PREDICTIVE MODELS IN NEAR-INFRARED SPECTROSCOPIC ANALYSIS FOR BLOOD HEMOGLOBIN PREDICTION

MOHD NAZRUL EFFENDY BIN MOHD IDRUS

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Master of Electrical Engineering

Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia

NOVEMBER 2018

DEDICATION

SPECIAL GRATITUDES TO:

Special for my parent,

SARIAH BINTI JAFFAR & MOHD IDRUS BIN ZAKARIA

Especially for my kindly supervisor who gives encouragement and support,

DR. CHIA KIM SENG

My lovely wife,

JURIAH BINTI ABDUL WARIF

My lovely sibling,

All my friends who give full support,

People who guide and help me,

Only Allah S.W.T can repay your kindness

ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to ALLAH Almighty for His blessing to this great achievement. Without His grace and comparison, none of this would have been possible.

I would like to express my deepest appreciation to my supervisor, Dr. Chia Kim Seng for the guidance, enthusiasm, and motivation given for the entire progress of this project.

Special thanks to my family members and friends who have been helpful by offering comment and advice ensuring the success of this research.

Kindly regards go to my educational sponsor bodies, Hadiah Latihan Persekutuan Kementerian Pengajian Tinggi Malaysia. I would like to thank the Faculty of Electrical and Electronic Engineering (FKEE), UTHM for providing facilities for this study.

Lastly, I would like to thank those who have contributed directly or indirectly toward the success of this research.

ABSTRACT

Determination of blood hemoglobin concentration is important to diagnose anaemia. Common clinical practice used to measure blood hemoglobin is by drawn some blood from the patient to be mixed with reagent chemicals for analysis. Alternatively, near-infrared spectroscopy (NIRS) technology can be used to measure blood hemoglobin level. NIRS is based on molecular overtone and combination vibrations produce absorption bands typically very broad, leading to complex spectral data makes the NIRS useful for analysis. Different types of predictive models such as linear, nonlinear, and hybrid predictive models were commonly used to predict component of interest from NIRS spectral data. However, different predictive model approached may achieve different accuracy of performance in predicting component of interest from NIRS spectral data. The aims of this study is to investigate the accuracy of linear partial least square (PLS), nonlinear artificial neural network (ANN), and hybrid partial least square - artificial neural network (PLS-ANN) predictive modelling in NIRS analysis. These predictive models were coupled with Savitzky-Golay (SG) preprocessing to remove unwanted signal from spectral data. The optimal numbers of frame length, latent variables, and hidden neurons used in SG preprocessing and predictive models were investigated. Results show ANN coupled with first order SG derivatives achieved the best prediction of performance with root mean square error of prediction (RMSEP) and the coefficient of determination of prediction (R_p^2) were 0.3517 gd/L and 0.9849, respectively, followed by PLS-ANN and PLS. Although ANN outperformed PLS-ANN, PLS-ANN is simpler than ANN with three hidden neurons compared with five hidden neurons.

ABSTRAK

Penentuan kepekatan hemoglobin darah adalah penting untuk mendiagnos anemia. Amalan klinikal yang biasa digunakan untuk mengukur hemoglobin darah adalah dengan mengambil beberapa darah dari pesakit untuk dicampur dengan bahan kimia reagen untuk dianalisis. Sebagai alternatif, teknologi *near-infrared spectroscopy* (NIRS) boleh digunakan untuk mengukur aras hemoglobin darah. NIRS berasal daripada molekul yang berlebihan tona dan gabungan getaran yang menghasilkan jalur penyerapan yang sangat besar untuk menghasilkan data spektrum yang kompleks menjadikan NIRS berguna untuk dianalisis. Model ramalan yang berbeza seperti model ramalan linear, tidak linear, dan hibrid lazimnya digunakan untuk meramalkan komponen dari data spektrum NIRS. Walaubagaimanapun, model ramalan yang berbeza boleh mencapai ketepatan prestasi yang berbeza dalam meramalkan komponen penting daripada data spektrum NIRS. Matlamat kajian ini adalah untuk mengkaji ketepatan ramalan model linear *partial least square* (PLS), bukan linear *artificial neural network* (ANN), dan hibrid *partial least square-artificial neural network* (PLS-ANN) dalam analisis NIRS. Model ramalan ini digandingkan dengan *Savitzky-Golay* (SG) untuk mengeluarkan isyarat yang tidak diingini daripada data spektrum. Selanjutnya, bilangan panjang bingkai, pemboleh ubah terpendam, dan neuron yang digunakan di prapemproses SG dan model ramalan dikaji. Dalam kajian ini, ANN digandingkan dengan prapemproses SG derivatif pertama memperoleh ramalan prestasi terbaik dengan 0.3517 gd/L *root mean square error of prediction* (RMSEP) dan 0.9849 *coefficient of determination* (R_p^2) diikuti oleh PLS-ANN dan PLS. Walaupun ANN lebih baik daripada PLS-ANN, senibina PLS-ANN adalah lebih ringkas daripada ANN dengan tiga neuron digunakan berbanding dengan lima neuron.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

nm	-	Nanometer
%	-	Percentage
SG	-	Savitzky-Golay
ANN	-	Artificial Neural Network
PLS	-	Partial Least Square
PLS-ANN	-	Partial Least Square - Artificial Neural Network
HB	-	Hemoglobin
ms	-	Millisecond
NIR	-	Near-infrared
NIRS	-	Near-infrared Spectroscopy
RMSECV	-	Root Mean Square of Cross-Validation
RMSEC	-	Root Mean Square of Calibration
RMSEP	-	Root Mean Square of Prediction
SCG	-	Scale Conjugate Gradient
LM	-	Levenberg-Marquardt
UTHM	-	Universiti Tun Hussein Onn Malaysia
FKEE	-	Fakulti Kejuruteraan Elektrik dan Elektronik

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

This chapter presents the background of the research, the problem that needs to be solved, the aim and objectives, and the scopes of the research.

1.1 Background of study

Anemia disease can be diagnosed from blood hemoglobin concentration [1]. Generally, the hemoglobin level that lower than 7.0 g/dL was severe anemia; that between 9.9 to 7.0 g/dL was moderate anemia; that between 10 to 11.9 g/dL was mild anemia; and that was 12.0 g/dL or higher was defined as non-anemia [2]. Cyanmethemoglobin method is a common clinical practice used to measure blood hemoglobin by drawn some blood from the patient to be mixed with reagent chemicals for analysis [3]. However, this conventional method is time consuming and invasive. Besides, the conventional method requires reagent chemical for the analysis.

Near-infrared spectroscopy (NIRS) measures overtones and combination tones of the fundamental molecular vibrations in especially the asymmetric vibrations of C–H, O–H, and N–H produces absorption bands make near-infrared (NIR) useful for analyzing in the biological system [4, 5]. NIRS is a simple and reliable technique widely used in various field including blood constituent [6–14], food [15–19], functional neuroimaging [12], agrochemical [15, 20, 21], and fuel [22–24]. The advanced NIRS sensing technology produces a large volume of spectral data make it have advantages in the analytical analysis [25]. Multivariate calibration modelling were used to establish a relationship between spectral data and component

of interest [9]. Preprocessing, calibration, and validations is process relies in developing multivariate calibration modelling [26].

A well-known Savitzky-Golay (SG) preprocessing method has rapidly used as preprocessing technique in NIRS analysis. Principle component regression coupled with second order SG derivative achieved the best accuracy in predicting quality assessment of soluble solids content [27]. In addition, SG preprocessing has successfully unwanted signal in raw spectral data in NIRS analysis [6, 11, 17, 21, 27].

Different types of multivariate calibration such as linear, nonlinear, and hybrid predictive model methods have been applied into NIRS spectral data to extract the relevant part of information for a large dataset to predict concentration from samples [22, 28]. Partial least squares (PLS) was the most commonly linear predictive model used due to its advantages of rapidity, simplicity, and practicability [29, 30]. With a linear method combination of principal multi linear regression (MLR) and principle component analysis (PCA), PLS is able to handle data with strong co-linearity and noise, as well as in situations with the number of variables more than the number of samples [29]. A nonlinear predictive model approach, artificial neural networks (ANN) was widely used in drawing out quantitative information from large spectroscopic non-linearity databases [7, 15, 16, 22]. ANN originally inspired by neurons in the brain can be viewed as a universal model-free approximator that can represent any non-linear function with sufficient accuracy. For this reason, ANN is often taken to solve problems with non-linear relationships. Partial least square - artificial neural network (PLS-ANN) is a hybrid predictive model approach was proved can increase the accuracy of prediction compared to PLS and ANN [31, 32]. The major concern for these multivariate calibration methods is to generalized the nonlinearity in NIR spectral data [28, 29].

Chemometrics shootout competition was carried out in the international diffuse reflectance conference (IDRC). The competition aims to challenge the ability of participates to develop the most accurate model for blood hemoglobin prediction in NIRS analysis.

1.2 Problem statement

Chemometrics shootout competition was carried out in the international diffuse reflectance conference (IDRC2010) aimed to challenge the ability of participates to develop a better calibration model for blood hemoglobin prediction based on given spectral data [33]. Although four selected models achieved good results, only linear models i.e. multiple linear regression (MLR) and partial least square (PLS) that coupled with wavelength selection or reduction approaches were investigated [34]. However, linear model may give non-optimal prediction results when the spectra exhibit nonlinearities [35]. Moreover, nonlinear and hybrid predictive models have shown a better performance compared to the linear predictive model [22], [31]. There is a need to investigate the performance of different types of predictive models i.e. linear, nonlinear, and hybrid predictive models to improve the accuracy of prediction for near-infrared spectroscopic analysis in blood hemoglobin prediction.

Savitzky-Golay preprocessing method was commonly used to remove the unwanted signal from spectral data and overcome the most common issues in raw spectral data from near-infrared [5, 10, 14, 15]. However, little studies have been conducted to investigate the effect of the different number of frame length affects the predictive accuracy of the modelling [27]. Therefore, an investigation of the optimum parameters in SG preprocessing and capability of the different predictive models such as PLS, ANN, and PLS-ANN is needed to predict blood hemoglobin level from NIR spectral data.

1.3 Research aim

To investigate the accuracy of the different predictive models with different Savitzky-Golay preprocessing in the near-infrared spectroscopic analysis.

1.4 Research objectives

The objectives of this study are as follows:

1) To develop different predictive models with different Savitzky-Golay preprocessing for near-infrared spectroscopic analysis in blood hemoglobin prediction.
2) To establish relationship between blood hemoglobin and near-infrared spectrum using different predictive models and Savitzky-Golay preprocessing.
3) To evaluate the predictive accuracy of the developed predictive models based on root mean square error of prediction (RMSEP) and coefficient of determination (R_p^2) of prediction using hold-out validation and cross-validation.

1.5 Research scopes

This research addresses several limitations based on study focusing area.

1) The spectrum datasets were adopted from IDRC ShootOut 2010 (blood hemoglobin) were available on the website http://www.idrc-chambersburg.org.
2) Smoothing, first order, and second order Savitzky-Golay preprocessing method were used for the different predictive models.
3) Three different predictive models that are artificial neural network (ANN), partial least square (PLS) and partial least square - artificial neural network (PLS-ANN) developed.
4) Root mean square error of calibration (RMSEC), Root mean square error of Predictions (RMSEP) and coefficient of determination prediction (R_p^2) used to evaluate the performance of the modeling.

CHAPTER 2

LITERATURE REVIEW

This chapter reviewed the related information about blood hemoglobin level in Section 2.1. The adopted near-infrared spectra data is described in Section 2.2 and Section 2.3. Then, a different type of preprocessing method is reviewed in Section 2.5. Next, a different type of predictive modelling is reviewed in Section 2.6. After that, the validation method used in this research is discussed in Section 2.7. Lastly, the chapter summary is summarized in Section 2.8.

2.1 Blood hemoglobin

Hemoglobin (Hb) is a protein molecule in red blood cells that contains an iron molecule to carries oxygen from the lungs to the rest of body [1]. Hemoglobin concentration can be used to diagnose anaemia, even though not all anaemia is caused by iron deficiency [2]. Cut-offs level of hemoglobin was used by the World Health Organization (WHO) in defining mild, moderate and severe anaemia as shown in Table 2.1.

Table 2.1: World Health Organization cut-offs defining mild, moderate and severe anaemia [2]

Population	Non -Anaemia[a] (g/dL)	Anaemia[a] (g/dL)		
		Mild	Moderate	Severe
Children 6 - 59 months of age	11.0 or higher	10.0-10.9	0.70-0.99	lower than 0.70
Children 5 - 11 years of age	11.5 or higher	11.0-11.4	0.80-10.9	lower than 0.80
Children 12 - 14 years of age	12.0 or higher	11.0-11.9	0.80-10.9	lower than 0.80
Non-pregnant women (15 years of age and above)	12.0 or higher	110-11.9	0.80-10.9	lower than 0.80
Pregnant women	11.0 or higher	10.0-10.9	0.70-0.99	lower than 0.70
Men (15 years of age and above)	13.0 or higher	11.0-12.9	0.80-10.9	lower than 0.80

[a] Haemoglobin levels to diagnose anaemia at sea level.

2.2 Near-infrared spectrum

Infrared spectroscopy is the study of the interaction between matter and electromagnetic radiation that deals with the infrared region of the electromagnetic spectrum. There are three regions infrared spectrum; the Near-infrared (NIR), middle infrared, and far infrared. Near-infrared spectroscopy (NIRS) is a spectroscopic method that uses the near-infrared region of the electromagnetic spectrum (from about 780 nm to 2500 nm) [26]. The NIRS measures overtones and combination tones of the fundamental molecular vibrations in especially the asymmetric vibrations of C–H, O–H, and N–H produces absorption bands make NIR useful for analyzing in the biological system [4, 5].

The NIRS is a simple and reliable technique widely used in medicine including blood hemoglobin, pulse oximetry, functional neuroimaging, sports medicine, elite sports training, ergonomics, and rehabilitation. There are also applications in other areas as well such as pharmaceutical, food and agrochemical quality control, and fuel research. Different reviews have demonstrated that fundamental reasons which limit the use of NIRS based on several factors; interference resulting in a poor signal to noise ratio, calibration issues, baseline drift, thermal noise and proper selection of wavelength [15]. The challenge is to model complex and high-dimensional spectral data. Multiple variables calibration techniques, for example, principal components regression (PCA), partial least squares (PLS) and artificial neural networks (ANN) are often employed to extract the desired chemical information [29]. The nonlinearity of spectral data lead to multivariable calibration method are increasingly used to extract relevant information from the different type of spectral data [22].

2.2.1 Nonlinearity in near infra-red spectral data

The nonlinearity in spectral data can influence the accuracy performance of the modelling. The main causes of the nonlinearity in spectral data is related to light scattering effect [36]. Molecular interactions combine light scattering can produce a deviation from linear additive relation to nonlinear relation between the analyte concentration level and the acquired absorbance signals [9]. In addition, fluctuations

in temperature, compactness, varying temperature, and other external variables would make the nonlinear variations into spectral data [36, 37]. The baseline shift, slope, and background effects can be observed as the major undesired variations in the sample set.

2.3 Blood hemoglobin Spectrum data

The spectrum dataset adopted from IDRC ShootOut 2010 provided by Karl Norris [33]. Blood samples were analyzed during the period from 1990 to 1992 with an NIRSystems 6500 spectrometer with a transmission amplifier mounted in the sample transport. All spectra have 700 variables, from 1100 to 2498 nm, with a 2 nm interval as shown in Figure 2.1. The instrument was configured to a vertical light-path mode using a platform supplied by the manufacturer.

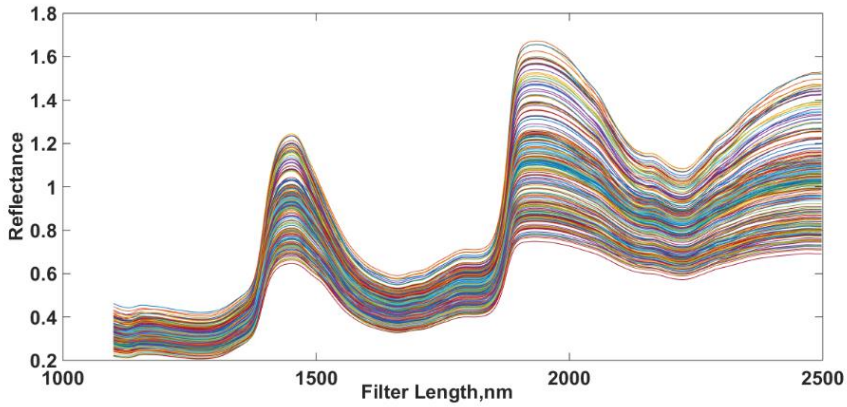

Figure 2.1: Blood hemoglobin reflectance raw spectral data

The sample cell was a 2 cm diameter stainless steel cylinder with a quartz window. For reflection measurements, the cell was filled to a sample thickness of at least 2 mm to provide a sample of infinite thickness for the NIR spectral region. The empty cell was used as a reference for transmission measurements, and a ceramic reference standard was used for reflection measurements. Table 2.2 shows the summary of descriptive characteristics of the calibration and testing references data. From 425 sets of samples, 231 samples for calibration and 194 samples for testing which means 54.35% and 45.65% for calibration and testing respectively ratio were used. From Table 2.2, the minimum value in testing set of data is lower than calibration set of data that is 6.5 g/dL and 10.30 g/dL respectively. Then, the

maximum value in testing higher than calibration set of data that is 18.20 g/dL and 17.30 g/dL respectively. These characteristics of calibration and testing data indicate that extrapolation samples were used for modelling. Therefore, an appropriate modelling technique needs to be implemented to predict the out of range data from the testing process. IDRC shootout 2010 has used the same extrapolation data in the competition [34].

Table 2.2: Descriptive statistics of the blood hemoglobin

Data	Sample	Min (g/dL)	Max (g/dL)	Mean (g/dL)	Median (g/dL)	Std
Calibration	231	10.30	17.30	13.78	13.70	1.66
Testing	194	6.50	18.20	12.20	12.25	2.83
Total	425	6.50	18.20	13.06	13.40	2.40

The blood hemoglobin reference that is 231 sets for calibration and 194 sets for blind testing were measured by Coulter STKS monitor, which is made by the Coulter Corporation of Hialeah, FL [39]. Previous research was used a similar technique in collecting a blood sample for calibration and reference sample showed near-infrared spectroscopy spectral data suggested the possibility of noninvasive measuring blood hemoglobin concentration [39]. While in another research shown three wavelengths (586, 816, and 1300 nm) were identified represented of possibility measurement blood hemoglobin by prototypic noninvasive hemoglobinometer. Based on these techniques to acquired blood sample spectra data, it indicated that blood hemoglobin concentration can be measured as a noninvasive measurement.

2.4 Chemometrics shootout

Chemometrics shootout competition was carried out in the international diffuse reflectance conference (IDRC) [34]. The competition aims to challenge the ability of participates to develop the most accurate model for blood hemoglobin prediction in NIRS analysis. Table 2.3 shows the root mean square error of prediction for reflectance transmission geometry. The results indicated that participant 1 was declared as the champions of the competitions followed by participant 4, participant 2, and participant 3. From the approach taken by the participants, only linear models, e.g. multiple linear regression (MLR) and partial least square (PLS) were used. Whereas there is other predictive modelling such as

artificial neural network and deep learning can use to overcome the nonlinearity issues in NIR analysis. In addition, the data preprocessing technique with wavelength selection or reduction approaches could be unappropriated to be used and consequently minimize the information of spectral data for modelling. Whereas, there is another common preprocessing technique used in NIR analysis such as Savitzky-Golay preprocessing.

Table 2.3: Statistics root mean square error of prediction for reflectance transmission geometry

Prediction statistic	Participant 1	Participant 2	Participant 3	Participant 4
Root mean square error of prediction (g/dL)	0.361	0.431	0.470	0.422
Predictive model method	Multi Linear Regression	Partial Least Square	Multi Linear Regression	Partial Least Square

2.5 NIR preprocessing method

In NIR spectroscopy signal analysis, preprocessing methods were always a concerning topic discussed. Signal to noise ratio, baseline effect, and slope effect are a most common issue in raw spectral data from NIR spectroscopy. Different kind of preprocessing method approach can reduce and remove the unwanted signal from spectral data to improve accuracy and the performance of prediction results of samples. There are several types of preprocessing methods such as multiplicative scatter correction (MSC), standard normal variate (SNV), direct orthogonal signal correction (DOSC), wavelet transforms, robust principal component analysis (ROBPCA) and Savitzky-Golay derivatives (SG) with different advantages were applied in previous research.

MSC is a preprocessing step needed to compensate for additive e.g. amplification and multiplicative e.g. offset in spectral data in order to avoid these effects dominate the information in the spectral data [36, 37]. Partial least square coupled with MSC preprocessing shown higher accuracy in the effectiveness identification of clinical bacterial pathogens using near-infrared spectroscopy compared with other preprocessing technique e.g. standard normal variate correction, first and second order Savitzky-Golay derivative [42]. However, MSC has shown the lowest performance compared with correction maximum reflectance preprocessing method [43]. SNV is a row-oriented transformation which centers and

REFERENCES

[1] F. B. Jensen, A. Fago, and R. E. Weber, "Hemoglobin Structure and Function," *Fish Physiol.*, vol. 17, no. March 2016, pp. 1–40, 1998.

[2] M. Chan, "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity," *Geneva, Switz. World Heal. Organ.*, pp. 1–6, 2011.

[3] J. Rosenblit *et al.*, "Evaluation of three methods for hemoglobin measurement in a blood donor setting.," *Sao Paulo Med. J.*, vol. 117, no. 3, pp. 108–112, 1999.

[4] L. Norgaard, B. Rasmus, and S. B. Engelsen, "Principal Component Analysis and Near Infrared Spectroscopy," *A white Pap. from FOSS*, pp. 2–7, 2014.

[5] G. Reich, "Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications," *Adv. Drug Deliv. Rev.*, vol. 57, no. 8, pp. 1109–1143, 2005.

[6] N. N. M. Yatim *et al.*, "Noninvasive glucose level determination using diffuse reflectance near infrared spectroscopy and chemometrics analysis based on in vitro sample and human skin," *Syst. Process Control (ICSPC), 2014 IEEE Conf.*, no. March 2016, pp. 30–35, 2014.

[7] S. Ramasahayam, K. S. Haindavi, B. Kavala, and S. R. Chowdhury, "Non invasive estimation of blood glucose using near infra red spectroscopy and double regression analysis," *Proc. Int. Conf. Sens. Technol. ICST*, pp. 627–631, 2013.

[8] P. Venkatesan, C. Dharuman, and Gunasekaran, "squares s regression with A comparative study of principal component regression and partial least square application to FTIR diabetes data," *Indian J. Sci. Technol.*, vol. 4, no. 7, pp. 740–746, 2011.

[9] M. Goodarzi, S. Sharma, H. Ramon, and W. Saeys, "Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring," *TrAC - Trends Anal. Chem.*, vol. 67, pp. 147–158, 2015.

[10] A. Al-Mbaideen and M. Benaissa, "Determination of glucose concentration from NIR spectra using independent component regression," *Chemom. Intell. Lab. Syst.*, vol. 105, no. 1, pp. 131–135, 2011.

[11] C. Wood *et al.*, "Near infra red spectroscopy as a multivariate process analytical tool for predicting pharmaceutical co-crystal concentration," *J. Pharm. Biomed. Anal.*, vol. 129, pp. 172–181, 2016.

[12] J. Hennrich, C. Herff, D. Heger, and T. Schultz, "Investigating deep learning for fNIRS based BCI," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2015–Novem, pp. 2844–2847, 2015.

[13] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharm. Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, 2000.

[14] M. R. Bozkurt, N. Yurtay, Z. Yilmaz, and C. Sertkaya, "Comparison of difierent methods for determining diabetes," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 22, no. 4, pp. 1044–1055, 2014.

[15] M. W. Lars Nørgaard, Martin Lagerholm, "Artificial Neural Networks and Near Infrared Spectroscopy - A case study on protein content in whole wheat grain," *A white Pap. from FOSS*, no. 1, pp. 1–6, 2013.

[16] K. S. Chia, H. A. Rahim, and R. A. Rahim, "Artificial Neural Network Coupled with Robust Principal Components in Near Infrared Spectroscopic Analysis," *IEEE 8th Int. Colloq. Signal Process. its Appl. Artif.*, no. 36, pp. 19–22, 2012.

[17] N. C. T. Mariani, G. H. De Almeida Teixeira, K. M. G. De Lima, T. B. Morgenstern, V. Nardini, and L. C. C. Júnior, "NIRS and iSPA-PLS for predicting total anthocyanin content in jaboticaba fruit," *Food Chem.*, vol. 174, pp. 643–648, 2015.

[18] M. Singh, S. Semwal, A. Kumar, and S. Singh, “Impedance modeling for classification of flavored green teas,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 23, pp. 2208–2214, 2015.

[19] A. SARIKAŞ and M. DOĞRUYOL BAŞAR, “An electronic portable device design to spectroscopically assess fruit quality,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, pp. 4063–4076, 2017.

[20] P. Lin, Y. M. Chen, Y. He, G. W. Hu, X. L. Fu, and C. L. Gu, “Study on Nonlinear Multivariate Methods Combined with the Visible Near-Infrared Spectroscopy (Vis/NIRS) Technique for Detecting the Protein Content of Cheese,” *Food Bioprocess Technol.*, vol. 7, no. 12, pp. 3359–3369, 2014.

[21] T. R. Viegas, A. L. M. L. Mata, M. M. L. Duarte, and K. M. G. Lima, “Determination of quality attributes in wax jambu fruit using NIRS and PLS,” *Food Chem.*, vol. 190, pp. 1–4, 2016.

[22] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, “Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction,” *Chemom. Intell. Lab. Syst.*, vol. 88, no. 2, pp. 183–188, 2007.

[23] G. Kökkülünk, E. Akdoğan, and V. Ayhan, “Prediction of emissions and exhaust temperature for direct injection diesel engine with emulsified fuel using ANN,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 21, no. SUPPL. 2, pp. 2141–2152, 2013.

[24] K. S. Chia, “Predicting the Boiling Point of Diesel Fuel using Adaptive Linear Neuron and Near Infrared Spectrum,” in *Control Conference (ASCC)*, 2015, pp. 0–2.

[25] R. M. Balabin, E. I. Lomakina, and R. Z. Safieva, “Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy,” *Fuel*, vol. 90, no. 5, pp. 2007–2015, 2011.

[26] M. Blanco and I. Villarroya, “NIR spectroscopy: A rapid-response analytical tool,” *TrAC - Trends Anal. Chem.*, vol. 21, no. 4, pp. 240–250, 2002.

[27] K. S. Chia, H. Abdul Rahim, and R. Abdul Rahim, "Evaluation of common pre-processing approaches for visible (VIS) and shortwave near infrared (SWNIR) spectroscopy in soluble solids content (SSC) assessment," *Biosyst. Eng.*, vol. 115, no. 1, pp. 82–88, 2013.

[28] H. Yang, P. R. Griffiths, and J. D. Tate, "Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra," *Anal. Chim. Acta*, vol. 489, no. 2, pp. 125–136, 2003.

[29] X. Shao, X. Bian, J. Liu, M. Zhang, and W. Cai, "Multivariate calibration methods in near infrared spectroscopic analysis," *Anal. Methods*, vol. 2, no. 11, p. 1662, 2010.

[30] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[31] S. Li, L. Li, R. Milliken, and K. Song, "Hybridization of partial least squares and neural network models for quantifying lunar surface minerals," *Icarus*, vol. 221, no. 1, pp. 208–225, 2012.

[32] K. Song *et al.*, "Using partial least squares-artificial neural network for inversion of inland water chlorophyll-a," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1502–1517, 2014.

[33] B. I. Fred McClure, "Chemometrics ShootOut Rules," *Int. Diffus. Reflectance Conf. 2012*, 2012.

[34] B. Igne *et al.*, "The 2010 IDRC software shoot-out at a glance," *NIR news*, vol. 21, no. 8, p. 14, 2010.

[35] W. Ni, L. Nørgaard, and M. Mørup, "Non-linear calibration models for near infrared spectroscopy," *Anal. Chim. Acta*, vol. 813, pp. 1–14, 2014.

[36] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009.

[37] J. Wan, Y. C. Chen, A. J. Morris, and S. N. Thennadil, "A Comparative

Investigation of the Combined Effects of Pre-Processing, Wavelength Selection, and Regression Methods on Near-Infrared Calibration Model Performance," *Appl. Spectrosc.*, vol. 71, no. 7, pp. 1432–1446, 2017.

[38] Z.-P. Chen and J. Morris, "Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization," *Analyst*, vol. 133, no. 7, p. 914, 2008.

[39] J. T. Kuenstner, K. H. Norris, and W. F. McCarthy, "Measurement of hemoglobin in unlysed blood by near-infrared spectroscopy," *Appl. Spectrosc.*, vol. 48, no. 4, pp. 484–488, 1994.

[40] M. R. Maleki, A. M. Mouazen, H. Ramon, and J. De Baerdemaeker, "Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy," *Biosyst. Eng.*, vol. 96, pp. 427–433, 2007.

[41] H. Yang, B. Kuang, and A. M. Mouazen, "Affect of Different Preprocessing Methods on Principal Component Analysis for Soil Classification," in *Third International Conference on Measuring Technology and Mechatronics Automation*, 2011, pp. 355–358.

[42] W. Lin, Q. Chai, X. Chen, and Y. Li, "Effective Identification of Clinical bacterial pathogens by Fourier Transform Near-Infrared Spectroscopy," in *Chinese Automation Congress (CAC)*, 2017, no. 2016.

[43] R. Vasat, R. Kodesova, A. Klement, and L. Boruvka, "Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation," *Geoderma*, vol. 298, pp. 46–53, 2017.

[44] E. Bertran, H. Iturriaga, S. Maspoch, and I. Montoliu, "Effect of orthogonal signal correction on the determination of compounds with very similar near infrared spectra," *Anal. Chim. Acta*, vol. 431, no. 2, pp. 303–311, 2001.

[45] X. Li and C. Li, "Study on the Application of Wavelet Transform to Non-invasive Glucose Concentration Measurement by NIRS," *2015 Fifth Int. Conf. Instrum. Meas. Comput. Commun. Control*, pp. 1294–1297, 2015.

[46] M. Hubert, P. Rousseeuw, and K. Branden, "ROBPCA: a new approach to

robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 1–34, 2005.

[47] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[48] R. W. Schafer, "What is a savitzky-golay filter?," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, 2011.

[49] G. Betta, D. Capriglione, G. Cerro, L. Ferrigno, and G. Miele, "The effectiveness of Savitzky-Golay smoothing method for spectrum sensing in cognitive radios," *Proc. 2015 18th AISEM Annu. Conf. AISEM 2015*, no. 4, pp. 1–4, 2015.

[50] M. Nazrul, E. Mohd, and K. S. Chia, "Partial Least Square with Savitzky Golay Derivative in Predicting Blood Hemoglobin Using Near Infrared Spectrum," *MATEC Web Conf.*, vol. 01001, pp. 1–6, 2018.

[51] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009.

[52] J. Subramanian and R. Simon, "Over fi tting in prediction models – Is it a problem only in high dimensions ?," *Contemp. Clin. Trials*, vol. 36, no. 2, pp. 636–641, 2013.

[53] K. Murakoshi, "Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps," *BioSystems*, vol. 80, pp. 37–40, 2005.

[54] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. Mcshane, "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification FOR AND," *J. Natl. Cancer Inst.*, vol. 95, no. 1, 2003.

[55] R. Bro, "Multivariate calibration: What is in chemometrics for the analytical chemist?," *Anal. Chim. Acta*, vol. 500, no. 1–2, pp. 185–194, 2003.

[56] T. Naes, T. Isakson, T. Fearn, T. Davies, and N. I. R. Publications, "A USER-

FRIENDLY GUIDE TO MULTIVARIATE CALIBRATIONANDCLAS-SIFICATION," *J. Chemom.*, vol. 17, pp. 571–572, 2003.

[57] J. Xie, T. Pan, J. Chen, H. Chen, G. Jiang, and L. Wu, "Wavenumbers Combination Optimization for NIRS Analysis of Glucose in Human Serum," *Symp. Photonics Optoelectron.*, no. v, pp. 1–4, 2010.

[58] MathWorks, "Partial Least Squares Regression and Principal Components Regression," *R2014a Documentation*. 2014.

[59] R. D. Tobias, "An introduction to partial least squares regression," *SAS Conf. Proc. SAS Users Gr. Int. 20 (SUGI 20)*, pp. 2–5, 1995.

[60] R. Rosipal, "Nonlinear partial least squares: An overview," *Chemoinformatics Adv. Mach. Learn. Perspect. Complex Comput. Methods Collab. Tech.*, pp. 169–189, 2011.

[61] C. Bai, "Noninvasive near infrared spectroscopy on living tissue with multivariate calibration approaches," *PhD (Doctor Philos. thesis, Univ. Iowa*, pp. 1–189, 2010.

[62] Z.-M. Chuah, R. Paramesran, K. Thambiratnam, and S.-C. Poh, "A two-level partial least squares system for non-invasive blood glucose concentration prediction," *Chemom. Intell. Lab. Syst.*, vol. 104, no. 2, pp. 347–351, 2010.

[63] J. Zupan, "Introduction to artificial neural network (ANN) methods: what they are and how to use them," *Acta Chim. Slov.*, vol. 41, no. September, pp. 327–352, 1994.

[64] M. Mia and N. R. Dhar, "Prediction of surface roughness in hard turning under high pressure coolant using Artificial Neural Network," *Measurement*, vol. 92, pp. 464–474, 2016.

[65] P. Yu, M. Y. Low, and W. Zhou, "Development of a partial least squares-artificial neural network (PLS-ANN) hybrid model for the prediction of consumer liking scores of ready-to-drink green tea beverages," *Food Res. Int.*, vol. 103, no. October 2017, pp. 68–75, 2018.

[66] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for

Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[67] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014.

[68] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimed. Tools Appl.*, pp. 1–17, 2017.

[69] R. Xie, A. Quitadamo, and J. Cheng, "A Predictive Model of Gene Expression Using a Deep Learning Framework," pp. 676–681, 2016.

[70] K. Diaz-Chito, K. Georgouli, A. Koidis, and J. Martinez del Rincon, "Incremental model learning for spectroscopy-based food analysis," *Chemom. Intell. Lab. Syst.*, vol. 167, no. December 2016, pp. 123–131, 2017.

[71] M. Khashei, A. Zeinal Hamadani, and M. Bijari, "A novel hybrid classification model of artificial neural networks and multiple linear regression models," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2606–2620, 2012.

[72] H. Ding, Q. Lu, H. Gao, and Z. Peng, "Non-invasive prediction of hemoglobin levels by principal component and back propagation artificial neural network.," *Biomed. Opt. Express*, vol. 5, no. 4, pp. 1145–52, 2014.

[73] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du, "Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling," *Ecol. Inform.*, vol. 10, pp. 25–36, 2012.

[74] S. Hosseinpour, M. Aghbashlo, M. Tabatabaei, and E. Khalife, "Exact estimation of biodiesel cetane number (CN) from its fatty acid methyl esters (FAMEs) profile using partial least square (PLS) adapted by artificial neural network (ANN)," *Energy Convers. Manag.*, vol. 124, pp. 389–398, 2016.

[75] K.-B. O. Pik-Yin Foo , Voon-Hsien Lee , Garry Wei-Han Tan, "A gateway to realising sustainability performance via green supply chain management practices: A PLS-ANN approach," *Inf. Softw. Technol.*, 2018.

[76] M. R. Smith, M. S. Gashler, and T. Martinez, "A Hybrid Latent Variable Neural Network Model for Item Recommendation," in *2015 International*

Joint Conference Neural Networks (IJCNN), 2015.

[77] J. Peng, L. Li, and Y. Y. Tang, "Combination of activation functions in extreme learning machines for multivariate calibration," *Chemom. Intell. Lab. Syst.*, vol. 120, no. January, pp. 53–58, 2013.

[78] Y. Han, J. Chen, T. Pan, and G. Liu, "Determination of glycated hemoglobin using near-infrared spectroscopy combined with equidistant combination partial least squares," *Chemom. Intell. Lab. Syst.*, vol. 145, pp. 84–92, 2015.

[79] D. M. L. W. Aung, "Soil Property Prediction: An Extreme Learning Machine Approach," *Neural Inf. Process.*, no. May, 2015.

[80] Z. Tian, B. Gu, L. Yang, and Y. Lu, "Hybrid ANN-PLS approach to scroll compressor thermodynamic performance prediction," *Appl. Therm. Eng.*, vol. 77, pp. 113–120, 2015.

[81] L. Zhang *et al.*, "Blood species identification using Near-Infrared diffuse transmitted spectra and PLS-DA method," *Infrared Phys. Technol.*, vol. 76, pp. 587–591, 2016.

[82] X. Bian *et al.*, "A boosting extreme learning machine for near-infrared spectral quantitative analysis of diesel fuel and edible blend oil samples," *Anal. Methods*, 2017.

[83] M. C. A. Costa, M. A. Morgano, M. M. C. Ferreira, and R. F. Milani, "Analysis of bee pollen constituents from different Brazilian regions: Quantification by NIR spectroscopy and PLS regression," *LWT - Food Sci. Technol.*, vol. 80, pp. 76–83, 2017.

[84] A. Al-Harrasi *et al.*, "Application of NIRS coupled with PLS regression as a rapid, non-destructive alternative method for quantification of KBA in Boswellia sacra," *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 184, pp. 277–285, 2017.

[85] P. S. Sampaio, A. Soares, A. Castanho, A. S. Almeida, J. Oliveira, and C. Brites, "Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms," *Food Chem.*, vol. 242, no. February

2017, pp. 196–204, 2018.

[86] M. Sajdak and M. Kotyczka-Morańska, “Development and validation of a fast method based on infrared spectroscopy for biochar quality assessment,” *Biomass and Bioenergy*, vol. 112, no. February 2017, pp. 99–109, 2018.

[87] H. Grossman, Robert; Seni, Giovanni; Elder, John; Agarwal, Nitin; Liu, *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, 2010.

[88] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016.

[89] T.-T. Wong, “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation,” *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.

[90] A. N. S. Ohail *et al.*, “Analysis of hepatitis C infection using Raman spectroscopy and proximity based classification in the transformed domain,” *Biomed. Opt. Express*, vol. 9, no. 5, pp. 2041–2055, 2018.

[91] P. García-Moreno, A. Artés-Rodríguez, and L. K. Hansen, “A Hold-out method to correct PCA variance inflation,” *2012 3rd Int. Work. Cogn. Inf. Process. CIP 2012*, vol. 2, no. 3, pp. 2–7, 2012.

[92] T. Fearn, “How reliable is an RMSEP from a validation set?,” *NIR news*, vol. 13, no. 4, pp. 20–21, 2002.

[93] G. Jiang and W. Wang, “Error estimation based on variance analysis of k -fold cross-validation,” *Pattern Recognit.*, vol. 69, pp. 94–106, 2017.

[94] C. L. Cheng, Shalabh, and G. Garg, “Coefficient of determination for multiple measurement error models,” *J. Multivar. Anal.*, vol. 126, pp. 137–152, 2014.

[95] X. Jintao, Y. Liming, L. Yufei, L. Chunyan, and C. Han, “Noninvasive and fast measurement of blood glucose in vivo by near infrared (NIR) spectroscopy,” *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 179, pp. 250–254, 2017.

[96] D. Svozil, V. Kvasnička, and J. Pospíchal, “Introduction to multi-layer feed-forward neural networks,” *Chemom. Intell. Lab. Syst.*, vol. 39, no. 1, pp. 43–62, 1997.

[97] S. Haykin, “FEEDFORWARD NEURAL NETWORKS: AN INTRODUCTION,” *Convergence*, pp. 1–16, 1998.

[98] Mathworks, “Multilayer Neural Network Architecture,” *MATLAB R2017b*, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ug/multilayer-neural-network-architecture.html. [Accessed: 12-Nov-2017].

[99] Mathworks, “Nguyen-Widrow layer initialization function,” *MATLAB R2017b*, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ref/initnw.html. [Accessed: 12-Nov-2017].

[100] Mathworks, “Levenberg-Marquardt backpropagation,” *MATLAB R2017b*, 2017. [Online]. Available: https://www.mathworks.com/help/nnet/ref/trainlm.html. [Accessed: 12-Nov-2017].

[101] M. R. Abdullah, A. B. H. M. Maliki, R. M. Musa, N. A. Kosni, and H. Juahir, “Intelligent Prediction of Soccer Technical Skill on Youth Soccer Player ’ s Relative Performance Using Multivariate Analysis and Artificial Neural Network Techniques,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 5, pp. 668–674, 2016.

[102] G. Binetti *et al.*, “Cultivar classification of Apulian olive oils: Use of artificial neural networks for comparing NMR, NIR and merceological data,” *Food Chem.*, vol. 219, pp. 131–138, 2017.

[103] N. Merh, V. P. Saxena, and K. R. Pardasani, “a Comparison Between Hybrid Approaches of Ann and Arima for Indian Stock Trend Forecasting,” *Bus. Intell. J.*, pp. 23–43, 2010.