RESEARCH ARTICLE Legume Research- An International Journal, Volume 46 Issue 2: 251-256 (February 2023)

Comparing Various Machine Learning Algorithms for Sugar Prediction in Chickpea using Near-infrared Spectroscopy

Madhu Bala Priyadarshi[1], Anu Sharma[2], K.K. Chaturvedi[2], Rakesh Bhardwaj[1], S.B. Lal[2], M.S. Farooqi[2], Sanjeev Kumar[1], D.C. Mishra[2], Mohar Singh[1]

10.18805/LR-4931

ABSTRACT

Background: Chickpea is the third major pulse produced globally, with 11.6 million tonnes produced per annum (Merga and Haji, 2019). Sugar alcohols, inulin, starch are all prebiotic carbohydrates found in chickpeas (Johnson *et al*., 2020). Near-Infrared (NIR) spectroscopy is a non-destructive, versatile and powerful analytical technique.

Methods: Spectral data obtained from NIR spectroscopy requires application of various techniques to extract useful information from spectral data which is further used for building various models for prediction of physical or chemical components presents in agricultural crops. The main aim of this study is to apply various machine learning algorithms especially effective in predicting sugar concentration in chickpea. Sugar prediction models are developed using Linear Regression (LR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR) and Decision Tree Regression (DTR) algorithms. Performance of the models is evaluated using measures namely, Root Mean Square Error (RMSE), Residual Standard Error (RSE), Coefficient of Determination (R^2) and Adjusted Coefficient of Determination (adjusted R^2).

Result: It was observed that, RF outperformed all other models in terms of accuracy for predicting sugar component from preprocessed spectra, with RMSE, RSE, R2 and adjusted R2 values of 0.054, 0.062, 0.954 and 0.937, respectively. The accuracy of the ANN model is similar to that of the RF, with minor differences in RMSE, RSE, R2 and adjusted R2, values of 0.057, 0.067, 0.952 and 0.935.

Key words: Algorithm, Artificial neural network (ANN), Chickpea, Machine learning, Near-infrared spectroscopy, Random forest (RF), Spectroscopy.

INTRODUCTION

Legumes are the most important crops due to their nutritional qualities. Legumes' seeds and powder are high in protein, carbohydrates, vitamins and minerals and dietary fiber (Baljeet *et al.,* 2014). Chickpea is the third-largest produced pulse in the world, with 11.6 million tonnes produced per annum, 80% of which is desi and the remaining 20% is kabuli (Merga and Haji, 2019). Near-infrared (NIR) spectroscopy is used to rapidly characterize quality parameters in desi chickpea flour using Partial Least Square Regression (PLSR) to determine protein, carbohydrate, fat and moisture concentrations of chickpea (Kamboj *et al.,* 2016). It is used to quantify the reflection rates of infrared light radiation within a sample, which is then used to estimate the chemical contents of the sample using machine learning modeling techniques. Food adulteration, authenticity control, the assessment of physicochemical qualities, rheological, or technological properties have all been cited as successful applications of NIR technology in analytical instrumentation and quality control (Porep *et al*., 2015).

In order to extract quantitative data from NIR spectra utilising a variety of wavelengths, numerous approaches have been devised. The most used calibration techniques for NIR spectroscopy are LR and PLSR. These methods, which avoid co-linearity problems, are ideally suited to situations needing information from a large number of wavelengths and may therefore be used when the number of variables exceeds the number of available samples (Pasquini, 2003). Non-linear algorithms such as ANN, RF, SVR and DTR are emerging as a viable alternative to LR and PLSR for near-infrared calibration. These techniques assume a linear relationship between the spectral data and the quantitative value to be assessed. Although not as often employed, these models have potential since they might provide greater results in particular circumstances (Pasquini, 2003).

This study was undertaken with an objective of applying (i) NIR reflectance spectroscopy, best techniques, variable selection techniques, extracting the wavelengths (750-2500 nm) associated with best-predicting sugar in chickpea,

[1]ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi-110 012, India.
[2]ICAR-Indian Agricultural Statistics Research Institute, Pusa Campus, New Delhi-110 012, India.

Corresponding Author: Madhu Bala Priyadarshi, ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi-110 012, India. Email:

How to cite this article: Priyadarshi, M.B., Sharma, A., Chaturvedi, K.K., Bhardwaj, R., Lal, S.B., Farooqi, M.S., Kumar, S., Mishra, D.C. and Singh, M. (2023). Comparing Various Machine Learning Algorithms for Sugar Prediction in Chickpea using Near-Infrared Spectroscopy. Legume Research. 46(2): 251-256. doi: 10.18805/LR-4931.

Submitted: 02-04-2022 **Accepted:** 03-10-2022 **Online:** 02-11-2022

(ii) Establish and compare five different machine learning prediction models of sugar in chickpea and the wavelength of choice germplasm flour within acceptable agreement with chemical laboratory methods. The machine learning methods considered in this study are LR, ANN, RF, SVR and DTR to predict the concentration of sugar in chickpea germplasm flour.

MATERIALS AND METHODS

Sample collection

A total of 237 chickpea germplasm accessions were chosen from the National Gene Bank of ICAR-National Bureau of Plant Genetic Resources (NBPGR), New Delhi, to construct NIR models for predicting sugar concentration in chickpea flour.

Collection of spectral reflectance data

The near-infrared spectroscopic examination used a near-infrared scanning monochromator in reflectance mode. Chickpea samples were homogenized in a Foss Cyclotec mill with a 0.5 mm filter to guarantee uniform particle size. Homogenized flour was placed in a circular cuvvete with a glass window and slightly squeezed with the back cover to produce consistent packing. At a 2 nm spacing, spectra from the wavelength range 750-2500 nm were captured using a Foss NIRS 6500 cuvette spinning model. The device was tested for wavelength accuracy and repeatability after a 30-min warm-up period. The instrument was calibrated against white mica each time the sample was scanned. The average spectrum was recorded after scanning the material 32 times. The reflectance logarithm (log 1/R) was used to record spectral data at 2 nm intervals in the 400-2498 nm wavelength range. The concentration of sugar in chickpea was measured in the chemical laboratory, which served as reference data for training and measuring the performance of prediction models.

Spectra preprocessing

Interval partial least squares (*i*PLS) is used to select a spectral interval that is particularly informative with respect to the parameter under consideration (Norgaard *et al*., 2000). This method frequently improves prediction ability when compared to standard full-spectrum PLS models (Borin and Poppi, 2005; Norgaard *et al*., 2000; Winning, *et al*, 2007). The *i*PLS algorithm was used by dividing the full spectra into 15 non-overlapping intervals of equal size. Mdatools package of R was used to run the *i*PLS 2.1 routine (Norgaard, 2005) developed by the Royal Veterinary and Agricultural University of Denmark. The best interval was chosen based on the root-mean-square error of cross-validation (RMSECV), which is defined as:

$$RMSECV = \sqrt{\frac{1}{N_{cal}} \sum_{n=1}^{N_{cal}} (y_{cal,n} - \hat{y}_{cal,n})^2}$$

Where;

$y_{cal,n}$ is the reference value of the parameter being considered for the n^{th} sample of the calibration set (N_{cal}=237 samples). The predicted value $y_{cal,n}$ is obtained by removing the n^{th} sample from the calibration set, building a model with the rem aining samples and applying this model to the sample that was removed.

Bala and colleagues noted in the literature in 2022 that several agricultural and food products had been assigned NIR bands in the ranges of 1400-1600 nm and 2000-2350 nm. According to Kasemsumran *et al*., (2004), the most significant wavelength range for the presence of glucose is between 1420 and 1480 nm and 1630 and 1730 nm.

Chemometric analysis

The steps involved in developing models are (1) outlier detection, (2) data preprocessing, (3) dimensionality reduction, (4) model development, and (5) performance estimation of developed models. The flowchart in Fig 1 depicts the model development process.

Outlier detection and preprocessing

Using box plots, outliers were identified and deleted. Prior to the development of the model, spectral datasets were preprocessed. To maximise the calibration results, several mathematical treatments using the raw spectrum data were used, with various combinations of smoothing and gap size. For example, in 2,2,2,1, the first number indicates the order of the derivative function (two is the second derivative), the second number is the gap (the length in nm) in data points over which the derivative is calculated, the third number is the number of data points (segment length) used in first smoothing, and the fourth number is the number of data points in second smoothing, which is normally set at 1 if no second smoothing is used (Shenk and Westerhaus, 1993). The optimal combination of data preprocessing was selected as the one providing a pls model with a good compromise of a low RMSE and high R^2 value.

Multicollinearity

This is an extreme case in which collinearity exists between three or more variables despite the fact that no two variables

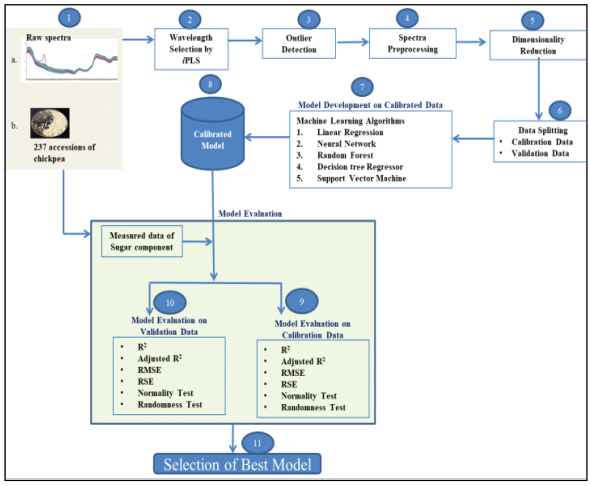

Fig 1: Flowchart of model development.

have a particularly strong correlation Statistics are less reliable due to multicollinearity. Correlation between variables is used to detect multicollinearity.

Model development

Machine learning algorithms, LR, ANN, RF, SVR and DTR were used to develop prediction models utilizing preprocessed spectra. All model development packages were obtained from the Comprehensive R Archive Network (CRAN). To create a linear regression model, it employs the *lm()* (Chambers, 1992) function. To develop neural networks models, the *neuralnet()* package from the Comprehensive R Archive Network (CRAN) has been used. The network employs *Tanh* and *sigmoid* activation functions. In this study, "*randomForest()"* function from the R package *"randomForest"* is used for models of random forest algorithm and "*svm()"* function from the R package "*e1071*". (SVM in R for Data Classification using e1071 Package, 2021) is used for SVR algorithm. For models of decision tree regression, *rpart()* function from the R package "*Recursive Partitioning and Regression Trees (RPART)*" (Breiman, 1984) is used.

Model evaluation

To evaluate the efficacy of the regression model, the RMSE, RSE, R^2 and adjusted R^2 statistical measures were used. The RMSE is the average difference between measured and predicted values (Kobayshi and Salam, 2000). Lower RMSE values improve the model's predictive ability: the lower the RMSE value for a model, the better the model's prediction ability.

Residual analysis is used to assess the suitability of a regression model. As the RSE decreases, the fit of a regression model to a dataset improves. However, the greater the RSE, the worse a regression model fits a dataset. To examine the randomness of residuals, a residual plot is created and evaluated. The Shapiro-Wilk (SW) test, was used to perform a normality test with a significance level of 0.05 for residues from all models. For predicting the result of a particular event and to measure changes in one variable due to differences in another variable, R^2 is calculated. R^2 measures the strength of a linear relationship between two variables. The optimum model for each component was chosen in this investigation based on the lowest prediction RMSE and RSE and the highest R^2 and adjusted R^2 value between measured and predicted values.

Model validation

With the help of validation data, each developed model was tested. With validation data, Table 3 shows evaluation statistics for each of the five models. The adjusted R^2 and RMSE values show that the model is capable of accurately predicting the sugar component.

RESULTS AND DISCUSSION

The descriptive statistics including mean, standard deviation (SD) for the sugar component of chickpea samples is shown in Table 1.

Out of 237 samples, it was discovered that 184 samples had reflectance for sugar component that could be studied in NIR spectra. In 184 samples, the mean value is 7.01. Since we had a small sample size, determining the distribution of the sugar component was important. An analysis using the Shapiro-Wilk test revealed that the distribution of the sugar component considerably deviated from normality (W = 0.96, p-value 0.001).

Table 2 shows effective wavelength ranges for sugar prediction. RMSECV and R^2 value was calculated on these wavelength ranges. The range of lower RMSECV and highest R^2 which is 1800-1938 is considered for developing models for sugar component. This range is characterised by O-H, N-H combinations.

Boxplots were used to detect outliers. Preprocessing on a specific wavelength range of 1800-1938 nm using various mathematical techniques led to the development of the models. Model developed with treatment 1,2,0,2 which means spectra is passed to first derivative with gap size of 2 nm. It is smoothed by moving average with 2 nm gap size. In addition to that spectra is passed to standard normal variate for scatter correction. There were 181 samples for 69 wavelengths after preprocessing.

Multicollinearity was detected and dimensionality reduction was done using PCA.

- A correlation matrix was created for NIR spectral data with 181 samples for 69 predictor wavelengths. It was found that the 44 predictor variables correlate more significant than 0.9 per cent indicating multicollinearity in data.
- The preprocessed spectra was decomposed into latent vectors that are ranked according to the amount of spectral variance explained by PCA. The first ten principal components (PCs) account for approximately 90% of the variation in NIR spectra of chickpea samples.

Prior to model development, data was split into two sets: a calibration set made up of 80% of the data (145 samples),

Table 1: Descriptive statistics of sugar component.

	Sugar
Mean	7.01
Standard error	0.18
Median	7.42
Mode	3.26
Standard deviation	2.45
Sample variance	6.02
Kurtosis	-0.92
Skewness	-0.34
Range	10.02
Minimum	2.24
Maximum	12.26
Sum	1289.50
Count	184
Normality (p-value)[a]	< 0.001

[a]Shapiro-Wilks test of normality was used to determine the normality of the data.

Table 2: Effective wavelength range for sugar prediction by *i*PLS.

Wavelength Index	Wavelength range	No. of wavelengths	RMSECV	R^2
701-770	1800-1938	71	3.40	0.120
281-350	960-1098	71	3.40	0.120
1-70	400-538	71	3.41	0.113
911-980	2220-2358	71	3.42	0.108
771-840	1940-2078	71	3.44	0.101
841-910	2080-2218	71	3.46	0.091

and a validation set made up of 20% of the data (36 samples). The calibration data set, which is split into two parts: training and testing data, contains 75% of the data (109 samples) and 25% of the data (36 samples) respectively. Model development take place with 10 PCs from PCA. The model generation process for each of the five algorithms produced a distinctive and durable model.

Predictor and response variables are connected by an equation in linear regression model, where the exponent (power) of both of these variables is 1. When plotted as a graph, a linear connection mathematically depicts a straight line. The general mathematical equation for a linear regression is:

$$y= ax+b$$

y= Response variable.
x= Predictor variable.
a, b= Constants.
Mathematical equation for LR model is given in following equation:

$$y= 0.914-0.734 \times 1+0.072 \times 2-0.006 \times 3-0.048 \times 4-0.008 \times 5+0.013 \times 6 +0.036 \times 7-0.111 \times 8-0.058 \times 9+0.039 \times 10 \quad .....(1)$$

Where;
y= Sugar component.
xi = 10 PCs from PCA.

Large-scale regression trees are combined by the RF algorithm. The random forest model has two parameters: ntree, which is equal to 500 trees and mtry, which is the number of input variables per node, which is 3.

Neurons are arranged in a neural network's three layers-input, hidden, and output. Equation 2 represents the neural network, where W stands for the weights vector, X for the inputs vector, and b for the bias. Equation 3 contains the sigmoid activation function that is applied.

$$y= \sum_{i=1}^{n}(wi*Xi+b) \quad .....(2)$$

$$f(z)= \frac{1}{1-e^{-y}} \quad .....(3)$$

Where,

f(z) and y= An activation function. The activation function's output ranges from 0 to 1. Here, the neural network was configured to The parameters of the neural network were set to 2, 8, 0.02, 0.3, 1500 for the hidden layers, nodes for each layer, learning rate, momentum and iteration.

Decision tree model was developed using *rpart* algorithm. There are two major processes of *rpart* (1) tree growing (2) splitting. Tree growing is expansion of the tree at specific decision points and tree pruning is to ignore the subtree with poor decision scores. To develop decision tree model, *rpart (formula= Sugar~.data= traindata)* command is used.

The accuracy of SVR models depends on how well the loss function, error penalty factor C, and SVR meta-parameters are configured. Additionally, the final models are significantly impacted by the choice of the kernel function. In this study, the commonly used radial basis kernel function (RBF), $K(x, x')= \exp(-|x-x'^2/\sigma^2)$, was used. SVR model was created with C as 1, ε as 0.1 and 0.1 for the RBF kernel parameter. Number of support vectors for SVR model is 89.

RSE values of all models lie in the range of 0.06 -0.09. To assess the randomness of residues obtained from generated models, a residual plot was created and discovered that all residues are roughly evenly distributed around zero in the plot with no apparent pattern, implying that residues are random. The Shapiro-Wilk test was used to determine the normality of residues in developed models. Table 3 displays the p-value obtained for each model. All models' p-values are greater than 0.05, indicating that the data did not show any evidence of non-normality.

It was discovered that RF and ANN prediction models perform best in the wavelength range 1800-1938 nm, to predict sugar concentration of chickpea. The model created by the RF algorithm was discovered to have the lowest RMSE value of 0.053 and RSE values of 0.062. The R^2 and adjusted R^2 values of 0.953 and 0.937, respectively, are the highest of the five models. According to (Sagar *et al.*, 2018), The ensemble-based RF approach improves model accuracy. The RF algorithm checks all variables at each node to assess how well they split due to the randomization.

The RMSE, RSE, R^2 and adjusted R^2 values of the ANN model are 0.056, 0.066, 0.951, and 0.935, respectively, with only minimal deviations from those of the RF model in terms of accuracy. Activation function gives NN their non-linearity and expressiveness. An ANN learns throughout the learning phase by changing the weights to forecast the value of the inputs' responses. It is worth noting that the tests were carried out in the lab, and the models' dependability can only be proven once they've been applied to real-world procedures. The specific performance measures for each model are shown in Table 3.

All models were assessed using validated data that included (36 samples). The accuracy of the random forest

Table 3: Comparison by performance metrics of all five models.

Model	Root mean squaree error (RMSE)		Correlation (r)		R^2		Adjusted R^2		Shapiro-wilk test on residue (p-value)		Residual standard Error (RSE)	
	Calibrated data	Validated data	Calibrated data	Validated data	Calibrated data	Validated data	Calibrated data	Validated data	Calibrated data	Validated data	Calibrated data	Validated data
Random forest (RF)	0.054	0.092	0.984	0.931	0.954	0.816	0.937	0.752	0.352	0.195	0.063	0.108
Artificial neural networks (ANN)	0.057	0.114	0.977	0.889	0.952	0.776	0.935	0.698	0.057	0.272	0.067	0.134
Support vector regression (SVR)	0.062	0.056	0.977	0.970	0.948	0.931	0.930	0.907	0.290	0.003	0.073	0.066
Decision tee regression (DTR)	0.065	0.104	0.972	0.889	0.944	0.788	0.924	0.714	0.045	0.003	0.077	0.122
Linear regression (LR)	0.082	0.096	0.962	0.897	0.910	0.798	0.880	0.728	0.658	0.022	0.096	0.114

model, measured as RMSE and RSE, was confirmed to be 0.09 and 0.108, respectively. The calculated R^2 and adjusted R^2 values were found to be 0.816 and 0.752, indicating that the RF model is effective at predicting chickpea sugar concentration in the chosen range of 1800-1938 nm.

CONCLUSION

The key conclusions are as follows: using the variable selection technique of *i*PLS regression, it was discovered that the wavelength range 1800-1938 nm is the best wavelength range from the entire range of 750-2500 nm for predicting chickpea sugar component. Five different machine learning predictive models were developed and compared in order to determine their efficiency. A comparison of these models shows that RF and ANN models have superior predictive statistics in terms of lower RMSE and RSE, as well as higher R^2 and adjusted R^2. This initiative has the potential to be scaled up to serve as a model for predicting other components in other leguminous crops. The non-destructive nature of near-infrared spectroscopy requires no or very little sample preparation. It works well as a quality control tool. It may be useful in detecting the amounts of several other components in a sample and can be used to fingerprint agricultural crops.

ACKNOWLEDGEMENT

The ICAR-Indian Agricultural Research Institute conducted this study as part of a Ph.D. programme (ICAR-IARI).

Conflict of interest: None.

REFERENCES

Bala, M., Sethi, S., Sharma, S. (2022). Prediction of maize flour adulteration in chickpea flour (*besan*) using near infrared spectroscopy. Journal of Food Science and Technology. https://doi.org/10.1007/s13197-022-05456-7.

Baljeet, S.Y., Ritika, B.Y., and Reena, K. (2014). Effect of incorporation of carrot pomace powder and germinated chickpea flour on the quality characteristics of biscuits. International Food Research Journal. 21(1): 217-222.

Borin, A., Poppi, R.J. (2005). Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. Vibrational Spectroscopy. 37: 27-32.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Wadsworth.

Chambers, J.M. and Hastie, T.J. (1992). Linear Models. Chapter 4 of Statistical Models in S. [(eds) J.M. Chambers and T.J. Hastie], Wadsworth and Brooks/Cole Advanced Books and Software. 608 pages. ISBN 0-534-16765-9.

Kamboj, U., Guha, P., and Mishra S. (2016). Characterization of chickpea flour by near Infrared Spectroscopy and Chemometrics. Analytical Letters. 50(11): 1754-1766.

Kasemsumran, S., Du, Y.P., Maruo, K. and Ozaki, Y. (2004). Comparing Various Machine Learning Algorithms for Sugar Prediction in Chickpea using Near-infrared Spectroscopy. Analytica Chimical Acta. 526-193.

Kobayshi, K., Salam, M.U. (2000). Comparing simulated and measured values using mean square deviation and its components. Agronomy Journal. 92: 345-352.

Merga, B., Haji, J. (2019). Economic importance of chickpea: Production, value and world trade. Cogent Food and Agriculture. 5: 1615718. doi: 10.1080/23311932.2019.1615718.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B. (2000). Interval partial least-squares regression (i-PLS): A comparative chemometric study with an example from near infrared spectroscopy. Applied Spectroscopy. 54: 413-419.

Norgaard, L. (2005). iPLS Toolbox Manual. Available from www.models.kvl.dk.

Pasquini, C. (2003). Near-infrared spectroscopy: Fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society. 14(2): 198-219.

Porep, J.U., Kammerer, D.R., and Carle, R. (2015). On-line application of near-infrared (NIR) spectroscopy in food production. Trends Food Science Technology. 46: 211-230.

Sagar, C., Potuganti, P. and Veetil, S. (2018). How to implement Random Forests in R. *R-bloggers*. https://www.r-bloggers.com/how-to-implement-random-forests-in-r/.

Shenk, J.S. and Westerhaus, M.O. (1993). Analysis of agriculture and food products by near-infrared reflectance spectroscopy Infrasoft International, MD, USA.

SVM in R for Data Classification using e1071 Package. (2021). TechVidvan's R tutorial series. https://techvidvan.com/tutorials/svm-in-r/.

Winning, H., Viereck, N., Norgaard, L., Larsen, J., Engelsen, S.B. (2007). Quantification of the degree of blockiness in pectins using H-1 NMR spectroscopy and chemometrics. Food Hydrocolloids. 21: 256-266.