KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca/index
ISSN: 2462-1927

Comparing Machine Learning Models for Adulterant Detection in Sago Via Visible Near-Infrared Hyperspectral Imaging

Mainak Das[1,*], Wan Sieng Yeo[1], Agus Saptoro[1]

[1] Department of Chemical and Energy Engineering, Faculty of Engineering and Science, Curtin University Malaysia, CDT 250, Miri 98009, Sarawak, Malaysia

ARTICLE INFO

Article history:
Received 11 October 2025
Received in revised form 19 November 2025
Accepted 25 November 2025
Available online 8 December 2025

Keywords:
Hyperspectral imaging; machine learning; regression; food safety; sago

ABSTRACT

Sarawak is the largest producer of sago flour in Malaysia. The assessment of flour quality is generally made by its noticeable color. The excessive use of whitening chemicals, such as calcium carbonate, to alter the color and nutritional value of the product has increased the potential for food fraud. Traditional laboratory methods used to detect adulteration are expensive and time-consuming; hence, this study employed a visible near-infrared (Vis-NIR) hyperspectral camera combined with machine learning models to quantify calcium carbonate levels in sago flour rapidly. The imaging was carried out in the 400-1,000 nm regions, and the calcium carbonate concentrations used ranged from 2 w/w% to 5 w/w%. Machine learning models considered in this study were Principal Component Regression (PCR), Partial Least Square Regression (PLSR), and Multiple Linear Regression (MLR). The mean reflectance from the spectral data was used to train and test these machine learning models. Upon optimizing the hyperparameters, the PLSR model outperforms both MLR and PCR models, where its training had R^2, RMSE, and MAE values of 0.99981, 0.00008, and 0.00006, respectively. These indicate that a visible near-infrared (Vis-NIR) hyperspectral camera coupled with PLSR has the potential to be deployed in detecting adulterants in sago flour.

1. Introduction

The agriculture sector plays a crucial role in Malaysia's economy, contributing 7.3% of the country's gross domestic product in the year 2019 [1]. It is an important sector as it supplies food to the communities. Food is one of the necessities for sustaining life and for survival. The core idea of it is the crucial concept of food security, which is an important concern for consumers worldwide. Food security ensures that individuals and communities have consistent access to safe and nutritious food, regardless of external challenges. Incidents of food safety breaches, such as melamine adulteration in baby milk powder [2], contaminated cantaloupes leading to a Listeria outbreak [3], and salmonella contamination in peanut butter products [4], have raised public awareness regarding the quality of food being consumed.

[*] *Corresponding author.*
E-mail address: nafiqah1003@gmail.com

https://doi.org/10.37934/arca.41.1.1623

Sago is one of the crops in the agriculture sector that are widely cultivated and consumed in Sarawak, Malaysia. Sago plants can provide between 150 and 300 kilograms of dry starch per plant, making it a high-yielding source of edible starch. Over 42,310 hectares of sago plantation are present in smallholders in Mukah, Sarawak, making it the capital with the largest sago plantation area [5]. Sago starch is the main raw product from the plant, and it is an important ingredient for both local and international food industries. Hence, the industry and relevant authorities should take the initiative to improve quality monitoring technologies and guarantee that the sago quality is not compromised.

Despite the Food and Agricultural Organization (FAO) setting standards for edible sago starch [6], there are compromises in its quality. On the other hand, according to a study conducted by Daniela Carboni *et al.*, [7], flour tends to have lesser minerals due to the milling process, which decreases its nutritional value. This effect is compensated by fortifying the flour with mineral salts of calcium-based salts. Calcium carbonate is one such salt that prolongs the flour's shelf-life, affecting its taste and quality [8]. To gain an economic benefit, the favorable white color of calcium carbonate is preferred over the oxidized and brown sago starch. Regions with limited regulatory enforcement of food quality standards create an environment where unscrupulous producers get away with the adulteration of food products. The misuse of additives and adulterants can pose significant risks to public health, and therefore, its usage should be restricted and carefully governed. According to the Department of Standards Malaysia [9], the total starch content for premium grade sago is ≥ 95% (dry basis).

In the agriculture sector, adulterants in flour are detected by organizations that specialize in standardizing food products in conjunction with national standards set by each nation. These organizations generally use traditional laboratory techniques like liquid chromatography-mass spectrometry [8] and titration [9] to determine adulteration in food products. Despite accurate techniques, they are often time-consuming and costly to run. The instrumentation requires high maintenance and depends on skilled personnel for operation. These analytical laboratory methods also generate chemical waste, which requires proper disposal. Hence, a rapid and non-invasive technology is desirable to monitor the quality of sago to solve the above-mentioned problems faced by the industry.

Hyperspectral imaging (HSI) has recently been used in food industries as a reliable, rapid, and non-invasive monitoring technique. HSI is a non-destructive method that integrates spatial and spectral data over various wavelengths, hence giving a three-dimensional output of the image [10]. HSI integrates traditional imaging with spectroscopy, enabling each spatial pixel to contain the spectral data of the sample. As reported by Aviara *et al.*, [11], HSI is environmentally friendly, considering no chemicals are used in the imaging and processing stages when assessing the quality and safety of food and agricultural products. Some hyperspectral cameras have near-infrared spectral ranges that provide an improved understanding of the chemical constituents of ingredients in the observed sample. Moreover, HSI saves time as compared to the conventional or chemical approach when regulating food grain storage and assessing food quality. It is accurate and reliable when determining the region of interest (ROI). In recent years, studies using HSI for food safety have progressed in flour and starch products, such as quantification in multigrain flour mixes [12], determining talcum powder and benzoyl peroxide in wheat flour [13], and predicting adulteration in tapioca starch [14].

HSI is often used in conjunction with machine learning models for image processing and multivariate data analysis. Widely used machine learning models in the food industry for quality monitoring are PLSR-based and PCR-based models. For instance, PLSR was used by Lim *et al.*, [15] to detect melamine adulteration in milk powder and Khamsopha *et al.*, [14] to determine limestone powder adulteration in tapioca starch. Su *et al*., [16] used both PCR and PLSR to detect cassava flour and corn flour adulteration in wheat flour. A study conducted by Ye *et al.*, [17] suggested using Multiple Linear Regression (MLR) to model the relationship between a response variable and multiple response variables.

However, minimal studies are found to use PCR, PLSR, and MLR for detecting calcium carbonate adulteration in sago flour. Moreover, there are limited studies of HSI being used on sago flour. The prior studies showed the effectiveness of HSI with a spectral range of 900-1,700 nm (near-infrared range, NIR) in quality monitoring of flour and starch products, but inadequate studies are conducted on the 400-1,000 nm spectral range (visible-NIR range, Vis-NIR). To address these research gaps, in this study, the Vis-NIR spectral range on the images of calcium carbonate adulteration of sago flour is used to develop the PCR, PLSR, and MLR models to predict the concentrations of calcium carbonate that replace the destructive analytical laboratory methods. This study aims to compare the results of these different models to identify the best model.

2. Material and Methods

The hyperspectral images of sago were taken to test the capacity of the machine learning models. Prior to the imaging, the adulterated sago samples were prepared in the lab.

2.1 Sample Preparation

Pure sago was sourced from CRAUN Research Sdn. Bhd. in Lemantak, Sarawak. The sago was washed to obtain a starch cake and was repeated until a clean starch cake was obtained. This was filtered through a muslin cloth to eliminate excess fiber. The washed sago starch cake was sun-dried until the moisture content dropped to 20%. But to achieve the standards set by FAO to below 13% moisture content [6] and to avoid browning from oxidation of the sago, the starch cake was further dried in an oven until moisture content dropped below 13%. The sago was sieved to obtain fine particles upon achieving the desired moisture content. The chemical additive used for this study was calcium carbonate, and the anhydrous calcium carbonate was purchased from Merck. In a previous study, Lee *et al*., [18] adulterated sago flour with calcium hypochlorite concentrations ranging from 0.005 w/w% to 2.0 w/w%. Their study focused on smaller calcium hypochlorite concentrations. However, this study used a different adulterant, namely calcium carbonate. Moreover, for this study, the higher concentrations of calcium carbonate added to sago flour were 2.0 w/w%, 2.5 w/w%, 3.0 w/w%, 3.5 w/w%, 4.0 w/w%, 4.5 w/w%, and 5.0 w/w%, respectively. The desired amount of sago and adulterant was loaded into a 50 mL centrifuge tube and mixed using a vortex mixer at 2,000 rpm for 15 minutes. Three replicates were prepared for each concentration to increase the size of the dataset and reduce the chances of any anomalies. Hence, with seven different concentrations, the number of adulterated sago samples prepared in this study was 21.

2.2 Hyperspectral Image Acquisition

After preparing the adulterated sago samples, a VIS-NIR HSI camera (Resonon Pika L, MT, USA) is used in this study. It has a spectral range of 400-1,000 nm measured across 300 spectral bands, with a spectral resolution of 3.3 nm. An objective lens with a focal length of 23 mm and an aperture of f/2.4 was attached to the spectrograph. The frame Rate and shutter time were set to be 50 frames/s and 16.49 ms, respectively. The HSI camera has a bench-top view of the sample with strong lighting provided by halogen lamps. The adulterated sago samples were placed on a petri dish to the brim.

The petri dish was kept at the center of the stage and scanned line by line at a speed of 4.5 mm/s, with 700 lines of capture. The distance between the camera and the sample was approximately 0.2 m.

The acquired hyperspectral images were loaded onto a laptop connected to the HSI setup with the 'Spectronon' software installed. This software was utilized to conduct radiometric calibration on the HSI data to reduce signal noise. A sample of the image output from this software is shown in Fig. 1, which is further discussed in the results section. This data was loaded into MATLAB R2023a in a computer with Intel Core i7 (6th Generation), GTX1060 (6GB GDDR5), and 16GB DDR4 memory for image segmentation and determining the ROI. The central region of the image was a suitable ROI, and hence, a dimension of 300×300 pixels was chosen. The selected ROI was further sub-sampled into nine different samples with dimensions of 100×100 pixels each to increase the size of the dataset. For each subsample, the mean reflectance across all the wavelengths was calculated to obtain the mean spectrum, and this was extracted into a separate file.

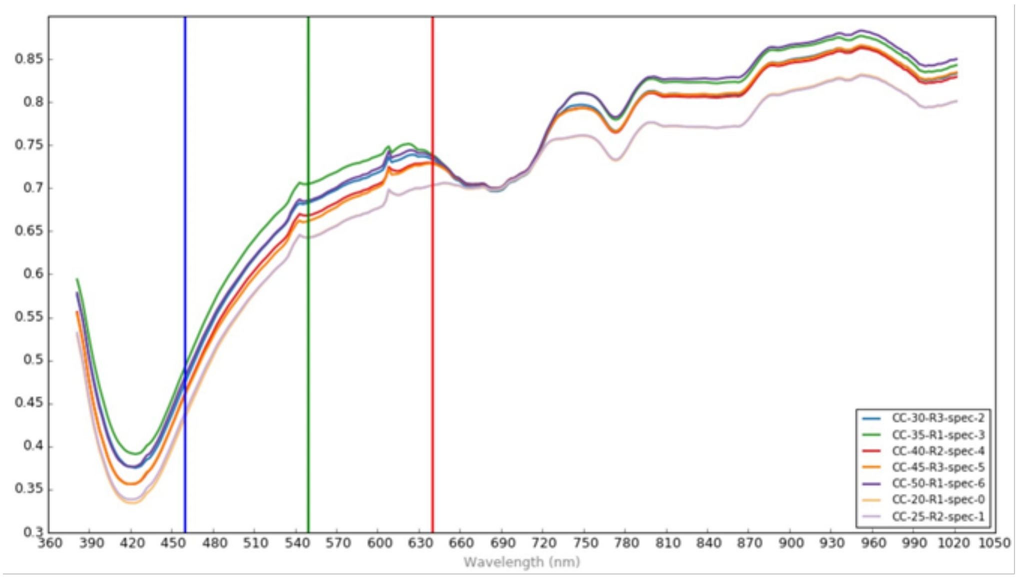


Fig. 1. Spectral profiles of sago samples adulterated with different concentrations of calcium carbonate

2.3 Regression Models

PCR and PLSR are modeling algorithms used to establish the quantitative relationship between the spectral data and the concentration of calcium carbonate in sago flour. These models are preferable when used with high-dimensional multicollinear data, but a simpler model like MLR works well with a small number of variables and assumes a linear relationship between the predictor and target variable [19]. The hyperparameters of the PCR and PLSR models, like Principal Components (PCs) and Latent Variables (LVs), were optimized using 189 sub-samples. The sub-samples were split for training and testing datasets, and the chosen splitting ratio was 7:3 for train:test [20]. These models were evaluated with error metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). The model development and evaluation process were performed using MATLAB version R2023a.

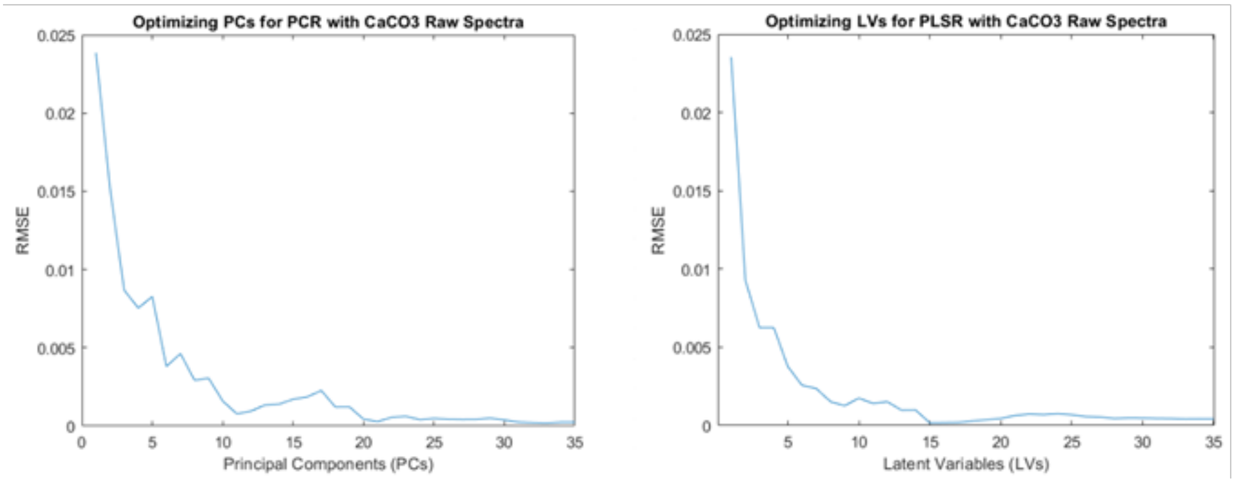
3. Results and Discussions

3.1 Reflectance Data

The reflective spectral data for different concentrations of calcium carbonate adulteration from 2 w/w% to 5 w/w% is shown in Fig. 1. The blue, green, and red lines in the graph represent the regions where the respective color is present in the visible light spectrum. The region of 660 nm to 730 nm is where the spectra of different concentrations converged. Additional peaks are also noticed at 420 nm, 540 nm, 610 nm, 780 nm, and 940 nm. 420 nm is between violet and indigo on the visible light spectrum, whereas 610 nm falls in the orange region. 780 nm is at the end of the red region in the visible light spectrum, whereas 940 nm is the near-infrared region. The peaks at these visible light regions do not quantitatively determine the relationship between the color and its influence on the visual characteristics of sago, but it suggests that these regions are more prominent than others when determining the visual characteristics of sago flour. The peak after the visible light spectrum is potentially from organic bonds within the sago flour that influence the spectral data.

3.2 Regression Analysis

The regression models used in this study were PCR, PLSR, and MLR. The hyperparameters like LV and PC were optimized for the PCR and PLSR models to create a precise and efficient model. Optimizing the number of LVs is a crucial step, as having a lesser LV would cause model underfitting, and having a higher LV would cause model overfitting. The high number of LVs also comes with computational efficiency issues due to increased load. Fig. 2 shows the hyperparameter optimization for the PCR and PLSR models, with RMSE as an error metric for indicating the predictive capability of the model. The RMSE value determined the optimum number of LVs and PCs. A lower RMSE value indicates high accuracy and good model performance [21]. Hence, as the RMSE value approached a minimum, the LVs and PCs were optimal. However, after a certain number of hyperparameters, the performance in the model got worse, potentially due to overfitting. Therefore, the optimal number of PCs was 21, whereas LVs were 15. This was chosen based on the presence of overfitting peaks in Figs. 2(a) and 2(b). After 21 PCs, there was a slight peak, and after 15 LVs, the RMSE value rose again. Once the optimal number of hyperparameters was determined, the PCR and PLSR models were run with the extracted hyperspectral data.



(a) (b)

Fig. 2. Hyperparameter optimization of (a) PCR, and (b) PLSR models

Table 1 shows the optimum number of LVs for PCR and PLSR and their respective error metrics compared to the performance of the MLR model. Upon examining the overall performances of the PCR, PLSR, and MLR models, the RMSE and MAE error metrics showed that the PLSR model performed the best. Although from Table I, MLR displayed the best results for training with an RMSE value as low as 4.0873×10^{-13} and an MAE value of 3.1645×10^{-13}. As for the PCR and PLSR models during the training dataset, they returned values of 2.8136×10^{-4} and 1.6942×10^{-4} for RMSE, while 2.1634×10^{-4} and 1.2639×10^{-4} for MAE, respectively. It is apparent from this table that there is a discrepancy between the testing dataset's RMSE and MAE values in PCR, PLSR, and MLR. For MLR, the testing dataset's RMSE is 0.3469, and MAE is 0.2714. Meanwhile, for the PCR and PLSR models, the RMSE and MAE values were much lower than MLR which are between 9.9402×10^{-5} and 6.0609×10^{-5}. These results also show that based on the testing dataset's RMSE and MAE values, PLSR performed slightly better than PCR because PLSR considers both input and output variables, while PCR only takes into account the input variables [22].

Additionally, as seen from Table 1, with the training datasets, the MLR had the highest R^2 of approximately 1, whereas the PCR and PLSR both had slight lower R^2 values of 0.9999. These results indicate that these models have similar predictive performance for the training datatset. However, a large degree of multicollinearity and redundancy occur in hyperspectral images [17]. MLR is not susceptible to the multicollinearity of data; therefore, when predicting the concentration of adulterant in sago flour with the testing dataset that is not used for model development, the R^2 value of MLR was much lower than PCR and PLSR. For testing datasets, the PCR and PLSR models with optimized hyperparameters returned values of 0.9995 and 0.9998 and outperformed the MLR model, which had an R^2 of 0.8817. From these results, it can be concluded that PLSR has better overall results as compared to MLR and PCR.

Table 1
Performance metrics of MLR, PCR, and PLSR models on adulterated sago samples

Model	RMSE (train)	RMSE (test)	MAE (train)	MAE (test)	R^2 (train)	R^2 (test)
MLR	4.0873×10^{-13}	0.3469	3.1645×10^{-13}	0.2714	~1	0.8817
PCR *PC = 21*	2.8136×10^{-4}	9.9402×10^{-5}	2.1634×10^{-4}	7.7420×10^{-5}	0.9999	0.9995
PLSR *LV = 15*	1.6942×10^{-4}	7.6768×10^{-5}	1.2639×10^{-4}	6.0609×10^{-5}	0.9999	0.9998

4. Conclusions

This study used a hyperspectral camera within the 400-1,000 nm Vis-NIR spectral range to capture hyperspectral images of sago flour samples adulterated with calcium carbonate. The raw spectral data was analyzed, and the mean reflectance spectrum was extracted from the ROI of this raw data. The images were further sub-sampled to increase the size of the dataset. The extracted data was then tested to predict the concentration of calcium carbonate in sago flour using MLR, PCR and PLSR models. For PCR and PLSR, their hyperparameters, PCs, and LVs were optimized, respectively, and the models produced better predictive results than MLR, especially for the testing dataset. To conclude the overall results for training and testing datasets, PLSR performs better than MLR and PCR in predicting calcium carbonate concentration in sago flour over a spectral range of 400-1,000 nm. These results demonstrate that PLSR is more suitable to be used for the efficacy of Vis-NIR hyperspectral imaging in providing detailed and non-destructive insights into sago flour properties, surpassing the limitations of conventional approaches.

Acknowledgement

The authors would like to acknowledge the financial support given by the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2022/TK05/CURTIN/03/2) from the Ministry of Higher Education (MOHE), Malaysia, and Curtin University Malaysia. The financial support provided by FRGS and Curtin University Malaysia is deeply appreciated.

References

[1] Shaari, Mohd Shahidan, Paul Anthony Mariadas, Nor Ermawati Hussain, and Uma Murthy. "The Effect of Energy Consumption in the Agricultural Sector on CO2 Emissions in Malaysia." International Energy Journal 21, no. 4 (2021).

[2] Sun, Fengxia, Liqiang Liu, Hua Kuang, and Chuanlai Xu. "Development of ELISA for melamine detection in milk powder." *Food and agricultural immunology* 24, no. 1 (2013): 79-86. https://doi.org/10.1080/09540105.2011.641170

[3] Laksanalamai, Pongpan, Lavin A. Joseph, Benjamin J. Silk, Laurel S. Burall, Cheryl L. Tarr, Peter Gerner-Smidt, and Atin R. Datta. "Genomic characterization of Listeria monocytogenes strains involved in a multistate listeriosis outbreak associated with cantaloupe in US." (2012): e42448. https://doi.org/10.1371/journal.pone.0042448

[4] Lathrop, Amanda A., Tiffany Taylor, and James Schnepf. "Survival of Salmonella during baking of peanut butter cookies." *Journal of food protection* 77, no. 4 (2014): 635-639. https://doi.org/10.4315/0362-028X.JFP-13-408

[5] Wahed, Zulhakim, Annie Joseph, Hushairi Zen, and Kuryati Kipli. "Sago Palm Detection and its Maturity Identification Based on Improved Convolution Neural Network." *Pertanika Journal of Science & Technology* 30, no. 2 (2022): 1219-1236. https://doi.org/10.47836/pjst.30.2.20

[6] Nations, Food and Agriculture Organization of the United Nations. "Codex Alimentarius - International Food Standards." In Regional Standard for Edible Sago Flour, 2011.

[7] Carboni, Angela Daniela, Andrea Gómez-Zavaglia, Maria Cecilia Puppo, and María Victoria Salinas. "Effect of Freezing Wheat Dough Enriched with Calcium Salts with/without Inulin on Bread Quality." *Foods* 11, no. 13 (2022): 1866. https://doi.org/10.3390/foods11131866

[8] Rossi, Matías G., Marina Soazo, Gisela N. Piccirilli, Emilce E. Llopart, Gilda C. Revelant, and Roxana A. Verdini. "Technological, nutritional and sensorial characteristics of wheat bread fortified with calcium salts." *International Journal of Food Science & Technology* 55, no. 10 (2020): 3306-3314. https://doi.org/10.1111/ijfs.14594

[9] Malaysia, Department of Standards. "Edible Sago Starch - Specification (Second Revision)." 2012

[10] Mendez, Jeffrey, Liz Mendoza, J. P. Cruz-Tirado, Roberto Quevedo, and Raúl Siche. "Trends in application of NIR and hyperspectral imaging for food authentication." *Scientia Agropecuaria* 10, no. 1 (2019): 143-161. https://doi.org/10.17268/sci.agropecu.2019.01.16

[11] Aviara, Ndubisi A., Jacob Tizhe Liberty, Ojo S. Olatunbosun, Habib A. Shoyombo, and Samuel K. Oyeniyi. "Potential application of hyperspectral imaging in food grain quality inspection, evaluation and control during bulk storage." *Journal of Agriculture and Food Research* 8 (2022): 100288. https://doi.org/10.1016/j.jafr.2022.100288

[12] Blanch-Perez-del-Notario, Carolina, Wouter Saeys, and Andy Lambrechts. "Fast ingredient quantification in multigrain flour mixes using hyperspectral imaging." *Food control* 118 (2020): 107366. https://doi.org/10.1016/j.foodcont.2020.107366

[13] Fu, Xiaping, Jinchao Chen, Feng Fu, and Chuanyu Wu. "Discrimination of talcum powder and benzoyl peroxide in wheat flour by near-infrared hyperspectral imaging." *Biosystems engineering* 190 (2020): 120-130. https://doi.org/10.1016/j.biosystemseng.2019.12.006

[14] Khamsopha, Duangkamolrat, Sahachairungrueng Woranitta, and Sontisuk Teerachaichayut. "Utilizing near infrared hyperspectral imaging for quantitatively predicting adulteration in tapioca starch." *Food Control* 123 (2021): 107781. https://doi.org/10.1016/j.foodcont.2020.107781

[15] Lim, Jongguk, Giyoung Kim, Changyeun Mo, Moon S. Kim, Kuanglin Chao, Jianwei Qin, Xiaping Fu, Insuck Baek, and Byoung-Kwan Cho. "Detection of melamine in milk powders using near-infrared hyperspectral imaging combined with regression coefficient of partial least square regression model." *Talanta* 151 (2016): 183-191. https://doi.org/10.1016/j.talanta.2016.01.035

[16] Su, Wen-Hao, and Da-Wen Sun. "Evaluation of spectral imaging for inspection of adulterants in terms of common wheat flour, cassava flour and corn flour in organic Avatar wheat (Triticum spp.) flour." *Journal of Food Engineering* 200 (2017): 59-69. https://doi.org/10.1016/j.jfoodeng.2016.12.014

[17] Ye, Xujun, Shiori Abe, and Shuhuai Zhang. "Estimation and mapping of nitrogen content in apple trees at leaf and canopy levels using hyperspectral imaging." *Precision agriculture* 21 (2020): 198-225. https://doi.org/10.1007/s11119-019-09661-x

[18] Lee, Ming Hao, Agus Saptoro, King Hann Lim, Han Bing Chua, Tuong Thuy Vu, Nurleyna Yunus, and Hasnain Hussain. "Feasibility of Visible Near-Infrared Hyperspectral Imaging in Detection of Calcium Hypochlorite in Sago Flour." In *MATEC Web of Conferences*, vol. 377, p. 01005. EDP Sciences, 2023. https://doi.org/10.1051/matecconf/202337701005

[19] Yeo, Wan Sieng, Agus Saptoro, Perumal Kumar, and Manabu Kano. "Just-in-time based soft sensors for process industries: A status report and recommendations." *Journal of Process Control* 128 (2023): 103025. https://doi.org/10.1016/j.jprocont.2023.103025

[20] Thien, Teck Fu, and Wan Sieng Yeo. "A comparative study between PCR, PLSR, and LW-PLS on the predictive performance at different data splitting ratios." *Chemical Engineering Communications* 209, no. 11 (2022): 1439-1456. https://doi.org/10.1080/00986445.2021.1957853

[21] Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate research* 30, no. 1 (2005): 79-82. https://doi.org/10.3354/cr030079

[22] Yeo, Wan Sieng. "Prediction of yellowness index using partial least square regression model." In *2021 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, pp. 1-5. IEEE, 2021. https://doi.org/10.1109/GECOST52368.2021.9538723