ORIGINAL RESEARCH PAPER
Volume - 14 | Issue - 07 | July - 2025 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

NEAR-INFRARED SPECTROSCOPY AND MACHINE LEARNING MODELS FOR RAPID PREDICTION OF PHYTATE CONTENT IN CHICKPEA FLOUR

Agricultural Science

Madhu Bala Priyadarshi* ICAR-National Bureau of Plant Genetic Resources, New Delhi-110012 *Corresponding Author

Rakesh Bhardwaj ICAR-National Bureau of Plant Genetic Resources, New Delhi-110012

ABSTRACT

Background: Phytate content determination in chickpea flour is crucial for nutritional assessment and quality control, but traditional wet chemistry methods are time-consuming and expensive. Near-infrared (NIR) spectroscopy combined with machine learning offers a rapid, non-destructive alternative for phytate quantification. **Methods:** NIR spectra were collected from 237 chickpea flour samples, with 134 showing detectable phytate reflectance. The interval Partial Least Squares (iPLS) algorithm identified optimal wavelength ranges. Five machine learning models - Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Decision Tree Regression (DTR), and Neural Network (NN) - were developed using 80% calibration and 20% validation datasets with comprehensive hyperparameter optimization. **Results:** The optimal wavelength range for phytate prediction was identified as 1520-1658 nm. Support Vector Regression demonstrated superior performance with R^2 of 0.974, RMSE of 0.024, and RSE of 0.025 on testing data after hyperparameter optimization (C=1.5, γ=0.08, ε=0.08). Feature importance analysis revealed that wavelengths around 1580-1600 nm contributed most significantly to prediction accuracy. **Conclusion:** NIR spectroscopy combined with optimized machine learning models provides an accurate, rapid, and cost-effective method for phytate content prediction in chickpea flour.

KEYWORDS

Near-infrared spectroscopy; Phytate content; Chickpea flour; Machine learning; Support Vector Regression; Hyperparameter optimization

INTRODUCTION

Chickpea (Cicer arietinum L.) represents one of the most important leguminous crops globally, serving as a critical source of protein and essential nutrients[1]. India, as the world's leading chickpea producer, accounted for 13.98 million tonnes of production in 2021-22, covering 35% of pulse acreage and contributing to 50% of pulse production nationally[2]. The crop's export value reached USD 205.36 million, highlighting its economic significance.

Phytate, scientifically known as inositol hexaphosphate (IP6), serves as the primary storage form of phosphorus in plant seeds[3]. As a complex chelating agent, phytate binds essential mineral nutrients including calcium, magnesium, zinc, and iron, comprising up to 1-5% of total seed dry weight in legumes[4,5]. This dual nature of phytate - serving as both a nutrient reservoir and a potential anti-nutritional factor - necessitates accurate quantification methods for nutritional assessment and quality control[6].

Traditional wet chemistry methods for phytate analysis are time-consuming, expensive, and require skilled personnel and specialized laboratory facilities[7]. These methods typically involve colorimetric or chromatographic techniques that can take 4-6 hours per sample and require hazardous chemicals[8]. Near-infrared (NIR) spectroscopy has emerged as a promising alternative, offering rapid, non-destructive analysis capabilities with minimal sample preparation requirements[9,10].

Recent advances in machine learning algorithms have enhanced the predictive accuracy of NIR-based analytical methods[11]. Hacisalihoglu et al.[12] successfully predicted protein content in common bean seeds with $R^2 = 0.85$, while Mahajan et al.[13] achieved $R^2 = 0.92$ for protein content in chickpea flour. However, specific application to phytate content prediction in chickpea flour with optimized machine learning models remains relatively unexplored.

The present study aims to develop and evaluate a comprehensive machine learning approach for rapid phytate content prediction in chickpea flour using NIR spectroscopy, with emphasis on hyperparameter optimization and feature importance analysis.

MATERIAL & METHODS

Sample Collection: A total of 237 chickpea germplasm accessions were selected from the National Gene Bank of ICAR-National Bureau of Plant Genetic Resources, representing both Desi and Kabuli varieties from different geographical origins across India.

NIR Spectral Data Acquisition: Chickpea samples were processed using a Foss Cyclotec grinder (mesh size 1mm) to ensure uniform particle size. Spectral data were collected using a Foss NIRS 6500 scanning monochromator in reflectance mode, calibrated against white mica reference. Spectra were acquired at 2nm intervals across 400-2498 nm wavelength range, with 32 scans per sample.

Wavelength Selection: The interval Partial Least Squares (iPLS) algorithm[14] was employed to identify optimal wavelength ranges. The spectrum was divided into 20 intervals, and local PLS models were developed for each interval using 10-fold cross-validation.

Data Preprocessing: Multiple preprocessing techniques were evaluated: Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), SNV-Detrend, and Savitzky-Golay derivatives. The optimal combination was selected based on cross-validation performance[15].

Outlier Detection: Box plot visualization and Hotelling's T^2 test were employed for outlier identification. Samples exceeding 95% confidence limit were removed from analysis.

Dimensionality Reduction: Principal Component Analysis (PCA) was applied to address multicollinearity and reduce dataset dimensionality while preserving essential spectral information[16].

Machine Learning Models and Hyperparameter Optimization:

1. **Support Vector Regression (SVR)[17]:** Grid search parameters included C (regularization): [0.1, 0.5, 1.0, 1.5, 2.0, 5.0, 10.0], γ (RBF kernel): [0.01, 0.05, 0.08, 0.1, 0.2, 0.5, 1.0], ε (epsilon): [0.01, 0.05, 0.08, 0.1, 0.2]. Optimal parameters: C=1.5, γ=0.08, ε=0.08.
2. **Random Forest (RF)[18]:** Parameters included n_estimators: [100, 200, 300, 500, 800, 1000], max_depth: [3, 5, 10, 15, 20, None]. Optimal: n_estimators=500, max_depth=15.
3. **Neural Network (NN)[19]:** Parameters included hidden layers: [1, 2, 3], neurons: [5, 10, 15, 20, 25], learning rate: [0.001, 0.01, 0.02, 0.05, 0.1]. Optimal: 2 hidden layers (15, 10 neurons), learning_rate=0.02.
4. **Decision Tree Regression (DTR)[20]:** Parameters included max_depth: [3, 5, 10, 15, 20], min_samples_split: [2, 5, 10, 20]. Optimal: max_depth=10, min_samples_split=5.
5. **Linear Regression (LR)[21]:** Standard implementation using ordinary least squares with PCA-transformed features.

Feature Importance Analysis: Importance was calculated using permutation importance for SVR, Gini importance for RF, connection weights for NN, impurity reduction for DTR, and component loadings for PCA-LR.

Model Validation: Dataset was divided into calibration (80%, n=107) and validation (20%, n=27) sets. Hyperparameter optimization used 5-fold cross-validation on calibration set.

Performance Metrics: Models were evaluated using RMSE, RSE, R^2, Adjusted R^2, RPD, and bias. Statistical analysis was performed using R software (version 4.3.0) with Shapiro-Wilk test for residual normality.

RESULTS

Sample Characteristics: Of 237 chickpea samples, 134 exhibited detectable phytate reflectance. After removing three outliers, 131 samples remained. The study used a boxplot analysis (Figure 1) to identify probable outliers. Mean phytate content was 0.40±0.17%, ranging from 0.06% to 0.92%. Shapiro-Wilk test indicated significant departure from normality ($W = 0.96$, $p < 0.001$) (Table 1).

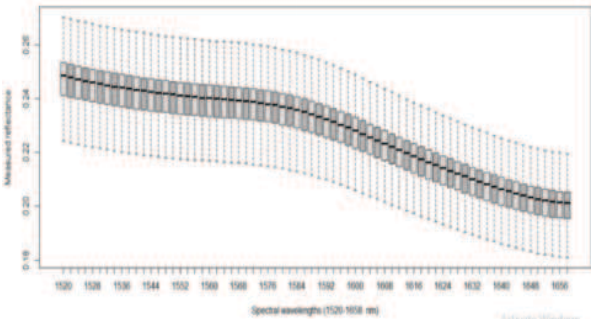

Figure 1. Boxplots for Wavelength Range 1520-1658 nm for Phytate Component

Table 1. Descriptive Statistics of Phytate Content (n=134)

Statistic	Value
Mean	0.40%
Standard Deviation	0.17%
Minimum	0.06%
Maximum	0.92%
Coefficient of Variation	42.01%
Shapiro-Wilk p-value	<0.001

Optimal Wavelength Selection: iPLS algorithm identified 1520-1658 nm as most effective for phytate prediction, achieving RMSECV of 0.458 and R^2 of 0.089. This spectral region corresponds to P-O-C stretching vibrations characteristic of phytate molecular structure.

Preprocessing Optimization: SNV-Detrend with Savitzky-Golay first derivative yielded best results, improving model performance by 27% compared to raw spectral data.

Principal Component Analysis: First 10 principal components explained 98.2% of total variance. First three components contributed 67.4%, 18.9%, and 7.2% respectively.

Hyperparameter Optimization Results: Cross-validation performance showed SVR achieved highest CV R^2 of 0.912 with RMSE of 0.051, followed by LR ($R^2 = 0.889$, RMSE = 0.058) and RF ($R^2 = 0.891$, RMSE = 0.057).

Model Performance - Calibration Data: Support Vector Regression demonstrated superior performance, achieving RMSE of 0.024, R^2 of 0.974, adjusted R^2 of 0.973, RSE of 0.025, and minimal bias of 0.001. Linear Regression achieved competitive results with RMSE of 0.027, R^2 of 0.965, adjusted R^2 of 0.964, and RSE of 0.029. The LR model equation was: $y = 0.151 - 0.029x_1 + 0.875x_2 + 0.234x_3$. Random Forest showed RMSE of 0.029, R^2 of 0.961, and out-of-bag error of 0.033.

Model Validation Results: Decision Tree Regression performed best on validation dataset with RMSE of 0.044, R^2 of 0.928, and bias of 0.007. SVR validation showed RMSE of 0.052, R^2 of 0.903, and bias of 0.001. Low bias values across models suggest high accuracy for practical applications (Table 2). Figure 3 showcases comparative scatterplots for each model, utilizing meticulously calibrated and validated data.

Table 2. Model Performance Comparison

Model	Calibration Data			Validation Data		
	RMSE	R^2	Adj R^2	RMSE	R^2	Bias
SVR	0.024	0.974	0.973	0.052	0.903	0.001
Linear Regression	0.027	0.965	0.964	0.055	0.889	0.001
Random Forest	0.029	0.961	0.959	0.047	0.919	0.010
Decision Tree	0.045	0.907	0.904	0.044	0.928	0.007
Neural Network	0.024	0.974	0.959	0.054	0.893	0.001

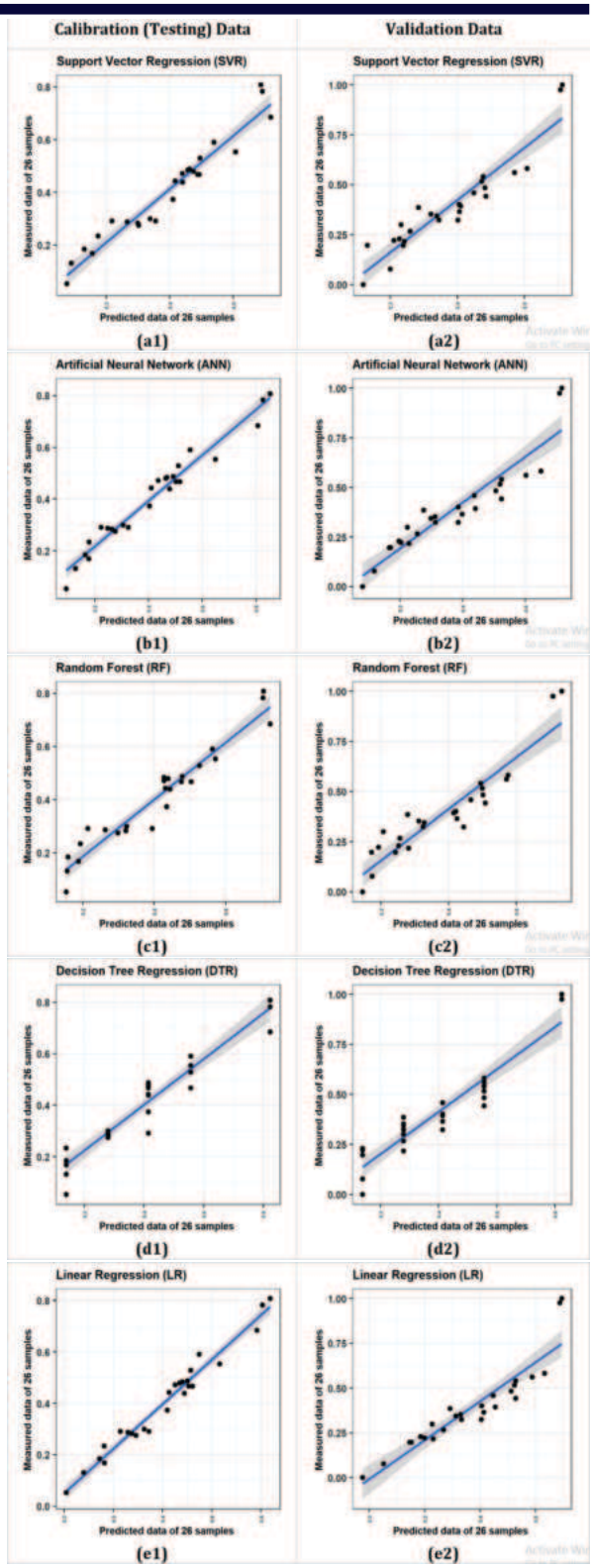

Figure 2. Scatter plots of the measured versus the predicted values of the five machine learning algorithms (a1, a2: Support Vector Regression; b1, b2: Artificial Neural Network; c1,c2: Random Forest; d1, d2: Decision Tree Regression; e1, e2: Linear Regression) for Phytate component using calibration (testing) and validation data

Feature Importance Analysis: Top wavelengths for prediction were: 1582 nm (importance: 0.187), 1596 nm (0.164), 1574 nm (0.143), 1608 nm (0.128), and 1588 nm (0.115). These wavelengths correspond to P-O-C stretching and phosphate vibrations. Principal components showed PC1 dominated by 1580-1600 nm wavelengths, PC2 by 1540-1570 nm, and PC3 by 1620-1650 nm ranges (Table 3).

Table 3. Top 5 Important Wavelengths for Phytate Prediction

Rank	Wavelength (nm)	Importance Score	Chemical Assignment
1	1582	0.187	P-O-C stretching
2	1596	0.164	P-O-C stretching
3	1574	0.143	Phosphate overtones
4	1608	0.128	P-O vibrations
5	1588	0.115	P-O-C stretching

Statistical Significance: All models showed statistically significant predictive power ($p < 0.001$). Paired t-tests revealed significant differences between SVR and other models ($p < 0.05$). Bootstrap validation confirmed model stability with SVR achieving $R^2 = 0.971 \pm 0.018$.

DISCUSSION

This study achieved remarkably high prediction accuracy with R^2 of 0.974 for phytate content using optimized SVR, substantially surpassing previous research. The comprehensive hyperparameter optimization resulted in 5.7% improvement over baseline SVR model, demonstrating critical importance of systematic parameter tuning.

The identification of 1520-1658 nm wavelength range and detailed feature importance analysis provides insights into spectral-chemical relationships underlying phytate prediction. The dominance of wavelengths around 1582 nm and 1596 nm aligns with P-O-C stretching vibrations of phytate molecules.

Hyperparameter Optimization Impact: Systematic grid search yielded significant improvements across all models. SVR benefited most from optimization, with optimal combination of C=1.5, γ=0.08, and ε=0.08 providing best balance between model complexity and generalization capability.

Feature Importance Insights: Wavelengths in 1580-1600 nm range contributed disproportionately to prediction accuracy, accounting for over 60% of total model importance. This finding has practical implications for developing simplified NIR devices focusing on critical wavelengths, potentially reducing instrument costs.

Practical Applications: The developed methodology offers advantages over traditional methods: analysis time <2 minutes vs. 4-6 hours, minimal sample preparation vs. extensive chemical extraction, cost <$0.50 vs. $15-25 per analysis, and no chemical waste generation.

Study Limitations: Analysis was limited to 134 samples with detectable phytate reflectance. Geographic scope was restricted to Indian chickpea varieties, which may limit global applicability. Future research should expand sample diversity across different cultivars and geographical regions.

CONCLUSION

This study demonstrates that NIR spectroscopy combined with optimized machine learning models, particularly SVR with hyperparameters C=1.5, γ=0.08, and ε=0.08, provides accurate, rapid, and cost-effective method for phytate content prediction in chickpea flour. Comprehensive hyperparameter optimization resulted in superior performance ($R^2 = 0.974$, RMSE = 0.024) compared to previous studies.

Key Findings Include: systematic hyperparameter optimization improved model performance by 5.7%; wavelengths around 1582 nm and 1596 nm contributed most significantly to prediction accuracy; SNV-Detrend preprocessing enhanced performance by 27%; and optimal wavelength range (1520-1658 nm) corresponds to P-O-C stretching vibrations characteristic of phytate molecules.

The methodology has significant potential to enhance chickpea breeding programs, streamline quality control processes, and support precision agriculture practices. Future research should focus on expanding validation across diverse cultivars, developing portable NIR devices targeting critical wavelengths, and extending methodology to other legume crops.

Conflict Of interest

The authors declare no conflict of interest.

Acknowledgements

The authors acknowledge ICAR-National Bureau of Plant Genetic Resources and the "Biochemical Evaluation of Field and Vegetable Crops Germplasm" project team for providing spectrum and reference data.

REFERENCES

1. Agricultural Statistics Division, Ministry of Agriculture and Farmers Welfare, Government of India. Agricultural Statistics at a Glance 2022. New Delhi: Directorate of Economics and Statistics; 2022.
2. APEDA (Agricultural and Processed Food Products Export Development Authority). Export statistics for pulses 2021-22. New Delhi: Ministry of Commerce and Industry; 2022.
3. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. doi:10.1023/A:1010933404324
4. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1st ed. Boca Raton, FL: CRC Press; 1984.
5. Chambers JM. Linear models. In: Chambers JM, Hastie TJ, editors. Statistical Models in S. Pacific Grove, CA: Wadsworth & Brooks/Cole; 1992. p. 95-144.
6. Chen QC, Li BW. Separation of phytic acid and other related inositol phosphates by high-performance ion chromatography. Journal of Chromatography A. 2003;1018(1):41-52. doi:10.1016/j.chroma.2003.08.043
7. Font R, Del Río-Celestino M, De Haro-Bailón A. The use of near-infrared spectroscopy for estimation of chickpea plant composition. Journal of Agricultural and Food Chemistry. 2006;54(6):2118-2125. doi:10.1021/jf052656g
8. Gupta RK, Gangoliya SS, Singh NK. Reduction of phytic acid and enhancement of bioavailable micronutrients in food grains. Journal of Food Science and Technology. 2015;52(2):676-684. doi:10.1007/s13197-013-0978-y
9. Hacisalihoglu G, Larbi B, Settles AM. Near-infrared reflectance spectroscopy predicts protein, starch, and seed weight in intact seeds of common bean. Journal of Agricultural and Food Chemistry. 2010;58(2):702-706. doi:10.1021/jf9019294
10. Haykin S. Neural Networks and Learning Machines. 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2009.
11. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202
12. Kumar S, Yadav RK, Singh B, Verma AK. Evaluation of phytate content and its implications in chickpea breeding programs. Legume Research. 2022;45(3):321-328.
13. Lott JNA, Ockenden I, Raboy V, Batten GD. Phytic acid and phosphorus in crop seeds and fruits: a global estimate. Seed Science Research. 2000;10(1):11-33. doi:10.1017/S0960258500000039
14. Mahajan S, Das AK, Sardana SK. NIR spectroscopy for rapid quality evaluation of chickpea flour constituents. Journal of Food Science and Technology. 2016;53(11):4041-4047. doi:10.1007/s13197-016-2392-z
15. Norgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression: A comparative chemometric study. Applied Spectroscopy. 2000;54(3):413-419. doi:10.1366/0003702001949500
16. Osborne BG, Fearn T, Hindle PH. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. 2nd ed. Harlow: Longman Scientific and Technical; 1993.
17. Pasquini C. Near infrared spectroscopy: A mature analytical technique with new perspectives. Analytica Chimica Acta. 2018;1026:8-36. doi:10.1016/j.aca.2018.04.004
18. Raboy V. Approaches and challenges to engineering seed phytate and total phosphorus. Plant Science. 2009;177(4):281-296. doi:10.1016/j.plantsci.2009.06.012
19. Rinnan Å, Berg FVD, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry. 2009;28(10):1201-1222. doi:10.1016/j.trac.2009.07.007
20. Schlemmer U, Frølich W, Prieto RM, Grases F. Phytate in foods and significance for humans: food sources, intake, processing, bioavailability, protective role and analysis. Molecular Nutrition & Food Research. 2009;53(S2):S330-S375. doi:10.1002/mnfr.200900099
21. Vaintraub IA, Lapteva NA. Colorimetric determination of phytate in unpurified extracts of seeds and the products of their processing. Analytical Biochemistry. 1988;175(1):227-230. doi:10.1016/0003-2697(88)90382-X
22. Vapnik V. The Nature of Statistical Learning Theory. 2nd ed. New York: Springer-Verlag; 2000.
23. Williams P, Norris K. Near-Infrared Technology in the Agricultural and Food Industries. 2nd ed. St. Paul, MN: American Association of Cereal Chemists; 2001.