; Siesler, H. W. New developments and applications of handheld raman, mid-infrared, and near-infrared spectrometers. *Appl. Spectrosc. Rev*. **2012**, 47, 83–115. https://doi.org/10.1080/05704928.2011.625748
8. Crocombe, R.A. Portable Spectroscopy. Applied Spectroscopy. **2018**, 72, 1701–1751.

9. MicroNIR OnSite-W from VIAVI Solutions. Available online: https://www.viavisolutions.com/en-us/osp/products/micronir-onsite-w (Accessed on 7 Jan 2021).
10. DLP NIRScan™ Nano EVM spectropmeter from Texas Instruments. Available online: https://www.ti.com/tool/DLPNIRNANOEVM (Accessed on 7 Jan 2021).
11. MEMS spectrometer from Fraunhofer. Available online: https://www.ipms.fraunhofer.de/en/Components-and-Systems/Components-and-Systems-Sensors/Optical-Sensors/MEMS-based-spectroscopy.html (Accessed on 7 Jan 2021).
12. Hone Lab Red from Hone. Available online: https://www.honeag.com/hone-lab (Accessed on 7 Jan 2021).
13. Osborne, B.G. Near infrared spectroscopy in food analysis. Encyclopedia of analytical chemistry: applications, theory and instrumentation. John Wiley & Sons Ltd, New York. **2006**. https://doi.org/10.1002/9780470027318.a1
14. Singh, C.B.; Paliwal, J.; Jayas, D.S.; White, N.D. Near-infrared spectroscopy: Applications in the grain industry. In: CSBE/SCGAB Annual Conference, Edmonton, Alberta, 16–19 July **2006**, Paper No. 06-189.
15. dos Santos, C.A.T.; Lopo, M.; Páscoa, R.N.M.J.; Lopes, J.A. A Review on the Applications of portable near-infrared spectrometers in the agro-food industry. *Appl. Spectrosc*. **2013**, 67, 215–1233.
16. Williams, P.C. Application of near infrared reflectance spectroscopy to analysis of cereal grains and oilseeds. *Cereal Chem*. **1975**, 52, 561–576.
17. Norris, K.H.; Barnes, R.F.; Moore, J.E.; Shenk, J.S. Predicting forage quality by infrared reflectance spectroscopy. *J. Anim. Sci.* **1976**, 43, 889–897.
18. Estienne, F.; Pasti, L.; Centner, V. et al. A comparison of multivariate calibration techniques applied to experimental NIR data sets: Part II. Predictive ability under extrapolation conditions. *Chemometr. Intell. Lab. Syst*. **2001**, 58, 195–211.
19. Esbensen, K.H.; Julius, L.P. Representative sampling, data quality, validation—A necessary trinity in chemometrics. In Comprehensive chemometrics (Eds. Brown, S., Tauler, R., and Walczak, R.), Volume 4. Oxford: Elsevier. **2009**, 1–20.
20. Agelet, L.E.; Hurburgh Jr, C.R. Limitations and current applications of near infrared spectroscopy for single seed analysis. *Talanta*. **2014**, 121, 288–299.
21. Chang, H.; Zhu, L.; Lou, X.; et al. A new local modelling approach based on predicted errors for near-infrared spectral analysis. *J. Anal. Methods Chem.* **2016**, 2016, 5416506.
22. Cheewapramong P. Use of near-infrared spectroscopy for qualitative and quantitative analyses of grains and cereal products. PhD thesis, University of Nebraska-Lincoln, Lincoln, NE. **2007**
23. Downey G. NIR and chemometrics in the service of the food industry. *NIR news*. **2007**, 18, 10–1.
24. Chen, J.Y.; Miao, Y.; Sato, S.; Zhang, H. Near infrared spectroscopy for determination of the protein composition of rice flour. *Food Sci. Technol. Res.* **2008**, 14, 132–138.
25. Kahriman, F. and Egesel, C.Ö. Development of a calibration model to estimate quality traits in wheat flour using NIR (Near Infrared Reflectance) spectroscopy. *Res. J. Agric. Sci.* **2011**, 43, 392–400.
26. Bagchi, T.B.; Sharma, S.; Chattopadhyay, K. Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. *Food Chem.* **2016**, 191, 21–27.
27. Lyu, N.; Chen, J.; Pan, T. et al. Near-infrared spectroscopy combined with equidistant combination partial least squares applied to multi-index analysis of corn. *Infrared Phys. Techn.* **2016**, 76, 648–654.
28. Sampaio, P.S.; Soares, A.; Castanho, A.; Almeida, A.S.; Oliveira, J.; Brites, C. Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chem*. **2018**, 242, 196–204.
29. Tomas, E.; Bayram, I. Establishing near infra red spectroscopy (NIR) calibration for starch analysis in corn grain. *Kocatepe Vet. J.* **2018**; 12: 7–14.
30. Chen, J.; Li, M.; Pan, T.; Pang, L.; Yao, L.; Zhang, J. Rapid and non-destructive analysis for the identification of multi-grain rice seeds with near-infrared spectroscopy. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2019,** 219, 179–185.
31. Lee, S.; Choi, H.; Cha, K.; et al. Random Forest as a non-parametric algorithm for near-infrared (NIR) spectroscopic discrimination for geographical origin of agricultural samples. *Bull. Korean Chem. Soc*. **2012**, 33, 4267–4270.
32. Kong, W.; Zhang, C.; Liu, F.; Nie, P.; He, Y. Rice seed cultivar identifcation using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors*. **2013**, 13, 8916–8927.
33. Chen, H.; Tan, C.; Lin, Z. Authenticity detection of black rice by near-infrared spectroscopy and support vector data description. *Int. J. Anal. Chem.* **2018**, 2018, 8032831.
34. Cui, C.; Fearn, T. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometr Intell. Lab. Syst.* **2018**, 182, 9–20.
35. Das, B.; Nair, B.; Reddy, V.K. et al. Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. *Int. J. Biometeorol*. **2018,** 62, 1809–1822.
36. Le, T.H.; Chen, H.; Babar, M.A. Deep learning for source code modeling and generation: models, applications, and challenges. *ACM Comput. Surv.* **2020**, 53, 62–100.
37. Sampaio, P.S., Castanho, A., Almeida, A.S. et al. Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods. *Eur. Food Res. Technol.* **2020**, 246, 527–537.
38. Kabir, M.H.; Guindo, M.L.; Chen, R.; Liu, F. Geographic Origin Discrimination of Millet Using Vis-NIR Spectroscopy Combined with Machine Learning Techniques. *Foods*. **2021**, 11, 2767.
39. Osborne, B.G.; Mertens, B.; Thompson, M.; Fearn, T. The authentication of Basmati rice using near infrared spectroscopy. *J. Near Infrared Spectrosc*. **1993**, 1, 77–83.

40. Burestan, F.N.; Sayyah, A.H.; Taghinezhad, E. Prediction of some quality properties of rice and its flour by near-infrared spectroscopy (NIRS) analysis. *Food Sci Nutr.* **2020**, 9, 1099–1105.
41. De Alencar Figueiredo, L.F.; Davrieux, F.; Fliedel, G.; Rami, J.F.; Chantereau, J.; Deu, M.; Courtois, B.; Mestres, C. Development of NIRS equations for food grain quality traits through exploitation of a core collection of cultivated sorghum. *J. Agric. Food Chem.* **2006**, 54, 8501–8509.
42. Levasseur-Garcia C. Updated overview of infrared spectroscopy methods for detecting mycotoxins on cereals (Corn, Wheat, and Barley). *Toxins.* **2018**, 10, 10010038.
43. Stubbs, T.L.; Kennedy, A.C.; Fortuna, A.M. Using NIRS to predict fiber and nutrient content of dryland cereal cultivars. *J. Agric. Food Chem.* **2010**, 58, 398–403.
44. Piaskowski, J.L.; Brown, D.; Campbell, K.G. Near-infrared calibration of soluble stem carbohydrates for predicting drought tolerance in spring wheat. *Agron J.* **2016**, 108, 285–293.
45. Norman, H.C.; Hulm, E.; Humphries, A.W. et al. Broad near-infrared spectroscopy calibrations can predict the nutritional value of >100 forage species within the Australian feedbase. *Anim. Prod. Sci.* **2020**, 60, 1111–1122.
46. Smartfood-International Year of millets. Available online: https://www.smartfood.org/international-year-of-millets-2023/millet (Accessed on 08 February 2022).
47. Sustainable Development Goal 3: https://in.one.un.org/page/sustainable-development-goals/sdg-3-2/ (Accessed on 08 February 2022).
48. Mainstreaming millets. https://pib.gov.in/PressReleasePage.aspx?PRID=1783716 (Accessed on 08 February 2022).
49. García, J.; Cozzolino, D. Use of near infrared reflectance (NIR) spectroscopy to predict chemical composition of forages in broad-based calibration models. *Agric. Téc.* **2006**, 66, 41–47.
50. Black, J.L.; Hughes, R.J.; Nielsen, S.G.; et al. Near infrared reflectance analysis of grains to estimate nutritional value for chickens. In: 20th Australian Poultry Science Symposium, Sydney, Australia, 9–11 February 2009. Poultry Research Foundation, **2009**, 31–34.
51. Tahir, M.; Shim, M.Y.; Ward, N.E.; et al. Phytate and other nutrient components of feed ingredients for poultry. *Poult. Sci.* **2012**; 91, 928–935.
52. Genebank of ICRISAT. Available online: https://www.genebank.icrisat.org (Accessed on 22 Dec 2021).
53. Upadhyaya, H.D., Pundir, R.P.S.; Dwivedi, S.L.; Gowda, C.L.L.; Reddy, V.G.; Singh, S. Developing a mini core collection of sorghum for diversified utilization of germplasm. *Crop Sci.* **2009**, 49, 1769–1780.
54. Jordan, D.R.; Mace, E.S.; Cruickshank, A.W.; Hunt, C.H.; Henzell, R.G. Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Sci.* **2011**, 51, 1444–1457.
55. Deshpande, S.; Rakshit, S.; Manasa, K.G.; Pandey, S.; Gupta, R. Genomic Approaches for Abiotic Stress Tolerance in Sorghum. In: The Sorghum Genome. Compendium of Plant Genomes. Springer International Publishing. **2016,** 169–187.
56. Kassahun, B.; Bidinger, F.R.; Hash, C.T.; Kuruvinashetti, M.S. Stay-green expression in early generation sorghum [*Sorghum bicolor* (L.) Moench] QTL introgression lines. *Euphytica*. **2010**, 172, 351–362.
57. Sehgal, D.; Skot, L.; Singh, R.; Srivastava, R.K.; Das, S.P.; Taunk, J.; Sharma, P.C., Pal, R.; Raj, B.; Hash, C.T.; Yadav, R.S. Exploring potential of pearl millet germplasm association panel for association mapping of drought tolerance traits. *PLoS One*. **2015**, 10, e0122165.
58. Handbook of Agriculture. Directorate of Publications and Information on Agriculture, by Indian Council of Agricultural Research, New Delhi. **2011**.
59. Association of Official Analytical Chemists (AOAC) International. Official Methods of Analysis (17th edition). Gaithersberg, MD, USA. **2000**.
60. Mativavarira, M., Dimes, J., Masikati, P., Van Rooyen, A.F., Mwenje, E., Sikosana, J.L.N. and Homann-Kee Tui, S. Evaluation of water productivity, stover feed quality and farmers' preferences on sweet sorghum cultivar types in the semi-arid regions of Zimbabwe. *J. SAT Agric. Res.* **2011**, 9, 9.
61. Samireddypalle, A., Boukar, O., Grings, E., Fatokun, C.A., Kodukula, P., Devulapalli, R., Okike, I. and Blümmel, M. Cowpea and groundnut haulms fodder trading and its lessons for multidimensional cowpea improvement for mixed crop livestock systems in West Africa. *Front. Plant Sci.* **2017**, 8, 30.
62. Jayawardana, S.A.S., Samarasekera, J.K.R.R., Hettiarachchi, G.H.C.M., Gooneratne, J., Mazumdar, S.D. and Banerjee, R. Dietary fibers, starch fractions and nutritional composition of finger millet varieties cultivated in Sri Lanka. *J. Food Compost. Anal.* **2019,** 82, 103249.
63. Hone Lab video. Available online: https://www.youtube.com/watch?v=c7f_p3p-SVg (Accessed on 08 February 2022).
64. Hone Create Platform. Available online: https://www.honecreate.com (Accessed on 08 February 2022).
65. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018,** 147, 70–90.
66. Fandango, A. Mastering TensorFlow 1. x: Advanced machine learning and deep learning concepts using TensorFlow 1. x and Keras. Packt Publishing Ltd. **2018**.
67. Savitzky, A.; Golay, M.J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, 36, 1627–1639.
68. Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc*. **1989**, 43, 772–777.
69. Hopkins, D.M. Using data pretreatments effectively. Seminar at International Diffuse Reflectance Conference, Chambersburg, PA. **2008**.

70. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv*. **2014**, 1412.6980.
71. Zhang, X.; Lin, T. ; Xu, J. et al. DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Anal Chim Acta*. **2019**, 1058, 48–57.
72. Assadzadeh, S.; Walker, C.K.; McDonald, L.S.; et al. Multi-task deep learning of near infrared spectra for improved grain quality trait predictions. *J Near Infrared Spectrosc*. **2020**, 28, 275–286.
73. Williams, P., Dardenne, P., & Flinn, P. Tutorial: Items to be included in a report on a near infrared spectroscopy project. Journal of Near Infrared Spectroscopy. **2017**, 25, 85–90.
74. Williams, P. Near-InfraRed Technoloy-Getting the Best Out of Light. PDK Projects Inc., **2003**, 109.
75. Williams, P. The RPD statistic: a tutorial note. *NIR News*, **2014**, 25, 22–26.
76. Asekova, S.; Han, S.I.; Choi, H.J.; et al. Determination of forage quality by near-infrared reflectance spectroscopy in soybean. *Turk J Agric For*. **2016**, 40, 45–52.
77. Dodevska MS, Djordjevic BI, Sobajic SS, et al. Characterisation of dietary fibre components in cereals and legumes used in Serbian diet. *Food Chem*. **2013**, 141, 1624–1629.
78. Belesova, K.; Gasparrini, A.; Sié, A.; et al. Household cereal crop harvest and children's nutritional status in rural Burkina Faso. *Environ Health*. **2017**, 16, 1–1.
79. Rankoana, S.A. The use of indigenous knowledge in subsistence farming: implications for sustainable agricultural production in dikgale community in Limpopo Province, South Africa. **2017**: 63–72.
80. Kumar, A.; Tomer, V.; Kaur, A.; et al. Millets: a solution to agrarian and nutritional challenges. *Agric. Food Secur.* **2018**, 7, 1–5.
81. Yang, Z.; Han, L.; Li, Q.; et al. Discriminant analysis of meat and bone meal content in ruminant feed based on NIRS. *Trans. Chin. Soc. Agric. Eng*. **2009**, 40, 124–128.
82. Hart, J.R.; Norris, K.H.; Golumbic, C. Determination of the moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. *Cereal Chem*. **1962**, 39, 94–99.
83. Cozzolino, D. An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Res. Int.* **2014**, 60, 262–265.
84. Yang, X.S.; Wang, L.L.; Zhou, X.R.; et al. Determination of protein, fat, starch, and amino acids in foxtail millet [Setaria italica (L.) Beauv.] by Fourier transform near-infrared reflectance spectroscopy. *Food Sci. Biotechnol.* **2013**, 22, 1495–1500.
85. Bhardwaj, R.; Yadav, S.; Suneja, P. NIRS based food quality assessment approaches for cereals, oilseeds, pulses, fruits and vegetables. In: 7th Indo-Global Summit and Expo on Food & Beverages, 8–10 October 2015, New Delhi, India. **2015**.
86. Diao, X. Production and genetic improvement of minor cereals in China. *Crop J*. **2017**, 5, 103–14.
87. McKevith, B. Nutritional aspects of cereals. *Nutr. Bull*. **2004**, 29, 111–142.
88. Girish, C.; Meena, R.K.; Mahima, D.; Mamta, K. Nutritional properties of minor millets: neglected cereals with potentials to combat malnutrition. *Curr. Sci*. **2014**, 107, 1109–1111.
89. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science*. **2015**, 349, 255–260.
90. Le, T.H.; Chen, H.; Babar, M.A. Deep learning for source code modeling and generation: models, applications, and challenges. *ACM Comput. Surv.* **2020**, 53, 62–100.
91. Huang, Z., Sha, S., Rong, Z., Chen, J., He, Q., Khan, D.M. and Zhu, S. Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed. *Ind. Crops Prod.* **2013**, 43, 654–660.
92. Wheat trading standards in Australia. Available online: https://www.graintrade.org.au/commodity_standards (Accessed on 08 February 2022).
93. Wheat quality and markets in Queensland, Department of Agriculture and Fisheries, Queensland. Available online: https://www.daf.qld.gov.au/__data/assets/pdf_file/0006/53799/Wheat-FactSheet-Quality-Markets-Qld.pdf (Accessed on 08 February 2022).
94. Pojić, M.M.; Mastilović, J.S. Near infrared spectroscopy—advanced analytical tool in wheat breeding, trade, and processing. *Food Bioprocess Technol*. **2013**, 6, 330–352.
95. Baeten, V.; Pierna, J.F.; Vermeulen, P.; et al. Performance comparison of bench-top, hyperspectral imaging and pocket near infrared spectrometers: the example of protein quantification in wheat flour. In: Engelsen SB, Sørensen KM and van den Berg F (eds) Proceedings of the 18th International Conference on Near Infrared Spectroscopy. Chichester: IM Publications Open. **2019**, 151–155.
96. Almendingen, K.; Meltzer, H.M.; Pedersen, J.I.; et al. Near infrared spectroscopy—a potentially useful method for rapid determination of fat and protein content in homogenized diets. *Eur. J. Clin. Nutr.* **2000**, 54, 20–23.
97. Bagchi, T.B.; Sharma, S.; Chattopadhyay, K. Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. *Food Chem*. **2016**, 191, 21–27.
98. Magwaza, L.S.; Naidoo, S.I.; Laurie, S.M.; et al. Development of NIRS models for rapid quantification of protein content in sweetpotato [*Ipomoea batatas* (L.) LAM.]. *LWT Food Sci. Technol.* **2016**, 72, 63–70.
99. Rosales, A.; Galicia, L.; Oviedo, E.; et al. Near-infrared reflectance spectroscopy (NIRS) for protein, tryptophan, and lysine evaluation in quality protein maize (QPM) breeding programs. *J. Agric. Food. Chem.* **2011**, 59, 10781–10786.
100. Wang, H.L.; Wan, X.Y.; Bi, J.C.; et al. Quantitative analysis of fat content in rice by near-infrared spectroscopy technique. *Cereal chem*. **2006**, 83, 402–406.

Supplementary Table S1: List of 328 grain samples used in the study along with the crop species, genotype ID and protein content (%, [g.100g^{-1}]) obtained from laboratory analysis.

S. no	Crop species	Genotype	Protein (%, [g.100g^{-1}])
1	Finger millet	IE 2043	7.21
2	Finger millet	IE 2296	7.65
3	Finger millet	IE 2572	7.37
4	Finger millet	IE 2606	9.59
5	Finger millet	IE 2790	6.76
6	Finger millet	IE 3077	6.76
7	Finger millet	IE 3470	7.61
8	Finger millet	IE 3475	7.54
9	Finger millet	IE 3614	5.99
10	Finger millet	IE 3618	7.58
11	Finger millet	IE 4057	8.01
12	Finger millet	IE 4073	7.39
13	Finger millet	IE 4115	7.55
14	Finger millet	IE 4121	8.78
15	Finger millet	IE 4671	9.07
16	Finger millet	IE 5066	9.50
17	Finger millet	IE 5106	8.88
18	Finger millet	IE 5165	8.65
19	Finger millet	IE 518	7.62
20	Finger millet	IE 5367	9.00
21	Foxtail millet	ISe 1251	10.96
22	Foxtail millet	ISe 1454	11.08
23	Foxtail millet	ISe 1468	9.08
24	Foxtail millet	ISe 1511	11.13
25	Foxtail millet	ISe 1664	12.39
26	Foxtail millet	ISe 1805	12.45
27	Foxtail millet	ISe 1881	12.66
28	Foxtail millet	ISe 1892	12.91
29	Foxtail millet	ISe 238	11.92
30	Foxtail millet	ISe 289	10.39
31	Foxtail millet	ISe 480	10.86
32	Foxtail millet	ISe 525	10.52
33	Foxtail millet	ISe 719	11.89
34	Foxtail millet	ISe 783	13.42
35	Foxtail millet	ISe 796	11.65
36	Foxtail millet	ISe 827	13.17
37	Foxtail millet	ISe 828	11.40
38	Foxtail millet	ISe 840	9.89
39	Foxtail millet	ISe 869	10.82
40	Maize	783527	9.34
41	Maize	4695575	9.97
42	Maize	9424780	9.62
43	Maize	18270413	9.04

44	Maize	22525674	9.10
45	Maize	900MG	8.98
46	Maize	X35D602	8.76
47	Maize	X35D612	8.53
48	Maize	X35D620	8.81
49	Maize	X35F833	9.21
50	Pearl millet	9444	12.53
51	Pearl millet	PUSA 322	9.69
52	Pearl millet	86 M 86	14.41
53	Pearl millet	86 M 88	11.42
54	Pearl millet	863B	11.65
55	Pearl millet	863B-P2	15.19
56	Pearl millet	APH 45	10.45
57	Pearl millet	Bio 451	10.29
58	Pearl millet	Bio 549	10.61
59	Pearl millet	BLMPH 105	10.16
60	Pearl millet	GB8735	10.87
61	Pearl millet	ICMV 93191	16.12
62	Pearl millet	GK 1183	11.75
63	Pearl millet	GK 1207	10.41
64	Pearl millet	GK 1235	11.02
65	Pearl millet	H77/833-2	9.76
66	Pearl millet	HT 416628	12.31
67	Pearl millet	HYMH 5	12.67
68	Pearl millet	HYMH 8	12.24
69	Pearl millet	ICMB 89111-P6	15.00
70	Pearl millet	ICMB 90111-P2	15.54
71	Pearl millet	ICMB 90111-P6	13.45
72	Pearl millet	ICML 22	14.72
73	Pearl millet	ICMP 451-P6	15.72
74	Pearl millet	ICMP 451-P8	14.97
75	Pearl millet	ICMS 7703	21.51
76	Pearl millet	ICMS 7704	19.84
77	Pearl millet	ICMV 155	15.05
78	Pearl millet	ICMV 221	19.04
79	Pearl millet	ICMV-IS 92222	21.16
80	Pearl millet	IP 10085	18.74
81	Pearl millet	IP 10394	15.42
82	Pearl millet	IP 10446	16.11
83	Pearl millet	IP 10539	15.33
84	Pearl millet	IP 10705	16.08
85	Pearl millet	IP 10761	14.73
86	Pearl millet	IP 10811	15.94
87	Pearl millet	IP 10953	13.54
88	Pearl millet	IP 11211	17.96
89	Pearl millet	IP 11275	16.95
90	Pearl millet	IP 11577	19.20

91	Pearl millet	IP 12116	15.82
92	Pearl millet	IP 12322	19.77
93	Pearl millet	IP 12840	18.46
94	Pearl millet	IP 13384	16.13
95	Pearl millet	IP 13964	18.86
96	Pearl millet	IP 15551	16.91
97	Pearl millet	IP 15946	18.04
98	Pearl millet	IP 16082	16.09
99	Pearl millet	IP 16096	17.03
100	Pearl millet	IP 16403	18.14
101	Pearl millet	IP 17611	16.02
102	Pearl millet	IP 17632	13.49
103	Pearl millet	IP 17720	18.00
104	Pearl millet	IP 18062	17.19
105	Pearl millet	IP 18132	16.65
106	Pearl millet	IP 18168	17.32
107	Pearl millet	IP 18293-P152	18.50
108	Pearl millet	IP 19386	15.97
109	Pearl millet	IP 19388	16.25
110	Pearl millet	IP 19405	14.59
111	Pearl millet	IP 19448	16.89
112	Pearl millet	IP 21517	15.34
113	Pearl millet	IP 22423	14.79
114	Pearl millet	IP 22424	14.57
115	Pearl millet	IP 22455	17.43
116	Pearl millet	IP 3108	18.05
117	Pearl millet	IP 3125	14.12
118	Pearl millet	IP 3175	17.14
119	Pearl millet	IP 3509	17.44
120	Pearl millet	IP 3616	16.10
121	Pearl millet	IP 3732	17.46
122	Pearl millet	IP 4020	21.46
123	Pearl millet	IP 4927	18.72
124	Pearl millet	IP 4979	14.92
125	Pearl millet	IP 5207	19.00
126	Pearl millet	IP 5253	12.76
127	Pearl millet	IP 5713	16.18
128	Pearl millet	IP 5923	17.39
129	Pearl millet	IP 6060	17.51
130	Pearl millet	IP 6102	18.34
131	Pearl millet	IP 6110	19.48
132	Pearl millet	IP 6112	16.76
133	Pearl millet	IP 6146	18.34
134	Pearl millet	IP 6179	17.86
135	Pearl millet	IP 6310	16.99
136	Pearl millet	IP 6460	18.65
137	Pearl millet	IP 6682	16.39

138	Pearl millet	IP 6769	16.41
139	Pearl millet	IP 6891	20.14
140	Pearl millet	IP 7470	16.50
141	Pearl millet	IP 7633	16.91
142	Pearl millet	IP 7762	16.45
143	Pearl millet	IP 7941	16.51
144	Pearl millet	IP 7970	15.79
145	Pearl millet	IP 8129	17.50
146	Pearl millet	IP 8166	14.77
147	Pearl millet	IP 8198	16.09
148	Pearl millet	IP 8210	15.88
149	Pearl millet	IP 8276	15.37
150	Pearl millet	IP 8426	17.13
151	Pearl millet	IP 8761	19.07
152	Pearl millet	IP 8767	17.37
153	Pearl millet	IP 8786	19.55
154	Pearl millet	IP 8972	16.53
155	Pearl millet	IP 9282	17.34
156	Pearl millet	IP 9347	17.27
157	Pearl millet	IP 9407	19.43
158	Pearl millet	IP 9426	20.31
159	Pearl millet	IP 9446	15.57
160	Pearl millet	IP 9651	15.56
161	Pearl millet	IP 9692	17.98
162	Pearl millet	IP 9710	15.67
163	Pearl millet	IP 9840	16.63
164	Pearl millet	IP 9854	15.93
165	Pearl millet	J 104	19.42
166	Pearl millet	JKBH 1352	9.72
167	Pearl millet	JKBH 1490	13.44
168	Pearl millet	KH 3022	11.19
169	Pearl millet	NBH 5863	10.48
170	Pearl millet	NU 399	10.62
171	Pearl millet	NU 409	10.65
172	Pearl millet	PRLT	10.09
173	Pearl millet	Super Boss	11.91
174	Pearl millet	Tift 383	15.97
175	Sorghum	Keninkeni	10.62
176	Sorghum	00-CZ-F5P-135	12.55
177	Sorghum	01-BE-F5P-15	12.07
178	Sorghum	02-SB-F4DT-275	13.09
179	Sorghum	296B	10.48
180	Sorghum	98-BE-F5P-84	11.67
181	Sorghum	AKSV BR (PKV KRANTI)	11.73
182	Sorghum	B2-3	9.04
183	Sorghum	B2-5	10.69
184	Sorghum	B35	11.07

doi:10.20944/preprints202202.0335.v1

185	Sorghum	BJV44	9.61
186	Sorghum	BTx623	10.62
187	Sorghum	C41-28-49-11(D3)	12.67
188	Sorghum	C41-28-49-14(D1)	11.18
189	Sorghum	C41-28-49-18-7	16.55
190	Sorghum	C41-28-49-18-9	13.10
191	Sorghum	C41-28-49-27-1	13.67
192	Sorghum	C41-28-52-13-5	11.50
193	Sorghum	C41-28-52-18(D3)	10.98
194	Sorghum	C41-28-52-26-1	12.93
195	Sorghum	C41-28-52-28-2	12.28
196	Sorghum	C41-28-52-28-3	12.21
197	Sorghum	C41-28-52-28-5	12.20
198	Sorghum	C41-28-75-21(D1)	13.32
199	Sorghum	C41-28-75-26-4	12.08
200	Sorghum	CIRAD406	12.00
201	Sorghum	CMDT45	11.42
202	Sorghum	CRS4	10.42
203	Sorghum	CSH16	11.25
204	Sorghum	CSM388	14.46
205	Sorghum	CSM63-E	11.46
206	Sorghum	CSV 14R	11.91
207	Sorghum	CSV 18	13.42
208	Sorghum	CSV 216R	9.23
209	Sorghum	CSV 26	13.01
210	Sorghum	CSV22	13.68
211	Sorghum	CSV29R	13.24
212	Sorghum	Doua-G	8.79
213	Sorghum	E36-1	10.09
214	Sorghum	E36-1	12.04
215	Sorghum	Framida	12.64
216	Sorghum	Gnossiconi	10.74
217	Sorghum	GPN01 S01 266-2-1-6-vr	8.68
218	Sorghum	GPN01 S01 267-9-3-1-1	10.51
219	Sorghum	GPN01 S01 267-9-3-3-vr	9.71
220	Sorghum	GRS1 (DSV5)	10.58
221	Sorghum	GS15-10	11.21
222	Sorghum	GS23	10.59
223	Sorghum	ICSB 370-2-9	10.38
224	Sorghum	ICSV745	10.65
225	Sorghum	ICSV93046-P1	10.29
226	Sorghum	IS 41397-3-P6	11.02
227	Sorghum	IS 8219-P1	9.34
228	Sorghum	IS10876	12.45
229	Sorghum	IS11026	17.90
230	Sorghum	IS11473	17.01
231	Sorghum	IS11919	17.25

232	Sorghum	IS12804	14.11
233	Sorghum	IS12883	13.43
234	Sorghum	IS12965	17.06
235	Sorghum	IS13893	13.54
236	Sorghum	IS14556	11.82
237	Sorghum	IS14779	14.05
238	Sorghum	IS15401	9.65
239	Sorghum	IS15466	11.41
240	Sorghum	IS15945	14.31
241	Sorghum	IS16528	15.58
242	Sorghum	IS25249	17.78
243	Sorghum	IS25548	16.24
244	Sorghum	IS25910	17.85
245	Sorghum	IS25989	18.22
246	Sorghum	IS26046	16.54
247	Sorghum	IS26222	16.05
248	Sorghum	IS26617	15.34
249	Sorghum	IS26694	14.40
250	Sorghum	IS26701	13.69
251	Sorghum	IS26737	14.44
252	Sorghum	IS27557	17.65
253	Sorghum	IS27786	17.10
254	Sorghum	IS27887	15.04
255	Sorghum	IS27912	14.45
256	Sorghum	IS28141	14.82
257	Sorghum	IS28313	17.35
258	Sorghum	IS28389	17.01
259	Sorghum	IS28449	17.40
260	Sorghum	IS28614	18.38
261	Sorghum	IS28747	17.88
262	Sorghum	IS28849	15.85
263	Sorghum	IS29091	12.29
264	Sorghum	IS29100	14.85
265	Sorghum	IS29187	12.50
266	Sorghum	IS29304	13.86
267	Sorghum	IS29314	14.64
268	Sorghum	IS29472	11.47
269	Sorghum	IS29568	17.50
270	Sorghum	IS29606	14.16
271	Sorghum	IS29689	15.39
272	Sorghum	IS30231	15.04
273	Sorghum	IS30460	12.38
274	Sorghum	IS30507	15.59
275	Sorghum	IS30572	16.80
276	Sorghum	IS30838	14.32
277	Sorghum	IS31186	12.59
278	Sorghum	IS31446	15.59

279	Sorghum	IS31681	16.64
280	Sorghum	IS31706	16.06
281	Sorghum	IS32787	13.46
282	Sorghum	IS33023	15.97
283	Sorghum	IS33090	14.84
284	Sorghum	IS393(411)695	10.93
285	Sorghum	IS7957	13.86
286	Sorghum	IS8012	13.09
287	Sorghum	IS9113	14.66
288	Sorghum	*Sevata jonna* (landrace)	12.66
289	Sorghum	M35-1	9.69
290	Sorghum	M35-1	10.75
291	Sorghum	M35-1	11.55
292	Sorghum	N13	10.26
293	Sorghum	Parbahani Moti	10.54
294	Sorghum	Parbhani Jyothi	10.62
295	Sorghum	PB15220-1	10.61
296	Sorghum	PB15881-3	10.45
297	Sorghum	Phule Anuradah (RSV 458)	13.79
298	Sorghum	Phule Chitra (SPV 1546)	13.36
299	Sorghum	Phule Maulee (RSLG262)	10.48
300	Sorghum	Phule Maulee	12.82
301	Sorghum	Phule Revati (RSV1006)	13.62
302	Sorghum	Phule Vasudha	17.99
303	Sorghum	Phule Vasudha	9.97
304	Sorghum	PVK 801-P23	11.75
305	Sorghum	R16	10.87
306	Sorghum	R16	16.39
307	Sorghum	R37-13-11-2-21-2	12.12
308	Sorghum	R37-13-11-2-26(D3)	13.35
309	Sorghum	R37-13-11-2-30(D2)	14.06
310	Sorghum	R37-13-11-2-40	13.34
311	Sorghum	R37-13-11-2-40(D1)	14.25
312	Sorghum	R37-13-11-2-40(D3)	14.08
313	Sorghum	R37-13-11-2-5-1	14.34
314	Sorghum	R37-13-30-1(D6)	12.60
315	Sorghum	R37-13-30-1(D7)	12.62
316	Sorghum	R37-13-30-11-4	11.91
317	Sorghum	R37-13-30-11-6	11.93
318	Sorghum	R37-13-30-15(D3)	12.56
319	Sorghum	R37-13-30-16(D2)	11.62
320	Sorghum	R37-13-30-28(D3)	15.09
321	Sorghum	Ribdahu	10.53
322	Sorghum	RSLG262	13.59
323	Sorghum	S35	11.65
324	Sorghum	S35	15.17
325	Sorghum	Sambalma	9.83

326	Sorghum	SP 2417-P3	10.41
327	Sorghum	SPV2217	10.43
328	Sorghum	SVD806	10.32

Supplementary Table S2: Descriptive statistics presenting the variability and range of protein content in the calibration and validation sets used in the study. [Legend: SD = standard deviation; CV% = coefficient of variation].

Details	Training set	Validation set
Number of samples	262	66
Range of protein content (%, [g.100g^{-1}])	5.99–21.51	6.76–20.14
Average (%, [g.100g^{-1}])	13.67	13.27
SD	3.25	3.37
CV%	23.79	25.36

Supplementary Table S3: Descriptive statistics [minimum (min), maximum (max), average (avg), standard deviation (SD), and standard error (SE)] of the protein content in each of the cereal species of the calibration and validation sets used in the study.

Set	n	Min	Max	Avg	SD	SE
Calibration	**262**	**5.99**	**21.51**	**13.67**	**3.25**	**0.20**
Finger millet	16	5.99	9.59	7.95	0.94	0.24
Foxtail millet	16	9.08	13.42	11.49	1.16	0.29
Maize	7	8.76	9.97	9.23	0.44	0.17
Pearl millet	99	9.69	21.51	15.88	2.74	0.28
Sorghum	124	8.68	18.38	13.17	2.45	0.22
Validation	**66**	**6.76**	**20.14**	**13.27**	**3.37**	**0.41**
Finger millet	4	6.76	9.07	7.84	1.26	0.63
Foxtail millet	3	10.52	13.17	11.59	1.40	0.81
Maize	3	8.53	9.21	8.91	0.35	0.20
Pearl millet	26	9.76	20.14	15.38	3.23	0.63
Sorghum	30	8.79	17.40	12.77	2.35	0.43
Total	**328**	**5.99**	**21.51**	**13.59**	**3.27**	**0.18**