Multidisciplinary Research Journal
ISSN: 3049-2718 (Print)

Volume 1, Issue 2, 2025
April - June 2025

Research Article

Near Infrared Spectroscopy and Machine Learning for Non-Destructive Estimation of Ageing of Komal Chaul

Shagufta Rizwana [1,2*], Manuj Kumar Hazarika [1],

1 Department of Food Engineering and Technology, Tezpur University, Assam, 784028, India.
2 The Assam Royal Global University, Guwahati, Assam, PIN- 78034, India.
* Correspondence: rizwanshagufta.sr@gmail.com

Scopus Author ID 55661416100

Received: 31 December 2024; Accepted: 14 April 2025; Published: 28 April 2025

Abstract: *Komal Chaul* is a traditional form of a parboiled rice product from rice varieties indigenous to the state of Assam and it constitutes a culturally significant dish served during the auspicious occasions. Ageing impacts its rehydration properties diminishing its value as a no-cooking rice product. To empower the consumers in selecting *Komal Chaul* of desired rehydration qualities, this study focused on developing a non-destructive tool based on near-infrared spectral data coupled with machine learning (ML) algorithm for distinguishing the aged *Komal Chaul*. An NIR spectral library of *Komal Chaul* samples was created, covering the spectral range of 740–1050 nm for samples stored for a period up to one year under ambient conditions. The methodology involved spectral preprocessing to enhance data quality, followed by partial least squares (PLS) regression modeling to predict storage time. Statistical metrics, including regression coefficient (R^2), relative error percentage (REP), and root mean squared error (RMSE), were used to validate the model. Feature selection based on coefficient weightage was performed to identify key wavelengths contributing to time prediction. Classification models, including LDA, KNN, CART, Naïve Bayes, SVM, and Random Forest, were employed to categorize samples into aging periods of 1, 3, and 6 months. Partial Least Squares (PLS) regression models predicted the ageing time with a validation score R^2 of 0.897 and RMSE of 19.41 days. Optimized with wavelength selection, the PLS regression model achieved significant accuracy in estimating the ageing time, with a prediction score R^2 of 0.89 and RMSE of 2.01 days. Similarly, using the same approach for cooking quality prediction resulted in satisfactory performance, achieving a validation R^2 of 0.79. Classification models further enhanced prediction accuracy, with the Random Forest model attaining the highest accuracy of 92% for six-month interval classifications. These results underscore the potential of integrating NIR spectroscopy and machine learning for efficient, non-destructive quality assessment of *Komal Chaul*, supporting its commercialization as a value-added traditional food product.

Keywords: Nondestructive; aged-rice; Komal Chaul; ready-to-eat; Machine Learning

1. Introduction

Chokuwa rice of Assam is an indigenous variety to the state with a GI tag used in the production of *Komal Chaul* [1]. *Chokuwa* rice of Assam is a low amylose variety grown majorly in the *Sali* season (Autumn-Winter). The word '*Komal*' means 'soft' and '*Chawal'* or '*Chaul'* means rice in Assamese. *Komal Chaul* is prepared by parboiling *Chokuwa* paddy. The unique characteristic of this rice is its no cooking requirement [2]. It can be consumed by soaking it in warm water at nearly 60-70°C temperature. This distinctive characteristic qualifies this rice in the ready-to-eat category, though it cannot be categorized as instant food exactly because *Komal Chaul* takes at least 10-12 min to gelatinize/cook in warm water fully [3]. In Assamese tradition, *Komal Chaul* is consumed as "*jalpan"*, a culturally significant dish served during auspicious occasions and fasting periods. The rice is consumed as brown rice and has been reported to have a lower glycaemic index, making it beneficial [4]. It is traditionally manufactured by soaking rice for 3-4 days to attain a desirable moisture content. Saoking water is drained, freshwater is added and cooked over a wood fire until the grain splits out of the husk. The water is drained off again, and the grains are dried in the sun on the same day. Drying is done on the same day to avoid retrogradation and maintain a soft structure [5]. The traditional method is weather-dependent, and unfavorable weather conditions result in delayed drying, affecting the final product's quality [6].

Komal Chaul, possessing a unique characteristic of no cooking requirement, has many scopes for marketability. Though it is not fully commercialized, there is potential for its application in the ready-to-eat food section. One issue with *Komal Chaul* is that its rehydration property after processing is affected. Ageing of the rice also results in lesser hydration. Moreover, there is neither a standard protocol nor any device developed for estimating rehydration quality concerning ageing. So, a handy tool can prove to be a reliable solution to this. In another context, governmental agencies like the Department of Food and Public Distribution under the Ministry of Consumer Affairs, Food and Public Distribution are looking for real-time scientific methods regarding the prediction of aged rice [7].

Scientific methods for predicting *Komal Chaul's* ageing will be helpful from a trading and purchasing perspective. Non-destructive techniques based on spectra and images are nowadays state-of-the-art methodologies in many disciplines. Near-infrared spectroscopy (NIRS) is usually preferred for its non-destructive nature and can easily record spectra from solid and liquid samples without any pre-treatment [8]. The Indian Agricultural Research Institute has done extensive research on the application of NIR and other non-destructive techniques for the quality evaluation of intact fruits and vegetables [9][10]. Unlike other non-destructive techniques, NIR is preferred for its cost-effectiveness. Its instrumentation involves simple mechanics and robust sensors, making it suitable for online process analysis [11]. NIR spectroscopy, being a secondary technique, requires chemometrics for calibration purposes. Machine Learning has been used as a calibration tool to map a link of near-infrared spectroscopy-based measurements into desirable extrinsic factors of analysis. Studies have employed NIR spectroscopy combined with ML models like partial least squares to identify aged-rice adulteration, highlighting its utility in monitoring rice ageing [12].

The most significant biochemical phenomenon influencing the texture and cooking quality of aged rice is starch retrogradation. Though it is desirable for most rice, it is undesirable for *Komal Chaul*. Therefore, there is a need for an ageing study of *Komal Chaul* to assess its shelf life in terms of cooking quality. NIR spectroscopy has proven reliable to calibrate biochemical changes. Therefore, the objective was to develop an NIR-based tool with the help of machine learning methods for predicting the time of ageing of *Komal Chaul* relating to its cooking properties.

2. Materials and Methods

2.1. Mapping of ageing time with NIR spectral data

Freshly harvested *Chokuwa* paddy was collected from a farm in Jamugurihat, Assam. *Komal Chaul* was prepared from brown rice by the process of parboiling [2] and was taken for the ageing study as shown in Fig. 1. The rice was stored in LDPE pouches and those packets were kept in a closed plastic box of approximate dimensions of 29 x 34 x 12 cm at room temperature. The temperature variation throughout the year was 23-37 °C and relative humidity was in the range of 85-90% during summers and 60-65% during winters. The target parameter here was the time of storage (degree of ageing) associated with quality attributes during ageing.

(a)

Figure 1. (a) Process flowchart of production of *Komal Chaul*; **(b)** NIR Sensor with sample holder

The respective spectrum of samples was taken over one year on a weekly basis. The procedure involves:

- Firstly, an NIR-based spectral library of 255 scans of aged and unaged rice samples with their respective storage time were recorded. The storage time was considered as the reference parameter or target parameter. The respective spectra were acquired in the range of 740-1050 nm using the NIR sensor (SCiO 1.2) (Fig. 1(b))[13].

➢ Initially, the whole spectrum was taken as a reference spectrum with a resolution of 5 nm, giving us 63 points for every spectrum. The total number of spectrums was 255 and the average scans per sample was 3. The dataset was divided into training (70%), validation(20%), and test sets(10%).

➢ Preprocessing of spectra is an essential step to enhance its quality, remove noise, and normalize the data for further analysis in regression.

➢ Developing an ML-based predictive model for estimating the time of ageing of Komal Chaul using the spectral library. The PLS process is explained in figure 2.

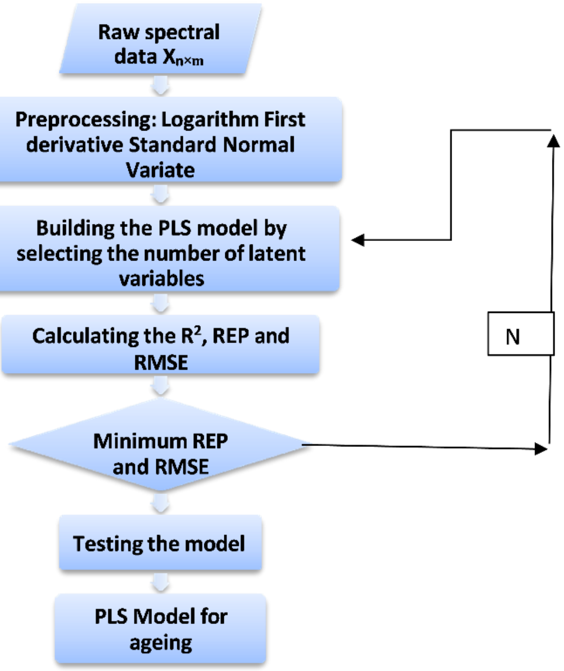


Figure 2. Partial Least Square (PLS) steps

The preprocessed data was then trained on PLS. The PLS algorithm iteratively calculated a set of latent variables over 40 components. The optimal number of components was chosen based on the Mean squared error.

➢ The performance of the model was evaluated using statistical metrics like Regression coefficient (R2), Relative Error percentage (REP), and Root mean squared error (RMSE) (Eq. 1- 3) using the SciPy library in Google Colab, Python version 3.8.

$$R^2 = \frac{n\sum_{i=1}^{n} x_{pi}x_{mi} - \sum_{i=1}^{n} x_{pi}x_{mi}}{\sqrt{\left(n\sum_{i=1}^{n} x_{pi}{}^2 - \left(\sum_{i=1}^{n} x_{pi}\right)^2\right)\left(n\sum_{i=1}^{n} x_{mi}{}^2 - \left(\sum_{i=1}^{n} x_{mi}\right)^2\right)}} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{1}^{n}(x_{pi} - x_{mi})^2}{n-1}} \tag{2}$$

$$REP = \frac{x_{pi} - x_{mi}}{x_{mi}} \times 100 \quad (3)$$

Where $\boldsymbol{x_{pi}}$ is the predicted ith value and $\boldsymbol{x_{mi}}$ is the measured ith value**.**

- The validated model was again used for prediction.
- There are specific regions in the spectrum that contributed more to the PLS regressor loading values, therefore, a coefficient weightage-based selection of optimal features was done to get the wavelengths more distinctively contributing to the PLS regression of time calculation.

2.2. Mapping of cooking time with NIR spectral data

The steps for mapping of cooking time with NIR data followed the same PLS procedure as mentioned in 2.1.

The cooking quality or softening properties of *Komal Chaul* from *Chokuwa* rice of Assam concerning ageing time was measured, and the parameter taken into consideration is the softening time. The time from starting to cook to the first sampling instance at which 90% of kernels are translucent is considered as the cooking time. To calculate the cooking time, 10 g of kernels were kept in a hot water bath at 60 ± 5 °C. Eight minutes onwards, at an interval of 1 min, 10 kernels were taken out and pressed between two glass slides. The percentage of translucent kernels was noted, and cooking time was determined when the percentage was above 90% [14].

2.3. Classification of NIR spectral on monthly basis

The next step involved the classification of spectral data based on their ageing time. The spectral database was coded according to a 1-month classification consisting of 12 classes, 3 months of classification, and 6 months of classification. The dataset was divided into training (70%), validation (20%), and testing (10%). The dataset was then trained, validated, and tested on classification models: LDA, KNN, CART, NB, SVM, and RF. Huang et al., 2023 used discriminant and ensemble tree algorithms like SVM and RF for distinguishing storage periods of rice [15]. The performances of the classification were measured using two metrics: classification accuracy (Eq.4) and F1_score (Eq.5).

$$Accuracy = \frac{Number\ of\ samples\ correctly\ classified}{\text{Total number of samples}} \quad (4)$$

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall} \quad (5)$$

3. Results and discussion

3.1. Estimation ageing time of Komal Chaul by ML model

Firstly, NIR-based spectral library (wavelength range: 740-1050 nm) of aged and unaged rice samples with their respective ages from the time of harvesting was created (Fig.

3). The raw data were preprocessed using the Standard Normal Variate and Savitzky Golay function, that takes the derivative of the data (Fig. 4). Then, a Partial Least Square (PLS), a ML-based regression model was trained on the pre-processed data, validated, and tested (k-fold selection) for estimating the time of ageing of rice using the spectral library. Later, the predictive model with best-fitted parameters. The least MSE values were obtained for 10 latent variables, so a PLS model with 10 components was trained (Fig. 5). The PLS model for ageing time estimation showed an R2 of 0.897 and an RMSE of 19.409 days for the validation set. The test dataset showed a lesser accuracy with an R2 value of 0.7013 and RMSE to be 32.477 days (Fig. 6). However, the relative percent deviation was less, that is, 3.08. The signal-to-noise ratio of raw reflectance data was observed to be 2.8 for average signal, while the signal-to-noise ratio is 0.01.

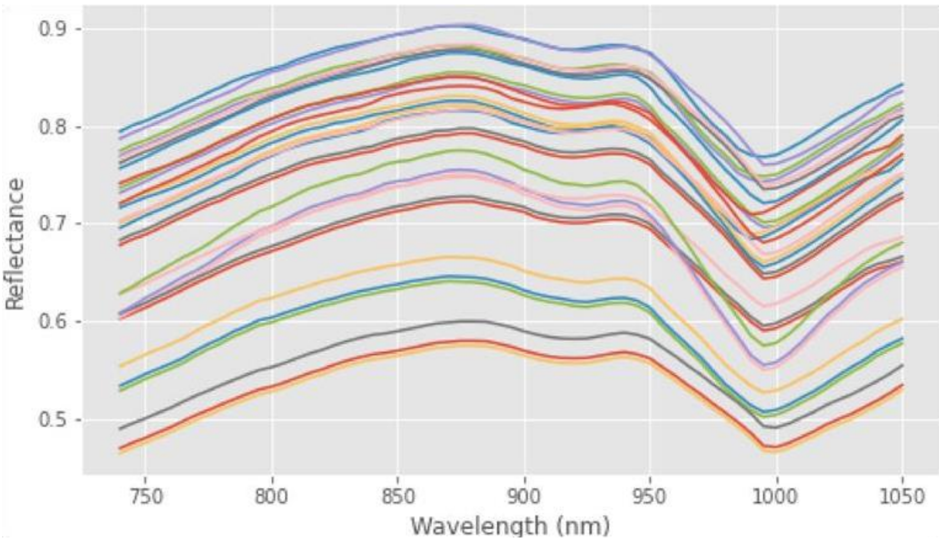


Figure 3. Raw reflectance vs wavelength

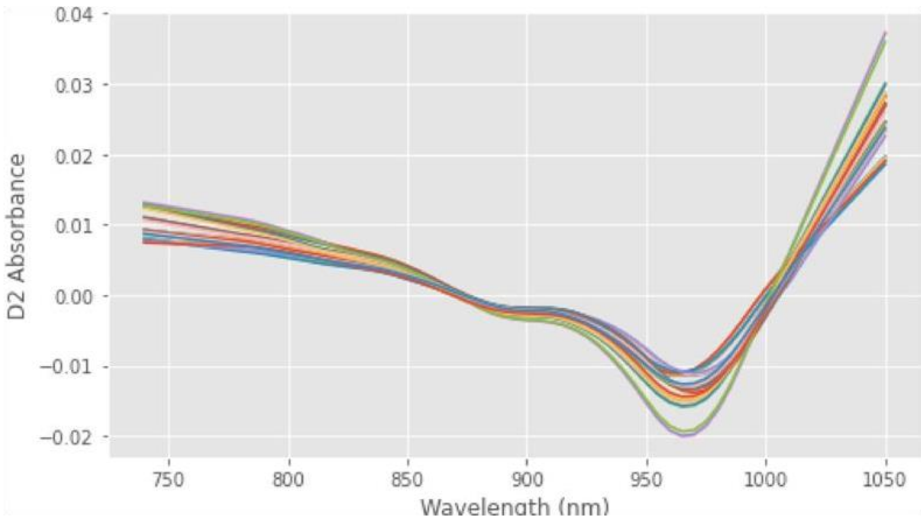


Figure 4. Pre- processed spectra of raw reflectance

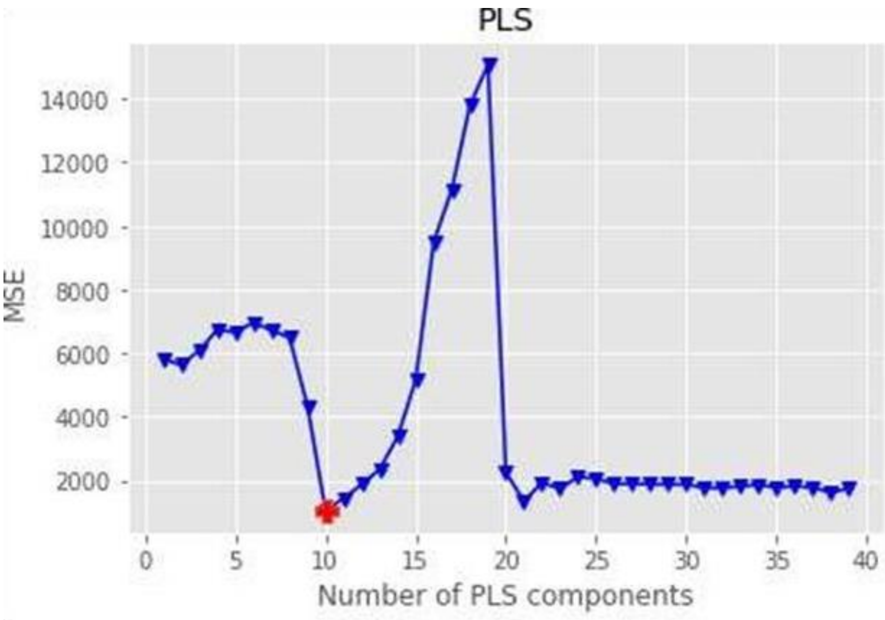


Figure 5. Hyper-tuning the Latent variables

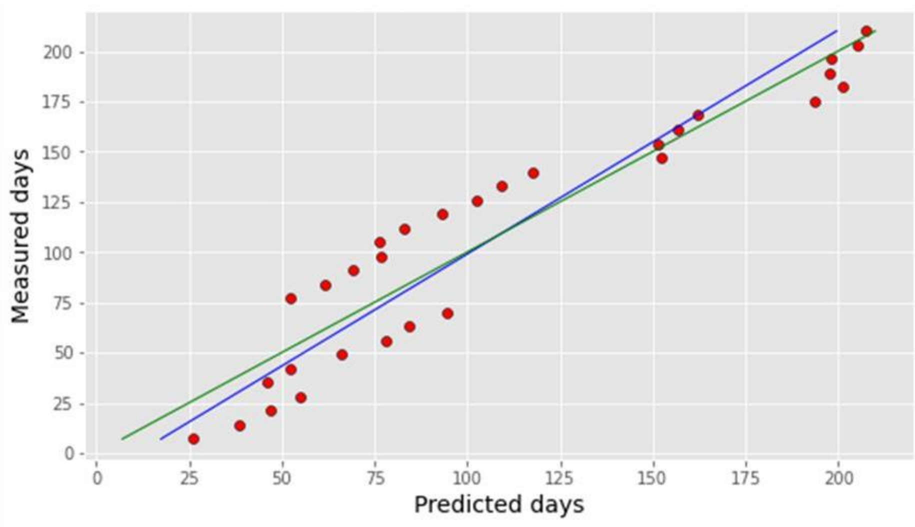


Figure 6. Cross-validation plot of test data for ageing time

To improve the performance of the model, a variable selection approach was made. Earlier studies had suggested that variable selection for high-dimensional data helped improve the performance of those with higher dimensions of data. There are various ways to approach selecting variables for PLS regression: filter method, wrapper method, and embedded method. For this study, a filter method was applied for wavelength selection [16]. The filtration was based on the regression coefficients of the PLS model. The association between PLS coefficients for every wavelength gave us an idea of largely contributing features (Fig. 7).

The number of PLS components after optimization was chosen to be 10. The application of a recursive method by eliminating one wavelength at a time and calculating the MSE value. The optimal number of discarded wavelength features was 37, based on an optimal MSE value of 11.29. The chosen wavelength values were 790, 795, 800, 805, 810, 830, 835, 840, 845, 850, 855, 860, 875, 880, 885, 900, 905, 910, 930, 935, 940, 945, 970, 975, 980 and 1000 (Fig. 8). The cross-validation plot between the measured and predicted showed better regression with an R2 value of 0.89 (Fig. 9). The regression for the test dataset was better than the previous PLS model suggesting better predictability. The RMSE values of validation and test set were found to be 2.01 and 3.24 days which is considerably low. The SE for validation and test set were found to be 1.75 and 14.24 days, and biases were -3.9 and -9.6 days, respectively. Recent study by Song et al. [17] has demonstrated the effective identification of aged-rice adulteration using support vector machine (SVM) classification combined with characteristic wavelength variable selection where features were eliminated to improve the prediction efficiency up to 98%.

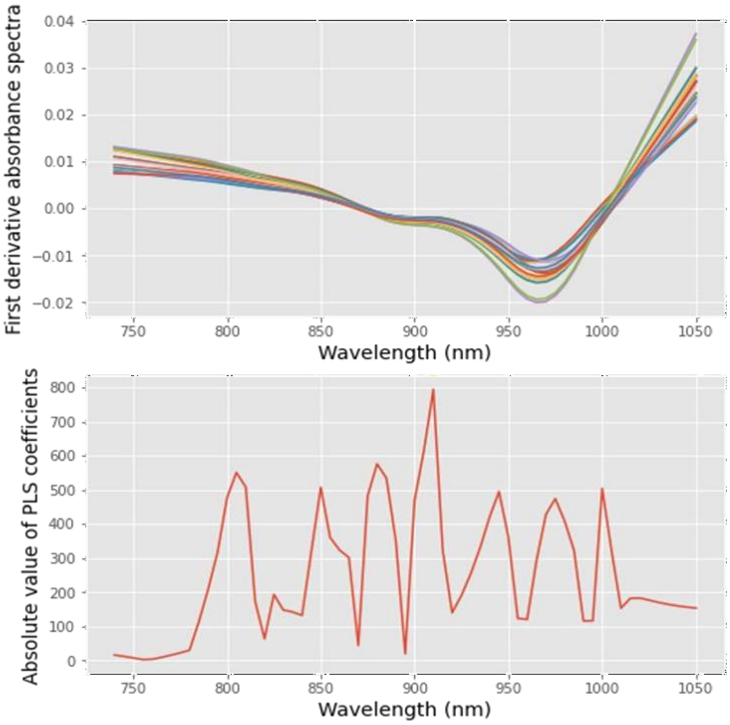


Figure 7. Fitting parameters for optimizing features

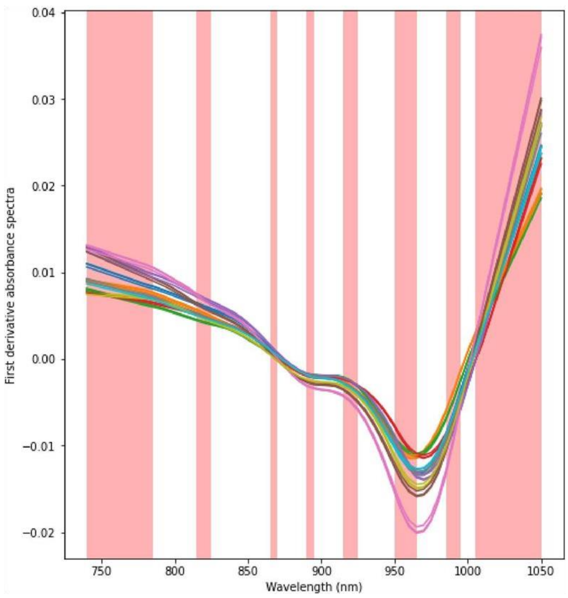


Figure 8. Optimal features selection

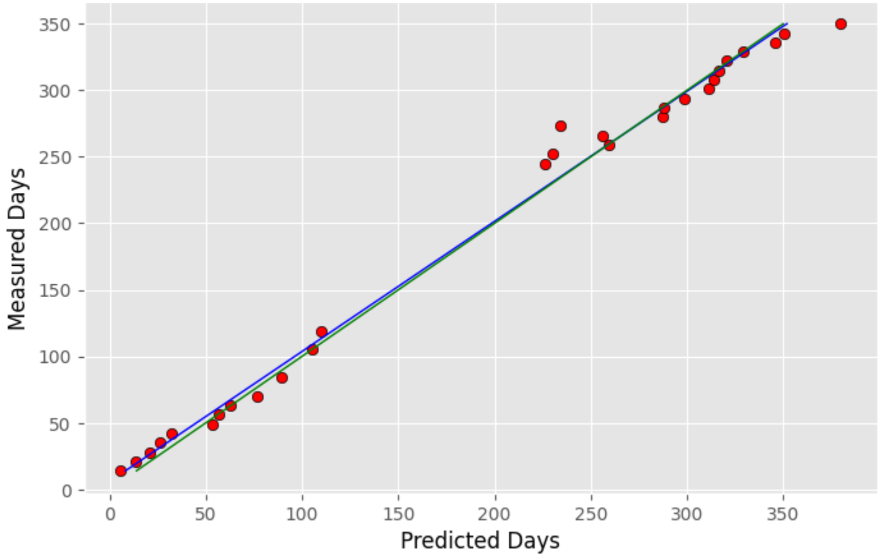


Figure 9. Cross validation plot after optimization

3.2. Estimation of cooking quality of Komal Chaul by ML model

The same spectral data (Fig. 10) were later mapped with the cooking time or softening time of *Komal Chaul* have changed linearly with time. Those raw data were preprocessed using the Savitzky Golay function, which takes the derivative of the data. PLS regression was used to model cooking with spectral data taken during the process of ageing. The training set was trained over 40 numbers of Latent variables. The least MSE values were obtained for 10 latent variables, so the PLS model with 10 components was validated. The validation score was found to be R2: 0.79 and RMSE of 1.09. The test dataset showed accuracy with an R2 value of 0.77 and an RMSE to be 2.77 (Fig. 11). The RMSE was higher because the endpoint-predicted deviated more for both the higher values and lower values.

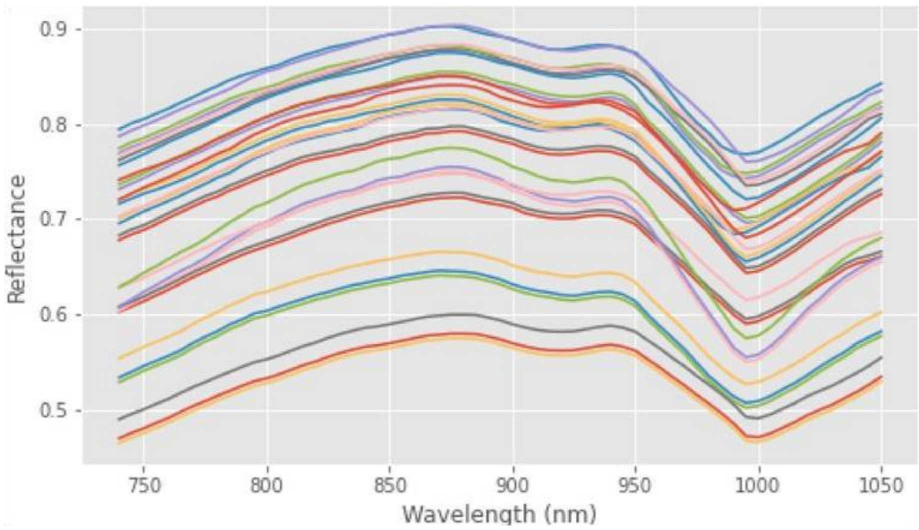


Figure 10. Raw reflectance vs wavelength

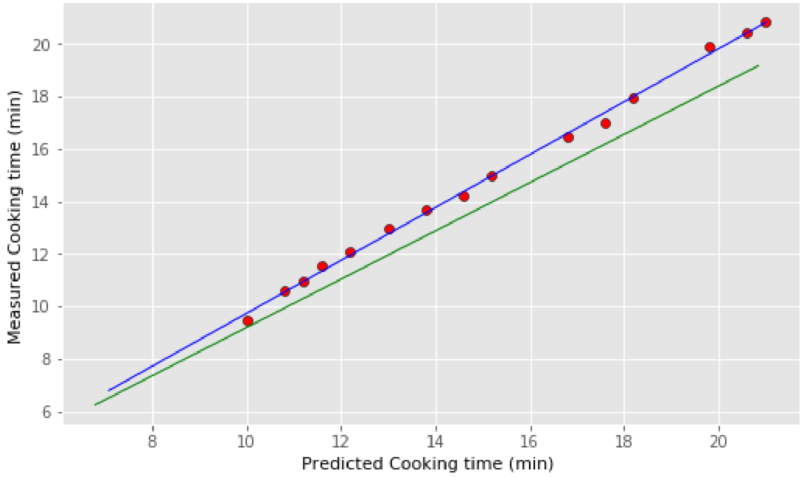
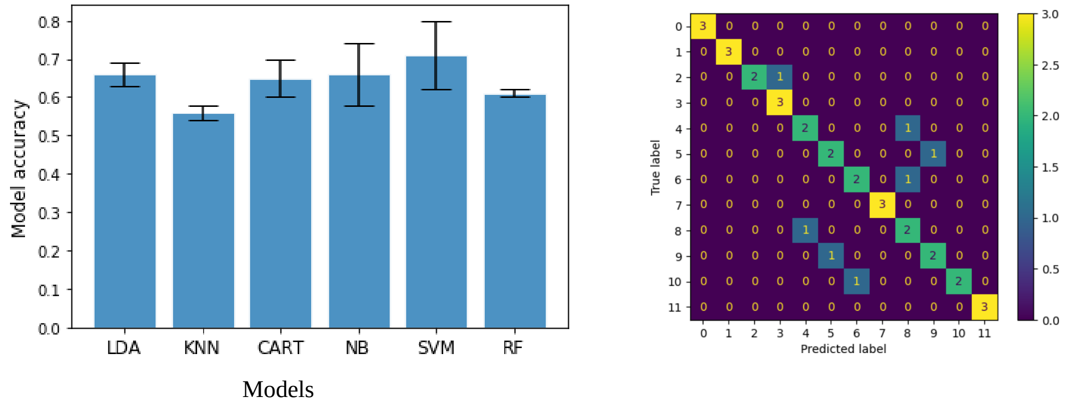


Figure 11. Cross validation plot for prediction of test data for cooking time

Therefore, ML-based PLS regression method proved quite a definitive technique in calibrating the ageing parameters of *Komal Chaul* as supported by previous work that worked on aged rice determination using NIRS [18]. This is because, in the case of limited data, PLS is more stable than other models as it can establish non-linear relationships between input and output variables by exploiting underlying latent factors.

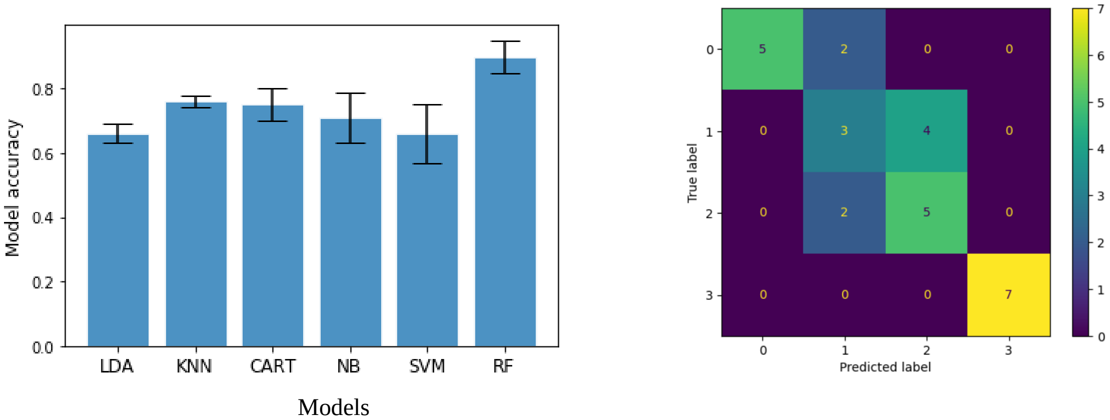
3.3. Prediction of the age of Komal Chaul by classification ML model

The ageing data were collected over a period of a year at an interval of 7 days. The spectral archive of ageing data for Komal Chaul was divided into 3 groups: 1 month, 3 months interval, 6 months interval. The classification models were then trained on the training dataset and tested over the test dataset using the k-fold validation technique. The validation accuracy was found to be highest for the SVM model (Fig. 12(a)) with a value of 0.77. This could be due to ability of SVM models to form collinearity among multilinear data [19]. The test accuracy was found to be 0.79 and the F1-score was found to be 0.69. The test dataset predictions were represented using a confusion matrix which showed that the later month predictions are coinciding (4th to 7th group) with that of the 6–9 months predictions. (Fig. 12(b)). Therefore, our next classification was based on three months classification, and the random forest classifier showed the best in classifying the data where each group consisted of spectral data from 3 months. The test accuracy was found to be 0.89 and the F1 score was found to be 0.81(Fig. 13(a)). This showed spectral data from 6-9 months (2nd group) coincides mostly, also the 0th and 1st group suggesting lower accuracy of prediction for the first 3 months (Fig. 13(b)). Lastly, a classification on the basis of six-month spectral data. The prediction accuracy was highest for RF classifier with an accuracy of 0.92 (Fig. 14(a)). for the test data, and the F1 score was highest. Therefore, the classification comparison suggests that Komal Chaul that are spectra obtained from apparently age groups of 6 months, were showing better predictability (Fig. 14(b)). This could also be due to the reason of a two-class classification problem.



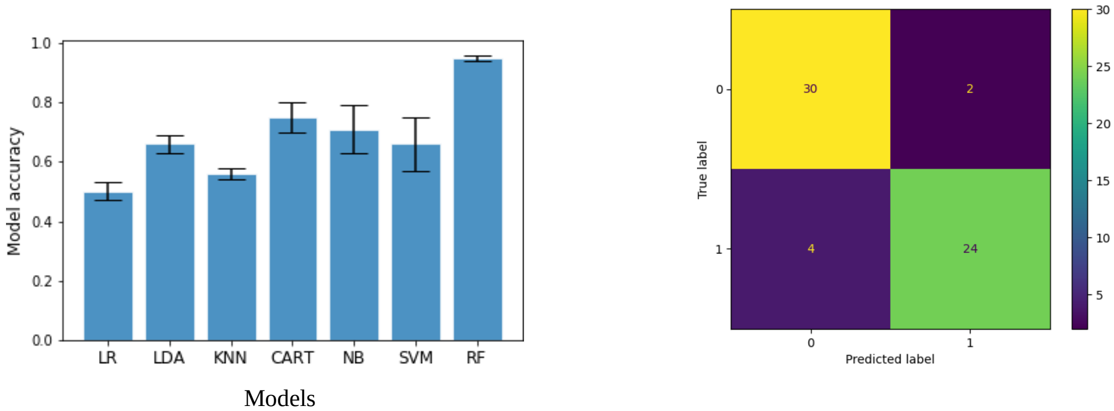
(a) (b)

Figure 12. (a) Comparison of classification of ageing time on 1-month basis; **(b)** Confusion matrix of classification of ageing time on 1-month basis



(a) (b)

Figure 13. (a) Comparison of classification of ageing time on 3-month basis; **(b)** Confusion matrix of classification of ageing time on 3-month basis



(a) (b)

Figure 14. (a) Comparison of classification of ageing time on 6-month basis; **(b)** Confusion matix of ageing time on 6-month basis

4. Conclusions

There is currently no non-destructive method that predicts the ageing of *Komal Chaul*. This study demonstrated that Near-Infrared (NIR) spectroscopy, combined with multi-regression and classification models, offers a non-destructive approach for monitoring rice quality during storage. Using wavelength selection optimization, the Partial Least Squares (PLS) regression model achieved a validation R2 of 0.89 and RMSE of 2.01 days in estimating ageing time within the tested dataset. While the model performed well, it is essential to note that PLS regression may be sensitive to multicollinearity and overfitting, especially when applied to complex or noisy datasets. Cooking quality predictions using the same model yielded a validation R2 of 0.79, indicating satisfactory but not definitive performance. Classification models provided further refinement, with Random Forest achieving the highest accuracy (92%) for distinguishing rice stored over six-month intervals. However, these findings are based solely on the Komal Chaul rice variety and specific storage conditions. Broader applicability requires further validation across diverse datasets and rice types. Future studies could explore the integration of Quantitative Structure-Activity Relationship (QSAR) models to better map storage time and cooking properties, though such efforts would require rigorous cross-validation and generalization testing.

Multidisciplinary Domains

This research covers the domains of Engineering solutions for societal needs.

Funding

The first author received financial support from the All-India Council for Technical Education (AICTE) under the AICTE National Doctoral Fellowship scheme during her Ph.D. program at Tezpur University.

Acknowledgment

The authors thankfully acknowledge the use of resources created under the All-India Council for Technical Education (AICTE) sponsored RPS project of the 2nd Author and the help and support received from the Department of Food Engineering and Technology, Tezpur University.

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.

Declaration on AI Usage

The authors declare that this article has been prepared without the use of AI tools.

References

[1] Pangging, G.; Sharma, M.B.; Sharma, C.L.; Rai, N.; Gogoi, J. Status of Geographical Indications in the Northeast region of India. *J. Intellect. Prop. Rights* **2023**, *28*, 304–311, https://doi.org/10.56042/jipr.v28i4.68113.

[2] Wahengbam, E.D.; Tongbram, T.; Hazarika, M.K. Drying characteristics of ready-to-eat komal chawal rice: processing and modeling. *J. Food Sci. Technol.* **2020**, *57*, 1698–1709, https://doi.org/10.1007/s13197-019-04199-1.

[3] Wahengbam, E.D.; Hazarika, M.K. Quality of ready-to-eat komal chawal produced by brown rice parboiling method. *J. Food Sci. Technol.* **2019**, *59*, 187–199, https://doi.org/10.1007/s13197-021-05001-1.

[4] Dutta, H.; Mahanta, C. Traditional parboiled rice-based products revisited: Current status and future research challenges. *Rice Sci.* **2014**, *21*, 187–200, https://doi.org/10.1016/S1672-6308(13)60185-1.

[5] Wahengbam, E. D., Das, A. J., Green, B. D., Shooter, J., & Hazarika, M. K. Effect of iron and folic acid fortification on in vitro bioavailability and starch hydrolysis in ready-to-eat parboiled rice. *Food chemistry*, **2019**, *292*, 39-46, https://doi.org/10.1016/j.foodchem.2019.04.044.

[6] Wahengbam, E. D., Tongbram, T., & Hazarika, M. K. Drying characteristics of ready-to-eat komal chawal rice: processing and modeling. *Journal of Food Science and Technology*, **2020**, *57*, 1698-1709.

[7] Gambhir, M. Smart India Hackathon 2022. Ministry of Education Innovation Cell. Available online: https://www.sih.gov.in/sih2022PS (accessed on 2 January 2025).

[8] Pasquini, C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Anal. Chim. Acta* **2018**, *1026*, 8–36, https://doi.org/10.1016/j.aca.2018.04.004.

[9] Jha, S.N.; Ruchi, G. Non-destructive prediction of quality of intact apple using near infrared spectroscopy. *J. Food Sci. Technol.* **2010**, *47*, 207–213, https://doi.org/10.1007/s13197-010-0031-7.

[10] Jha, S.N.; Jaiswal, P.; Narsaiah, K.; Kumar, R.; Sharma, R.; Gupta, M.; Bhardwaj, R.; Singh, A.K. Authentication of Mango Varieties Using Near-Infrared Spectroscopy. *Agric. Res.* **2013**, *2*, 229–235, https://doi.org/10.1007/s40003-013-0075-0.

[11] Jha, S.N. Near Infrared Spectroscopy. In *Nondestructive Evaluation of Food Quality: Theory and Practice*; Narsaiah, K., Ed.; John Wiley & Sons: Hoboken, NJ, USA, **2010**; pp. 141–212, https://doi.org/10.1002/9780470602636.ch5.

[12] Zhanming, L.; Jiahui, S.; Yinxing, M.; Yue, Y.; Xueming, H.; Yuanxin, G.; Jinxin, D.; Hao, D. Identification of aged-rice adulteration based on near-infrared spectroscopy combined with partial least squares regression and characteristic wavelength variables. *Food Chem.: X* **2023**, *17*, 100539, https://doi.org/10.1016/j.fochx.2023.100539.

[13] Li, M.; Qian, Z.; Shi, B.; Medlicott, J.; East, A. Evaluating the performance of a consumer scale SCiO™ molecular sensor to predict quality of horticultural products. *Postharvest Biol. Technol.* **2018**, *145*, 183–192, https://doi.org/10.1016/j.postharvbio.2018.07.017.

[14] Dutta, H.; Mahanta, C.L. Laboratory process development and physicochemical characterization of a low amylose and hydrothermally treated ready-to-eat rice product requiring no cooking. *Food Bioprocess Technol.* **2014**, *7*, 212–223, https://doi.org/10.1007/s11947-013-1061-0.

[15] Huang, F., Peng, Y., Li, L., Ye, S., & Hong, S. (2023). Near-infrared spectroscopy combined with machine learning methods for distinguishing the storage years of rice. *Infrared Physics & Technology*, **2023**, *133*, 104835.https://doi.org/10.1016/j.infrared.2023.104835.

[16] Song, J., Yu, Y., Wang, R., Chen, M., Li, Z., He, X., Ren, Z., & Dong, H. The identification of aged-rice adulteration by support vector machine classification combined with characteristic wavelength variables. Microchemical Journal, **2024**, 199, 110032. https://doi.org/10.1016/j.microc.2024.110032.

[17] Xu, L.; Zhou, Y.P.; Tang, L.J.; Wu, H.L.; Jiang, J.H.; Shen, G.L.; Yu, R.Q. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Anal. Chim. Acta* **2008**, *616*, 138–142, https://doi.org/10.1016/j.aca.2008.03.024.

[18] Shi, S.; Feng, J.; Yang, L.; Xing, J.; Pan, G.; Tang, J.; Jiang, Y. Combination of NIR spectroscopy and algorithms for rapid differentiation between one-year and two-year stored rice. *Spectrochim. Acta A* **2023**, *122343*, https://doi.org/10.1016/j.saa.2023.122343.

[19] Teye, E.; Amuah, C.L.Y.; McGrath, T.; Elliott, C. Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics. *Spectrochim. Acta A* **2019**, *217*, 147–154, https://doi.org/10.1016/j.saa.2019.01.049.