Indian J. Plant Genet. Resour. 35(1): 44–48 (2022)
DOI 10.5958/0976-1926.2022.00007.9

RESEARCH ARTICLE

Machine Learning Algorithms for Protein Physicochemical Component Prediction Using Near Infrared Spectroscopy in Chickpea Germplasm

Madhu Bala Priyadarshi[1*], Anu Sharma[2], KK Chaturvedi[2], Rakesh Bhardwaj[1], SB Lal[2], MS Farooqi[2], Sanjeev Kumar[2], DC Mishra[2] and Mohar Singh[1]

[1]*ICAR-National Bureau of Plant Genetic Resources (NBPGR), Pusa Campus, New Delhi-110012, India*
[2]*ICAR-Indian Agricultural Statistics Research Institute (IASRI), Pusa Campus, New Delhi-110012, India*

(Received: 19 December, 2021; Revised: 19 February, 2022; Accepted: 21 February, 2022)

Prediction of physicochemical components of chickpea flour using near infrared spectroscopy requires discovering exact wavelength regions that provide the most useful data before preprocessing. This study used six essential machine learning techniques to develop models for predicting proteinphysicochemical component in chickpea: Linear Regression (LR), Artificial Neural Network (ANN), Partial Least Squares Regression (PLSR), Random Forest (RF), Support Vector Regression (SVR) and Decision Tree Regression (DTR). Performance measurements such as Root Mean Square Error and Karl Pearson's Correlation Coefficient and Coefficient of Determination were used to validate the models. RF and ANN models showed significant improvement over all other models in terms of accuracy.

Key Words: Artificial Neural Network, Chickpea, Machine learning, Near infrared spectroscopy, Random Forest, Spectroscopy

Introduction

Near infrared spectroscopy (NIS) is an efficient method for identifying and analyzing many components in a sample (Acquah *et al.,* 2016) and can be an excellent predictive germplasm evaluation procedure. Combining bands of numerous hydrogen-containing groups in moisture, protein, fat, and carbohydrate, and the vibrational information in these organic molecules are used to assess the chemical composition of samples (Batten and Berardo*,* 1998). Several techniques have been developed to extract quantitative information from Near Infra-Red (NIR) spectra using wavelengths. Some of the most used calibration methods for NIR spectroscopy are Principal Component Regression (PCR), Linear Regression (LR) and Partial Least Squares Regression (PLSR). On the other hand, Non-linear algorithms such as Artificial Neural Network (ANN), Support Vector Regression (SVR), Decision Tree Regression (DTR) and Random Forest (RF) models are not as extensively used, but they may deliver superior results when the spectral data and the quantitative value of interest have a non-linear connection (Pasquini, 2003). Such supervised predictive modelling techniques use known data to develop models capable of predicting values for future events. Selecting the most effective predictive modelling technique at the start saves considerable time.

Chickpea is rich in protein and evaluating the physicochemical components of protein in the chickpea germplasm is fundamental to identifying superior genotypes and use them further in breeding programmes. Here, we report a comparative study of six essential machine learning techniques to develop models for predicting concentrations of protein physicochemical components in chickpea germplasm with an objective to identify the best model to facilitate improved use of NIS in biochemical assay of germplasm.

Materials and Methods

Plant sample and spectral data: A random set of 237 chickpea germplasm accessions were obtained from the National Genebank, ICAR-National Bureau of Plant Genetic Resources, New Delhi. Chickpea seeds were homogenized in a Foss Cyclotec mill and the flour was transferred to a circular cuvette. Spectra in the wavelength range 400-2498 nm were captured, with 2 nm spacing, using a Foss NIRS 6500 cuvette spinning model. The instrument was calibrated against white mica each time the sample was scanned. The average spectrum was

**Author for Correspondence: Email- madhu74_nbpgr@yahoo.com*

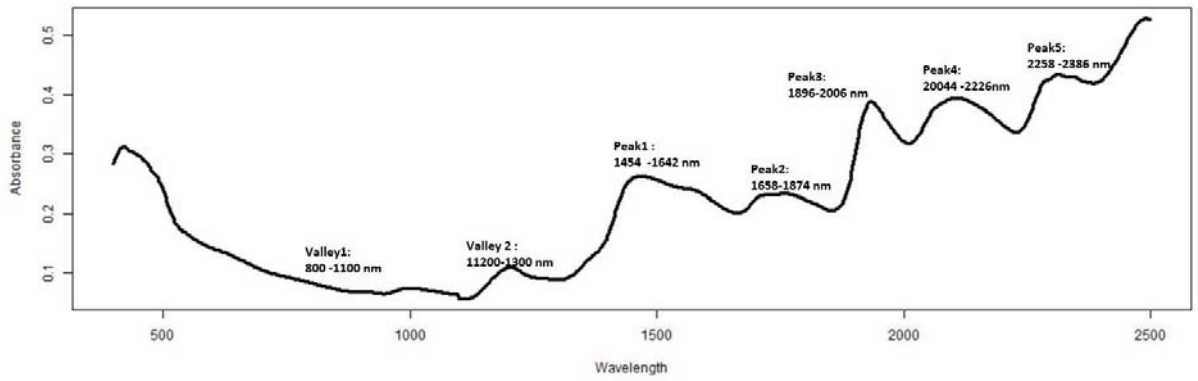

Fig. 1. Average spectrum of chickpea

recorded after scanning the material 32 times (Fig. 1). The concentration of protein physicochemical in chickpea seedsamples was measured in the chemical laboratory, which served as reference data for training and measuring the performance of prediction models.

Development of machine learning models: Machine learning techniques LR, ANN, PLSR, RF, SVR, and DTR were used to develop near infrared spectroscopy prediction models utilizing preprocessed spectra. The Comprehensive R Archive Network (CRAN) provided all model development packages. All models would be iterated 5000 times. The 237 samples were randomly separated into two groups: 75% (176 samples) as a training data set and 25% (61 samples) as a testing data set. All spectra were scaled so that the resulting model could be interpreted in terms of variance around the mean. To identify the best suited combination for all six machine learning methods, 27, 776 combinations were used. The preprocessed combination with the highest *r* and R^2 and the lowest RMSE value was chosen as the best. To evaluate the efficacy of the regression model, the RMSE, r, and R2 between measured and predicted concentration levels of protein physicochemical component in chickpea crop were calculated. The ideal model for each component was chosen based on the lowest RMSE value of prediction, highest *r*, and highest R^2 between measured and predicted values.

Table 1. Wavelength range for preprocessing

S.No.	Wavelength Range	No. of Wavelengths	File Name	Wavelength Characteristic
1	800-1100 nm	300	P6	N-H2nd overtone O-H2nd overtone C-H3rd overtone
2	1100-1300 nm	200	P7	C-H2nd overtone O-Hcombinations
3	1404-1642 nm	232	P1	C-H combinations, O-H, N-H, and 1st overtone
4	1658-1874 nm	216	P2	C-H and 1stovertone
5	1896-2006 nm	110	P3	O-H, N-H combinations
6	2004-2226 nm	222	P5	C-H, O-H, N-H combinations
7	2258-2386 nm	128	P4	C-H combinations

Results and Discussion

Different algorithms produced optimum prediction for protein with different RMSE for specific wavelength range at different *r* and R^2 values (Table 1 and Table 2). For instance, the RF algorithm produced the best prediction in a wavelength range of 1404-1642 nm with an RMSE of 0.09, *r* of 1.00, and R^2 of 0.87. On the other hand, the ANN algorithm predicted at 2258-2386

Table 2. Preprocessing methods applied on spectra with performance measures

Algorithm	RMSE	r	R^2	Moving Average[1]	Binning	Derivative	SG Polynomial	SG Window	Scatter Correction
RF	0. 09	1. 00	0. 87	0	0	0	0	0	SNV
SVR	0. 08	0. 95	0. 86	6	0	0	0	0	SNV-Detrend, MSC, SNV
ANN	0. 08	0. 94	0. 86	12	8	3	4	7	SNV-Detrend, SNV, MSC
DTR	0. 08	0. 93	0. 85	2	2	0	0	0	MSC, SNV
LR	0. 08	0. 93	0. 85	8	2	1	2	3	MSC, SNV
PLSR	0. 11	0. 84	0. 70	0	0	0	0	0	SNV-Detrend, MSC, SNV

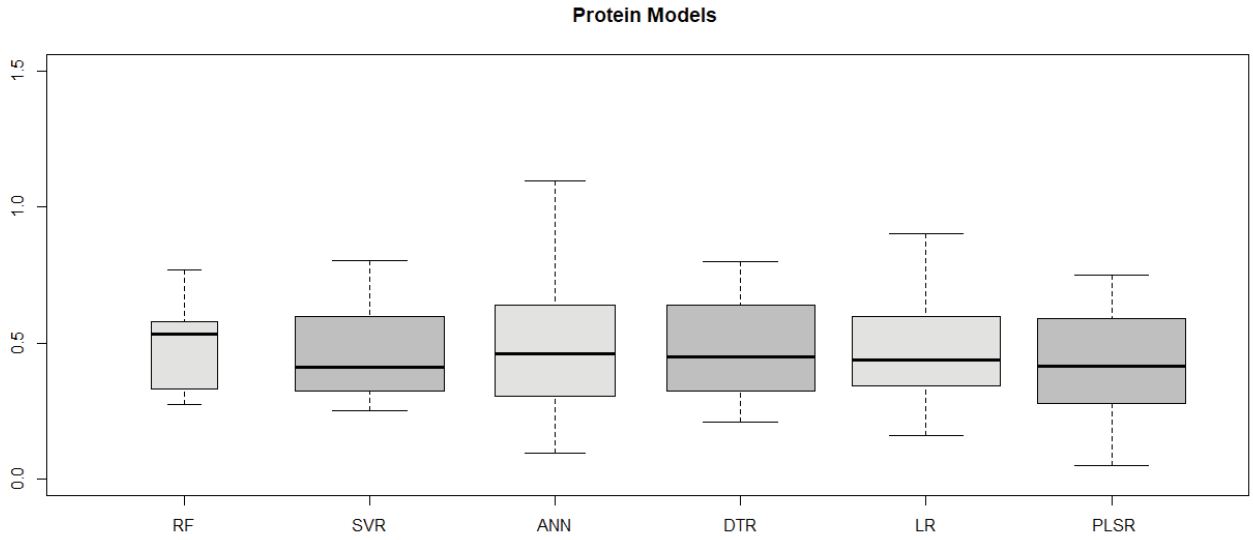

Fig. 2. Boxplots for six machine learning models for protein physicochemical component of chickpea

nm wavelength with an RMSE of 0.08, *r* of 0.95, and R^2 of 0.86.

Boxplots showed no outliers in the datasets, and the median value of all component models was observed to be close to 0.5. Compared to other models, ANN model for protein had maximum range, indicating that ANN models have maximum variability. Data variability of the DTR and SVR models were similar with practically equal coverage areas. The RF model had the least data variability with a negatively skewed median line, whereas the LR model was found to be positively skewed, whereas the PLSR model was symmetric.

Preprocessing is challenging to judge prior to model validation. All preprocessing techniques aim to minimize unmodeled variability in the data to improve the feature sought in the spectra (Rinnan *et al*., 2009). It has a linear relationship with a response variable such as a constituent. This is possible with the correct preprocessing technique, but there is always the risk of applying the wrong kind or over-processing, which will result in the loss of essential data. In the present study, experiments with 27, 776 preprocessing combinations were planned to achieve an appropriate preprocessing approach. They were put to the test, and the optimum preprocessing combination that gave the best prediction value from a model was identified.

It was discovered that RF performed best in the wavelength range 1404-1642 nm. Comparing all six models reveals that the RF models outperform all others. The RF technique is ensemble-based enhancing the model accuracy. Due to the randomness, the RF algorithm examines all variables at each node outperforms all other machine learning algorithms in processing speed and resistance to over-fitting noise. RF reduces the significant variance of a flexible model like a decision tree by combining numerous trees into one ensemble model. RF is less computationally intensive and does not require a graphics processor to complete training. A closer look at the data showed that the ANN model could also predict the physicochemical component of proteins within 500 iterations. It is worth noting that the tests were carried out under the laboratory condition and the models' dependability can only be proven once they've been applied to real-world procedures.

Conclusion

NIR, as a non-destructive technique, requires no or minimal sample preparation. Its use to find concentrations of physicochemical component provides excellent predictive methodology for germplasm evaluation. Results of the present study and application of machine learning algorithms is expected to scale up predicting physicochemical components in chickpea as well as other leguminous crops.

Acknowledgements

MadhuBala conducted this study as part of Ph.D. with study leave from ICAR-NBPGR at ICAR-IARI, New Delhi.

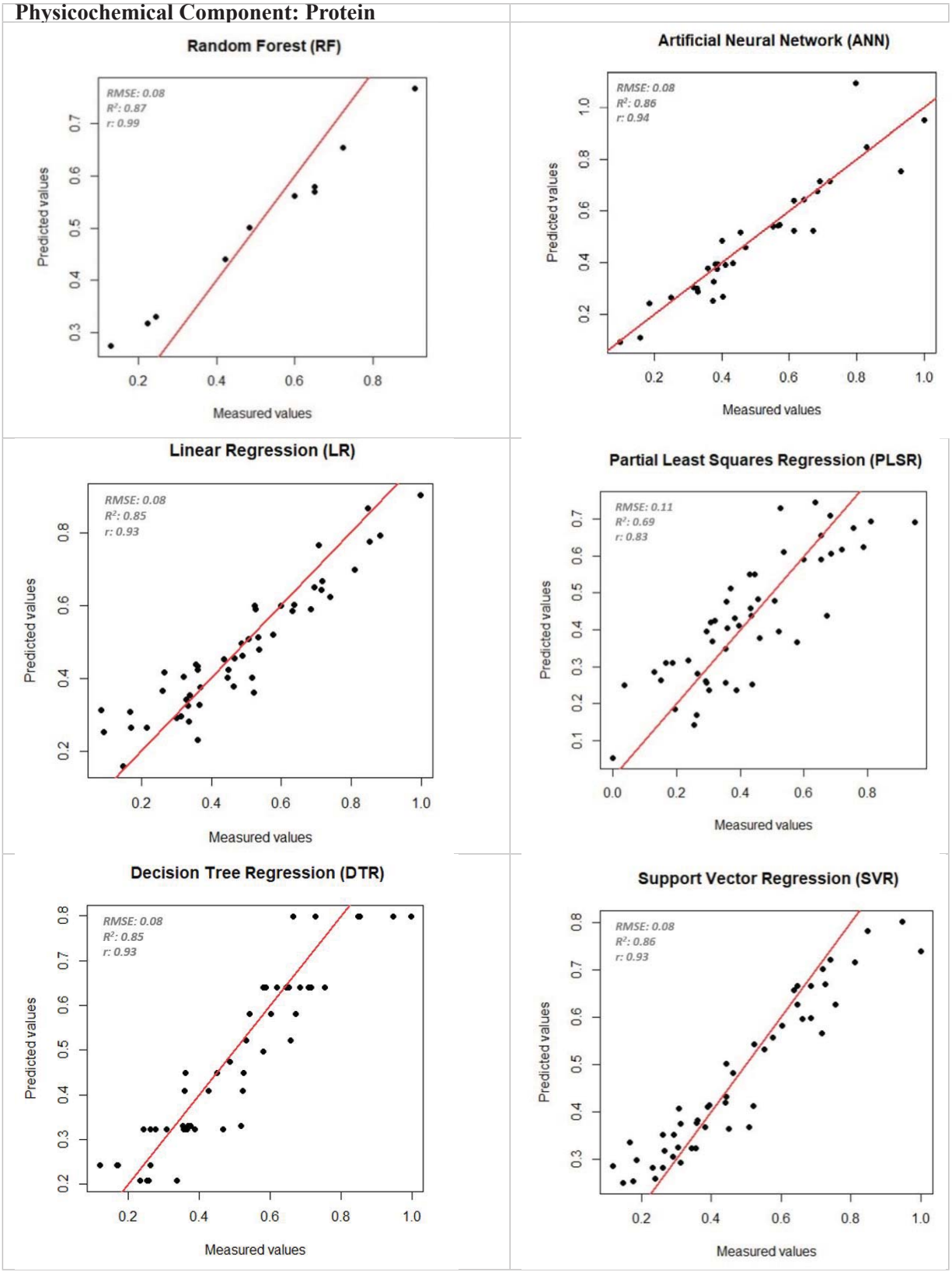

Fig. 3. Correlation between the measured and predicted value of protein physicochemical component of chickpea using all six prediction models. RMSE: root mean square error, R^2: coefficient of determination, *r*: correlation coefficient.

References

Acquah G, B Via, N Billor, O Fasina and L Eckhardt (2016) Identifying plant part composition of forest logging residue using infrared spectral data and linear discriminant analysis. *Sensors.* **16**: 1375. https://doi.org/10.3390/s16091375.

Batten GD (1998) Plant analysis using near-infrared reflectance spectroscopy: the potential and the limitations. *Aust. J. Expl. Agric.* **38**: 697-706.

Pasquini C (2003) Near infrared spectroscopy: Fundamentals, practical aspects, and analytical applications. *J. Braz. Chem. Soc.* **14(2):** 198-219.

Rinnan A, FVD Berg and SB Engelsen (2009) Review of most common pre-processing techniques for near infrared spectra. *Trends Anal. Chem.* **28**:1201-1222.