Classification of crop flours based on protein contents using near infra-red spectroscopy and principle component analysis

I Yuliyanda¹, R E Masithoh¹,², N Khuriyati¹,³, A D Saputro¹

¹Department of Agricultural and Biosystems Engineering, Faculty of Agricultural Technology, Universitas Gadjah Mada, Indonesia
²Corresponding author
³Department of Agroindustrial Technology, Faculty of Agricultural Technology, Universitas Gadjah Mada, Indonesia

E-mail: evi.ugm@gmail.com

Abstract. Indonesia has abundant crops that can be used as carbohydrate sources. Those crops are made into flour to prolong the shelf life, as well to ease for handling and cooking. Crops as carbohydrate sources are usually high energy but low protein. The aim of this research was to classify flours made of various crops using near infra-red spectroscopy (NIRS) and principle component analysis (PCA). The samples used in this study were six types of flour made of banana, breadfruit, taro, arrowroot, purple sweet potato, and modified cassava (mocaf). The reflectance data were taken using the NIRFlex N500 Fiber Optic Solids Cell at wavelengths of 1000-2500 nm. The spectral obtained were pre-processed and analyzed using The Unscrambler X version 10.5.1. Three pre-processing methods were used, i.e. 1st Savitzky Golay Derivative, Normalization, and Standard Normal Variate (SNV). PCA was able to classify flours based on types of crops. The best transformation was SNV which was able to classify all groups of samples with 100% success rate. PCA model was also able to differentiate low and high protein level of samples aligned with the chemical analysis.

1. Introduction
Rice consumption in Indonesia is categorized high although there is a slight decline in rice consumption due to conversion to wheat based products such as bread and instant noodle [1]. On the other hand, Indonesia produces other carbohydrate source crops, such as arrowroot, cassava, sweet potato, banana, corn, or breadfruit. Those crops are high in carbohydrate but lower glycemic index (GI) compared to conventional rice [2]. However, carbohydrate sources such as cassava and tubers tend to have lower protein [3,4] compared to rice. Therefore, it is a challenge to find carbohydrate sources which have high carbohydrate with low GI but high protein.

In general, determination of protein uses chemical analysis such as Kjeldahl’s method, Bradford’s method and a modified version of the Lowry method [5] or can be measured using HPLC method [6]. Another method is using UV spectrometer [7] which requires the determination of specific absorption value for a given protein. Those methods involve sample preparation and time consuming.
Currently, rapid and non-destructive methods for quantity and quality measurement is developing which utilize near infra-red spectroscopy (NIRS) [8]. NIRS is a spectroscopic method that utilizes the NIR range of electromagnetic spectrum, around 700 - 2,500 nm. When a radiation hits a sample, the radiation will be reflected, absorbed, or transmitted. The amount of radiation depends on the chemical composition and physical properties of the product. Most absorption in the NIR spectrum are overtones or combinations of several absorption classes caused by vibrational or rotational transitions. In some agricultural products, the spectrum produced is very similar to each other due to similar chemical composition in the products. The similarity requires reliable calibration techniques to extract information from the NIR spectrum such as multivariate analysis including partial least square (PLS) regression or principle component analysis (PCA). Several studies to analyze chemical compounds using NIRS for fruits and vegetables have been carried out. NIRS was used to detect citric acid and tartaric acid in oranges [9] or total acids in lemons and oranges [10], or to measure mango soluble solid content (SSC) [11], carotenoids in raw salads [12] and lycopene in watermelon and tomato porridge [13].

PCA is one method in multivariate analysis developed to reduce data dimensions. PCA reduces number of dimensions or multivariate variables, by simulating spectra data. The results of this variable reduction are used as input variables for further analysis, such as classification. [14] applied multivariate analysis including PCA to classify green coffee beans on continents and countries origin. [15] used Fourier transform infrared (FTIR) spectroscopy and PCA were applied to classify biomasses through the composition of their bio-oil. By applying PCA, classification of samples or products can be used based on their NIR spectra. This research was aimed to classify flour made of various crops, i.e. banana, breadfruit, taro, arrowroot, purple sweet potato, and modified cassava flour, as well as to classify those flours based on the level of protein by using NIRS spectra information.

2. Materials and Methods

2.1. Materials and Methods

2.1.1. Materials. Total sample for this research were 72 flour samples made of 6 types of crops such as banana, breadfruit, taro, arrowroot, purple sweet potato, and modified cassava flour. Samples were obtained from Women Farmer Group (Kelompok Wanita Tani Melati) at Yogyakarta.

2.1.2. Determination of protein. Each crop consisted of 12 samples, so in total there were 72 samples. Protein content was determined using Kjeldahl method [16] which in short consisted of conversion wet digestion nitrogen to ammonium sulfate, neutralization to free ammonia, distillation ammonia into boric acid, titration with alkali, conversion nitrogen concentration to protein using conversion ratio, then finally determination of ‘true protein’ by precipitating out protein, analyzing remaining nitrogen, subtracting from total nitrogen content. All experiments were conducted in triplicate.

2.1.3. Spectral acquisition. Approximately 4 g of a sample was weighed and transferred to a glass sample container (with a height of 10 mm and a diameter of 2 cm). To maintain uniformity of sample during spectra measurement, the sample was pressed with a glass slide beforehand. Near-infrared reflectance spectra of the samples were obtained in the reflectance mode on a NIRFlex N500 Fiber Optic Solids Cell spectrometer (Büchi Labortechnik AG, Flawil, Switzerland) with InGaAs detector. Spectra were acquired in reflectance mode in the long-wavelength region of 4000 – 10,000 cm⁻¹ (2500 – 1000 nm) with a 2 nm sampling interval resulted in total 1501 wavelengths. The NIRS were recorded at ambient temperature and a total of 72 spectra of samples were collected.
2.1.4. Data analysis. Spectra collection was obtained using the NIRWare 1 Software which was then converted into Excel file for worksheet calculation and data plot. The raw data obtained from the NIR instrument were in reflectance values which were then converted into absorbance values. The absorbance data was then imported to the Unscrambler X version 10.5.1 (CAMO Software AS, Oslo, Norway) for multivariate analysis, including spectra pre-treatment and PCA analysis. Before applying PCA, several pre-treatments, i.e. 1st Savitzky-Golay Derivative, 2 Polynomial Order, 25 Smoothing Points at 1350 – 2500 nm, Normalize, and Standard Normal Variate (SNV) were conducted.

3. Results and Discussion

3.1.1. Total protein content of flours. Results of total protein and moisture content of all samples are shown in Table 1. Bread fruit have the highest moisture content, which are 9.80%, whereas purple sweet potato flour has the lowest moisture content of 7.95%. In term of protein content, taro flour contained the highest protein of 5.15% and modified cassava (mocaf) flour contained the lowest protein content of 0.8%.

![Spectra acquisition using NIRFlex N500 Fiber Optic Solids Cell spectrometer](image)

**Figure 1.** Spectra acquisition using NIRFlex N500 Fiber Optic Solids Cell spectrometer

| Parameter (\%wb) | Modified cassava | Banana | Purple sweet potato | Arrow root | Taro | Bread fruit |
|------------------|------------------|--------|--------------------|------------|-----|-------------|
| Total protein    | 0.80 ± 0.14      | 1.77 ± 0.27 | 2.53 ± 0.06      | 4.70 ± 0.3 | 5.15 ± 0.27 | 3.67 ± 0.2  |

Protein values as shown in Table 1 are then classified into two classes based on protein content as shown in Table 2. Of all samples, flour made of taro, arrowroot, and bread fruit are categorized as ‘high protein’ having protein higher than 3% (wb), while purple sweet potato banana and mocaf were categorized as ‘low protein’ content less than 3% (wb).

| High >3% | Low <3% |
|----------|---------|
| Taro, Arrowroot, Bread fruit | Purple sweet potato Banana, Modified Cassava |

| Table 2. Class of flour based on total protein (in % wet basis) |

3
3.2. Raw spectra

Figure 2(a) showed average spectra of all types of flours. It can be seen that the spectra were similar. Those trends also similar to spectra of ground wheat showing absorption band of moisture at 1440 – 1470 nm and 1920-1940 nm as well as protein at 2148-2200 nm as shown in Figure 2(b) [17][3].

![Average spectra (a) of mocaf, arrowroot, bread fruit, banana, sweet potato, and taro flour (data from the experiments), and typical NIR spectra of ground wheat (Manley, 2014) (b)](image)

3.3. Principle Component Analysis (PCA)

3.3.1. Raw data spectra. PCA analysis on raw spectra in Figure 3(a) showed the scores plot in the first principal component (PC) for the flour samples explaining 98% of the total variability. Figure 3(b) shows the loadings plot using PC1. In the scores plot can be seen the uniformity of samples, in which the uniform samples are closed in the plot, as for taro, sweet potato, and arrowroot. On the other hand, samples of banana, mocaf, and bread fruit flours are far away from each other as indication of non-uniformity. The non-uniformity flour may be caused by difference in variety or process. The loading plot indicates the absorption bands mostly contributed to flour classification according the PC1 value.

By using untreated spectra data, percentage of classification of flour were 100% for banana flour, bread fruit, and cassava flour, as well as 92% for purple sweet potato flour. On the other hand, there were overlapped on sample groups for taro and arrowroot. For raw untreated data, the classification was obtained using 1 PC indicated by absorption band of protein at wavelength of 2110 nm.

![Scores plot in the first principal component (PC) (a) and loadings plot using PC1 (b)](image)
3.3.2. Pre-treatment spectra. Spectra pre-treatment techniques were used to remove any irrelevant spectra information from which regression technique cannot handle (4). There were four spectra pre-treatments used in this research, namely 1st Savitzky-Golay (SG) derivatives, Normalization, and Standard Normal Variate (SNV), as shown in Figure 4(a), (b), and (c). In general, all pre-treatments can classify flours with maximum two PCs (PC1 and PC2) as shown in Figure 4 which explain 87%, 100%, and 89% of variability for 1st Savitzky-Golay derivative, normalization, and SNV, respectively.

By using 1st Savitzky-Golay (SG) derivative (Figure 4(a)), the PC2 with 15% of total variance is able to classify sweet potato, banana, taro, and mocaf flour, but is not really clear in separating arrowroot and bread fruit flour. Sweet potato flour has a high positive score in the PC2, while arrowroot and bread fruit flour have the negative score in the PC2. Taro flour has the score around 0 (zero) on the PC2. PC1 with 72% of total variance, clearly separate banana flour in the negative as well as sweet potato and bread fruit in the positive side. Other flours are on the zero value of PC1. The loadings of PC 1 (Figure 4(b) (blue line)) have positive effect at the wavelength of 1956 and 2164 nm which are linked to the moisture and protein. There are several peaks according to PC2 loadings (Figure 4(b) (red line)) at wavelength of 1424 – 2251 nm which are associated with moisture, amylose, and protein [18, 19]. PCA using 1st SG derivative is able to classify banana and purple sweet potato flours without integrated with other flours.

![Figure 4](image_url)

Figure 4. PCA classification (a) and loading values of PC1 (blue line) and PC2 (red line) (b) using the 1st Savitzky-Golay Derivative, 2 Polynomial Order, 25 Smoothing Points at 1350 – 2500 nm

It is clear that PC1 in Figure 5(a) after normalization transformation can describe 99% of variance of the samples. PC1 is able to separate taro and bread fruit flours in the negative value, as well as sweet potato, banana, and mocaf in the positive value. Arrowroot cannot clearly distinguished by PC1, as it spreads around negative value and zero. PC2 only accounts for 1% of variance. It can distinguished sweet potato and taro flours based on its positive value, and arrowroot flour on zero value. However, PC2 score of breadfruit and banana flours are on the same value. This also applies on some samples of mocaf flour by which are grouped with breadfruit and banana flours. Both, the PC1 and PC2 loadings (Figure 5(b)) cannot show significant peaks except they have a highest negative value at 1944 nm which is assigned to moisture. The PC1 loading shows a positive value at 2214 nm which is assigned to protein. PCA using Normalization transformation is able to classify tato, sweet potato, mocaf, and bread fruit flours successfully. Only small amount of arrowroot flour which is joined together with taro flour (Figure 5(a)).
Variance due to sample particle size is higher than chemical composition. The multiplicative interferences of particle size can be effectively removed by SNV transformation (5). The PC1 and PC2 in Figure 6(a) explains 64% and 25% of variance, respectively. PC1 separates sweet potato, arrowroot, and taro flours clearly on the negative values, in opposite with bread fruit, banana, and mocaf on the positive values. Only few samples of arrowroot and bread fruit which are on the zero value of PC1 axis. The PC2 which accounts for 25% of variance is able to separate taro, bread fruit, and arrowroot flours on the positive values, while sweet potato, banana, and mocaf flours on negative values. The PC1 and PC2 loadings (Figure 6(b)) shows significant peak at 1937 nm which is related to moisture as well as 2107 and 2228 nm which are assigned to protein. SNV transformation is able to classify all flours successfully (Figure 6(a)).

Since SNV provide the best PCA classification result, it is then used to make a model for classification of low and high level of protein from crops. The value of low and high level protein is presented in Table 2. Flours made of taro, arrowroot, and bread fruit are categorized as ‘high protein’ having protein higher than 3% (wb), while purple sweet potato banana and mocaf were categorized as ‘low protein’ content less than 3% (wb). Figure 7(a) shows that PCs (PC1 and PC2) explains 89% of variability. By using the selected PCA model, low level protein (blue sign) are located on the negative side of PC2 and high level protein (red sign) are located on the positive side of PC2.
4. Conclusions

Scores and loadings of PC1 and PC2 can be used to classify type of flours based on their crop source. SNV transformation followed by PCA method is able to differentiate low and high level protein of various crop flour using NIR spectra which conforms with chemical protein analysis.

Acknowledgement

This research was funded by Universitas Gadjah Mada under Recognisi Tugas Akhir program 2019 No: 2129/UN1/DITLIT/DIT-LIT/LT/2019. Great appreciation is delivered to Nafisatul Maghfiroh, Win Pribadi, and Tafip Hariyanto (UKSW Salatiga) for their support and assistance in finishing this work.

References

[1] McDonald G and Meylinah S 2018 Indonesia Grain and Feed Annual Report 2018. Vol. ID1808, GAIN Report Number: 2018.
[2] Ebere RA, Imungi JK and Kimani VN 2017 Glycemic Indices of Cassava and Sweet Potatoes Consumed in Western Kenya Food Sci Qual Manag 63 p 7–12.
[3] Aryee FNA, Oduro I, Ellis WO and Afuakwa JJ 2006 The physicochemical properties of flour samples from the roots of 31 varieties of cassava Food Control 17(11) p 916–22.
[4] Ugwu FM 2009 The potential of cassava as weaning food Pakistan J Nutr. 8(10) p 1701–5.
[5] Maehre HK, Dalheim L, Edvinsen GK, Elvevoll EO and Jensen IJ 2018 Protein Determination—Method Matters. Foods. 7 (2) p 5.
[6] Bietz JA 1985 High performance liquid chromatography: how proteins look in cereals. Cereal Chem. 62(3).
[7] Aitken A and Learmonth MP 2002 Protein determination by UV Absorption. In: Walker JM, editor. The protein protocol handbook (Humana Press).
[8] Nicolai BM, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron KI and Lammertyn J 2007 Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review Postharvest Biol Technol 46 p 99–118.
[9] Cen H, Bao Y, He Y and Sun D 2007 Visible and near infrared spectroscopy for rapid detection of citric and tartaric acids in orange juice J of Food Eng. 82 p 253–60.
[10] Masithoh R, Haff R and Kawano S 2016 Determination of soluble solids content and titratable acidity of intact fruit and juice of satsuma mandarin using a hand-held near infrared instrument in transmittance mode J Near Infrared Spectrosc. 24(1) p 83.
[11] Jha SN, Chopra S and Kingsly ARP 2005 Determination of Sweetness of Intact Mango using Visual Spectral Analysis Biosyst Eng. 91 p 157–61.
[12] Niizu PY and Rodriguez-Amaya DB 2005 New data on the carotenoid composition of raw salad vegetables J Food Compos Anal. 18 p 739–49.
[13] Choudhary R, Bowser TJ, Weckler P, Maness NO and Mcglynn W 2009 Rapid estimation of lycopene concentration in watermelon and tomato puree by fiber optic visible reflectance spectroscopy. Postharvest Biol Technol 52(1) p 103-109
[14] Giraudo A, Grassi S, Savorani F, Gavoci G, Casiraghi E and Geobaldo F 2019 Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. Food Control 99 137–45.
[15] Lazzari E, Schena T, Caetano M, Marcelo A, Tatiane C, Nunes A, Ferrão MF, Bjerk T and Caramão EB 2018 Classification of biomass through their pyrolytic bio-oil composition using FTIR and PCA analysis. Ind Crop Prod. 111 p 856–64.
[16] Horwitz W and Latimer GW 2005 Official Methods of Analysis of AOAC International. AOAC International (Maryland).
[17] Manley M 2014 Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. Chem Soc Rev. 43(24) p 8200–14.
[18] Bagchi TB, Sharma S and Chattopadhyay K 2016 Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. Food Chem. 191 p 21–7.
[19] Barnes RJ, Dhanoa MS and Lister S 1989 Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc. 43(5) p 772–7.