Caffeine content calibration model on green beans arabica Mandailing Natal coffee using NIRS and artificial neural network

J Akhir, A Rindang* and P C Ayu

Department of Agricultural Engineering, Faculty of Agriculture, Universitas Sumatera Utara, Medan, Indonesia.

E-mail: *adian.rindang@usu.ac.id

Abstract. Generally, Near Infrared Reflectance (NIR) spectroscopy is a non-destructive method that can identify the chemical content of agricultural products. In this study, several pretreatment data were used to reduce some noises from original NIR spectra, namely normalization (0-1) and Multiple Scatter Correction (MSC) and calibration model of caffeine content using Artificial Neural Network (ANN). Result from wavelength analysis showed that the caffeine content of Mandailing Natal green bean arabica coffee was in the wavelengths of 1,208.9 nm and 1,728.91 nm. A good ANN architecture result was obtained with normalization/MSC which was 13-3-1 at 1,000 iterations, which was indicated by $r = 0.9845$, $CV = 4.80\%$ and $\mid RMSEC-RMSEP \mid = 0.0383$.

1. Introduction
Coffee has a different chemical content in each variety [1], one of the chemical contents in coffee is caffeine. The content of caffeine chemical compounds in coffee is in the form of crystals with the main constituent being xanthine purine compounds which are derivatives of protein compounds [2]. The caffeine content in green bean coffee ranges from 0.9-1.5% [3-5]. Generally, the determination of caffeine content in coffee is carried out in a chemical laboratory destructively which takes a long time to obtain results. Therefore, a more effective and efficient method is needed that does not destroy the coffee or non-destructively. 

NIRS is a technology that can be used to determine the chemical content of objects with a non-destructive method by utilizing near infrared light waves which creates a complex spectrum that can be processed qualitatively and quantitatively. The use of NIRS has several advantages such as simple sample preparation, fast detection process, and environmental friendliness because no chemical additives are used. Moreover, NIRS has the potential ability to define several quality parameters simultaneously [3-7], also to characterize and classify a product quality (sortation) based on its chemical content [6, 8].

2. Materials and methods

2.1. Materials and tools
The material used in this study was Mandailing green coffee beans obtained from Mandailing Natal...
Regency. In addition, the tools used in this study were digital scales, petri dishes, grinders, 50 mesh sieves, micro pipettes, measuring cups, ovens, desiccators, FT-NIR spectrometers (NIRFlex N-500, produced by BUCHI Labortechnik AG, Switzerland), computers which was equipped with softwares of unscrambler v 10.3 (CAMO, Norway), Matlab R2017a, unscrambler v.10.5, minitab 19 and Microsoft Excel.

2.2. Research methods acquisition of coffee beans

2.2.1. Destructive determination of caffeine content. Coffee beans are ground using a blender and filtered using a tea filter. The sample in the form of powder was dripped with 50 ml of distilled water and then heated for 30 minutes in a water bath with a temperature of 80-100°C. Then 2 drops of saturated Pb acetate were added to the solution, after the cooling process, then filtered in a 100 ml flask and added distilled water and 5 grams of sodium oxalate. Then 20 ml of the sample was put into HPLC at a temperature of 28-30°C to obtain the caffeine concentration. The caffeine concentration in the sample was calculated using the regression equation obtained and compared with the standard solution. Chemical analysis was carried out to obtain the percentage of chemical content of coffee beans. Measurement of chemical content that was carried out were the measurement of caffeine using the thermogravimetric method, AOAC 2006 and by difference (SNI 01-2892-1992) [9].

Coffee beans are ground using a blender and filtered using a tea filter. The sample in the form of powder was dripped with 50 ml of distilled water and then heated for 30 minutes in a water bath with a temperature of 80-100°C. Then 2 drops of saturated Pb acetate were added to the solution, after the cooling process, then filtered in a 100 ml flask and added distilled water and 5 grams of sodium oxalate. Then 20 ml of the sample was put into HPLC at a temperature of 28-30°C to obtain the caffeine concentration. The caffeine concentration in the sample was calculated using the regression equation obtained and compared with the standard solution. Chemical analysis was carried out to obtain the percentage of chemical content of coffee beans. Measurement of chemical content carried out is the measurement of caffeine using the thermogravimetric method, AOAC 2006 and by difference (SNI 01-2892-1992) [9].

2.2.2. Acquisition of coffee beans. The acquisition of reflectance spectra was carried out on each sample of green coffee beans. A total of 96 grams of coffee beans were placed in petri dishes evenly and tightly with 4 layers of piles [3]. Reflectant measurements are carried out by scanning each sample 3 times. The NIR wavelength range used were 1000-2500 nm with an interval of 2 nm interval.

2.2.3. Research data analysis. The spectral data obtained from measurements were reflectance data, which then transformed into absorbent spectra (log (1 / R)) to obtain a linear correlation between the NIR absorption value and chemical data [3]. The spectral data was still raw data, so a pretreatment methods were needed to reduce noise caused by external influences. In this study, standardize, multiple scatter sorrection (MSC) and normalization methods were used. Normalization serves to increase the reflectance value into the range 0-1 so that the difference will be clearer and eliminate the effect of differences in the particle size of the test sample. The multiple scatter sorrection (MSC) method and the combination of the two pretreatment methods.

2.2.4. Calibration and validation. Data processing was continued with the PCA method to obtain several main components. The main component (PC) was the result of data compression which contains information about the reflectance and absorbance values of NIR to avoid overfitting problems and obtain new variables without losing initial information (Osborne et al., 1993). The PC value was used as the input value in the ANN model. The ANN model developed in this study uses a supervised learning method with (Backpropagation) with the basic training algorithm Gradient Descent with Momentum and Adaptive (traingdx) to build a calibration model. The ANN architecture consists of three layers: input layer, shrouded layer and the output layer. The ANN input layer consisted of 10 and
13 PC variations of the absorbance data of the NIRS spectra. The number of shrouded layer neurons used was a variation of 10 and 13 neurons. The output of ANN is in the form of proximate content of caffeine content with ANN separately (the output is the value of the content of each component).

In this study, simulation of ANN was using the Matlab R2017a application. PC obtained from PCA is used as input data on ANN as calibration data and validation data with a ratio of 2/3 for calibration and 1/3 for validation carried out in Microsoft Excel software. Calibration data is called training data and validation data is called data test. After that, import training data and test data from Microsoft Excel in the matlab r2017a software to create input and target variables on the Matlab r2017a workspace. The next step was to give the command "nntool" in the command window to bring up the neural network tools. Then enter training data, training targets, test data and test targets by importing data from the workspace. To bring out the ANN training is to create a neural network.

Moreover, to train a network that has already been formed a network still needs to be set with several parameters prior to training network training. To get a good prediction result, it is necessary to enter the correct parameter values. The training parameters can be set the desired number of iterations to get the optimal calibration value. Validation values can be obtained by simulating the results of the calibration on the train network.

The regression model performance was evaluated with Root Mean Square Error Prediction (RMSEP). After obtaining the calibration regression equation model, the validation stage was carried out using the rest of the data. Validation aims to test the accuracy of the estimated chemical content of coffee with the calibration regression equation that has been built. The parameters for determining the suitability of the calibration model are the correlation coefficient (r), coefficient of variation (CV). The range of r values ranges from 0 to 1. A small RMSE value indicates a low error in the prediction of data calibration, if the value is closer to 0 then the prediction has good accuracy. CV shows the comparison of RMSE with the average chemical data of materials.

3. Results and discussion
The interactions were seen clearly between the waves and the molecular bonds in the material can be seen in the peaks and valleys formed in the absorbent spectra graph (Figure 1). In this study, the wave peaks indicating the caffeine content were at wavelengths of 1.208.9 nm and 1,728.9 nm [10].

![Figure 1. Average absorbance of NIRS on green beans Arabica Mandailing.](image)

3.1. Pretreatment data
Data changes were clearly visible before pretreatment and after pretreatment. In Figure 2 (a) the pretreatment using MSC spectrum is closer to the mean spectrum value, which is different from the data value before pretreatment (Figure 1). This is because pretreatment with MSC can reduce the spread of light on the diffuse reflectance and transmission spectrum so that the spectrum will dock (coincide) to the data mean value. From Figure 2 (b) with pretreatment normalization, the spectrum value is clearly seen in the difference compared to Figure 1, the highest value in the data in Figure 2
(b) is 1 and the lowest value is 0. The normalized value is the boundary value set when performing the pretreatment process.

3.2. Analysis using PCA

Figure 3 (b) is a graph of the average MSC pretreatment with segments 6, 8, and 10. The PC retrieval process was based on the cumulative percent of the PC by looking at the cumulative percent variation value until the value did not increase anymore. For PC in segment 6 it was taken from PC 1 with a value of 71% to PC 10 with a value of 99.9%. This was based on the cumulative percent value until PC 10 was considered to represent the entire data value because it has stagnated in the cumulative percent value on PC 11. For PCs in segments 8 and 10, the same thing applies to segment 6, namely 10 PCs from PC 1 with 71% to PC 10 with a value of 99.9% for segment 8, while for segment 10 from PC 1 with a value of 70.1% to PC 10 with a value of 99.9%. This must be done carefully because it will have an impact on the ANN input during the ANN simulation training. Then, as the input value of the ANN calibration and validation results from the analysis of the main components, namely PC 1 to PC 10 for each segment, the multiplication process of the matrix is carried out first between PC 1 to PC 10 with the results of the NIR spectra segmentation, namely segments 6, 8 and 10.

The pretreatment process with MSC with segmentation of 6, 8 and 10 could be seen to produce different cumulative percent values compared to normalization. Based on Figure 3 (b) for the cumulative percent value of PC 1 segment 6, namely 60.7% to PC 11 with a value of 99.8%, in segment 8 the cumulative percent value of PC 1 was 60.7% to PC 13 with a value of 99.9% and in segment 10 the cumulative percent value on PC 1 was 60.8% to PC 13 with a value of 99.9%. Treatment in normalization pretreatment was also carried out on MSC to get the calibration input value and validation of ANN.
3.3. Model validation calibration using Artificial Neural Networks (ANN)

Table 1 shows the results of the best predictions at pretreatment Normalization and MSC for each segmentation. In this study, underfitting and overfitting occurred during processing with ANN training which can affect the results of ANN training to be worse such as the same training results even with different inputs, the RMSEP value was higher than the RMSEC. This happened because there was no certainty that determines the number of hidden layers which was one of the causes of underfitting and overfitting [2]. The worst ANN model among the best models for each pretreatment and each segmentation is found using normalization pretreatment segment 8 with architecture 10-7-1 at 1,000 iterations.

Table 1. Results of the best predictions at pretreatment Normalization and MSC for each segmentation

| Architecture | Normalization | MSC |
|--------------|---------------|-----|
|              | Segment 6     | Segment 8 | Segment 10 | Segment 6 | Segment 8 | Segment 10 |
| In           | 10            | 10       | 10          | 11         | 13         | 13         |
| HL           | 3             | 7        | 3           | 3          | 3          | 7          |
| Out          | 1             | 1        | 1           | 1          | 1          | 1          |
| Iteration (Thousand) | 10       | 1       | 1       | 5          | 1          | 5          |
| $R^2$        | 0.833772      | 0.893283 | 0.7121148  | 0.8500283  | 0.9691983  | 0.8441057  |
| $r$          | 0.9131111     | 0.9451364 | 0.843869   | 0.9219698  | 0.9844787  | 0.9187523  |
| RMSEC        | 0.0201122     | 0.0170452 | 0.0277783  | 0.0207092  | 0.0091305  | 0.0204789  |
| RMSEP        | 0.0615624     | 0.085402  | 0.0721246  | 0.0578565  | 0.047465   | 0.0453193  |
| cv (%)       | 6.2330537     | 8.6467542 | 7.3024509  | 5.8578377  | 4.8057167  | 4.5884707  |
| RMSEC-RMSEP  | 0.0414502     | 0.0683568 | 0.0443463  | 0.0371474  | 0.0383344  | 0.0248404  |

**Figure 4.** Plot the best model using pretreatment MSC.

$$y = 0.9756x + 0.0233$$

$R^2 = 0.9692$
From the results of research shows Table 1 for the calibration and validation of pretreatment ANN using MSC with segments 6, 8, and 10, the best results can be seen in the plot of Figure 4, namely with an $R^2$ value of 0.9692, an $r$ value of 0.9845 for the calibration and for the validation value with percent cv amounting to 4.80% with an $r$ value of 0.6097. Selection of the best calibration and validation research results is not only seen from the value of r, R2 and percent CV alone, the difference between the RMSEC and RMSEP values was included in influencing the selection of the best results. In Table 1, the difference between RMSEC and RMSEP values was close to 0, namely 0.0383. The best results were obtained with input network architecture 13, output 1, hidden layer 3 nodes and 1,000 iterations in segment 8.

4. Conclusions
From this research, the prediction of Mandailing Natal caffeine content resulted in the best calibration using NIRS with the ANN method in the MSC pretreatment, namely with the input 13 architecture, output 1, hidden layer 3 node iterations 1,000 in the segmentation reduction process 8. This architecture has a value of $r = 0.9845$, $R^2 = 0.9692$, CV = 4.80% and the difference between the RMSEC and RMSEP values was 0.0383.

References
[1] Sahfitri, Rindang A and Ayu P C 2020 Prediction of Water Content in Lintong Green Bean Coffee Using FT-NIRS and PLS Method IOP Conf. Ser.: Earth and Environ. Sci. 454 012047
[2] Burnham T A 2001 Drug Fact and Comparison (St Louis: A Wolters Kluwers Company)
[3] Ayu P C 2017 Pengembangan Model Penentuan Kandungan Kimia Utama Pembentuk Flavor Biji Kopi Java Preanger Menggunakan FT-NIRS [Development of Determination Model of The Main Flavour Attributes in Java Preanger Green Coffee Bean Using FT-NIRS] Thesis (Bogor: IPB University)
[4] Ayu P C, Budiastra I W and Rindang A 2020 NIR Spectroscopy Application for Determination Caffeine Content of Arabica Green Bean Coffee IOP Conf. Ser.: Earth and Environ. Sci. 454 012049
[5] Budiastra I W, Sutrisno, Widytomoto S and Ayu P C 2018 Prediction of Caffeine Content in Java Preanger Coffee Beans by NIR Spectroscopy Using PLS and MLR Method IOP Conf. Ser.: Earth and Environ. Sci. 147 012004
[6] Kurniawan F 2017 Characterization and classification of Java Arabica coffee beans based on geographical indications using NIR spectroscopy and discriminant analysis methods Thesis (Bogor: IPB University)
[7] Budiastra I W, Sutrisno, Widytomoto S and Ayu P C 2020 Determination of Trigonelline and Chlorogenic Acid (CGA) Concentration in Intact Coffee Beans by NIRS Spectroscopy AgricEngInt: CIGR Journal 22 (1) pp 162-8
[8] Siregar S D, Rindang A and Ayu P C 2020 Principle Component Analysis (PCA) – Classification of Arabica Green Bean Coffee of North Sumatera Using FT-NIRS IOP Conf. Ser.: Earth and Environ. Sci. 454 012046
[9] [AOAC] Association of Official Analytical Chemists. (2016). Official Method 948.22: Fat (crude) in nuts and nut products and Official Method 950.48: Protein (crude) in nuts and nut products. Official Methods of Analysis of AOAC International, 20th ed., AOAC International, Gaithersburg, MD, USA
[10] Ribeiro J S, Ferreira M M C and Salva T J G 2011 Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverage using near infrared spectroscopy Talanta 83 pp 1352-8

Acknowledgements
The research was fully supported by Universitas Sumatera Utara based on contract of TALENTA USU research implementation for fiscal year 2019 No: 4167/UN5.1.R/PPM/2019 date: 01 April 2019.