Prediction of HB, TW, and LDH in Normal and Leukemia Subjects Using PLSR Analysis

Herlina Abdul Rahim[1]*, Intan Maisarah Abd Rahim[1], and Muhd Halalluddin Abdul Rahim[2]

[1]School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
[2]Electrical Engineering Department, Politeknik Sultan Azlan Shah, 35950 Behrang Perak

Corresponding author* email: herlina@utm.my

Accepted 3 March 2021, available online 31 March 2021

ABSTRACT

There are numerous applications of Near-Infrared spectroscopy (NIRs) in the medical field. For instance, the application of NIRs can be found in the process of determining a blood parameter which acts as biomarker of specific diseased. This paper proposed a non-invasive leukemia early stage screening technique based on blood hemoglobin (HB), total white cell (TW), and lactate dehydrogenase (LDH). The spectral data was acquired by applying NIRs directly at human fingertip and laboratory data was acquired from standard blood drawing procedure. During PLSR analysis the spectral data was compared with laboratory data for validation. The correlative coefficient of prediction (RP) for normal dataset is 0.9168, 0.8876, and 0.8307 for HB, TW, and LDH respectively. The correlative coefficient of prediction (RP) for leukemia dataset is 0.8588, 0.8868, and 0.8307 for HB, TW, and LDH respectively. The blood parameter with the highest R and smallest difference between RMSEC and RMSEP for both normal and leukemia data sets are HB, TW, and LDH accordingly. The high correlation level between the predicted values and the reference values proves the potential of using NIR spectral information for non-invasive early stage leukemia screening.

Keywords: Near Infrared spectroscopy (NIRs); Near Infrared (NIR); Partial Least Square Regression (PLSR); leukemia screening; non-invasive

1. Introduction

Leukemia is a blood forming cells cancer in the bone marrow [1]. It is a cancer of white blood cells (WBC) and characterized by an abnormal multiplication or mitosis of white cells in the bone marrow [2-3]. Many studies done related to NIR applied directly at human body but none in leukemia screening. Moreover, the earlier studies related to leukemia only focus on spectral biochemical characterization between normal and leukemia samples or specimens (full blood, blood plasma, cell, etc) instead of directly applied on human body [4-7]. The common biochemistry parameter measured using spectroscopic technique in characterization between normal and leukemia cell are lipids, protein, and DNA [4, 7]. However, it requires sample preparation (centrifuges process, replicate films preparation, and involved the use of some chemical reagent) and highly trained person [7]. Hence, this study proposed a new non-invasive approach for leukemia screening based on the blood parameter/biomarker.

2. Methodology

The methodology flow diagram for noninvasive leukemia screening using NIR spectroscopy method is shown in Figure 1. Firstly, the cases and control subjects must fulfill the exclusion and inclusion criteria required to be enrolled in this study. Then, NIR will be applied non-invasively at the subject fingertip for spectral data acquisition. The same subjects will also undergo the invasive procedure of blood draw for laboratory/reference data acquisition. Blood test is performed (full blood count (FBC) and liver function test (LFT)) to measure the laboratory data. The NIR spectral data obtained will be pre-processed and the standard laboratory data is used for validation in the prediction modeling.

There will be 60 data (consist of 30 spectral data and 30 laboratory data) acquired for each case (leukemia patient) and control (blood donor) group. The spectral data was acquired non-invasively using Reflectance NIRs (900nm-2600nm) from ARC Optix and the laboratory data was acquired invasively through blood drawing procedure. The analysis is performed using the full spectrum data. The data acquisition begins after the doctor in charge explained and discussed the required criteria to the staffs in the wards to assist in patient recruitment process. Ethical issues regarding the experimental procedure in data acquisition is already reviewed and approved by the Human Research Ethics Committee of USM.

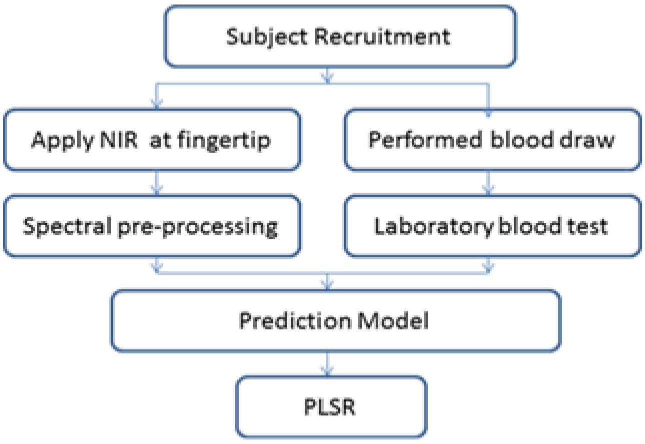


Figure. 1. Methodology flow diagram for non-invasive leukemia screening using NIR spectroscopy

2.1 Spectral Data Acquisition

Firstly, the experimental purposed and procedure involved during data acquisition is explained to the patients or blood donors. After consented to be enrolled in this study, NIR spectroscopy is set up and spectrum is acquired by applying the fiber optic at patient fingertip. In the case of normal subjects NIR is performed after the blood donating procedure is finished and donor is ready for the spectral data acquisition. The measurement is repeated for 5 times for each subject inthe period of 10 minutes (including process of consent application and setup preparation). The scanning period for eachspectral acquisition depends on the ability of the instrument used. This NIR spectroscopy requires around one minute per each spectral scanning. Each subject will provide 5 set of spectral data. Therefore, 30 spectral were acquired from each normal (six subjects) and leukemia (6 subjects) group. In medical study, 30 spectral data for a modelling is enough as the patients are limited. Previous study related to NIRs applied in the study of chronic lymphocytic leukemia cells also usingless than 40 data for each normal (28 data) and patients (38 data) group [4]. There were 30 raw NIR spectral acquired directly from human fingers of 6 normal subjects (blood donor). The spectral will be read using the provided software inthe suitable format before continuing with data analysis.

2.2 Laboratory/Reference Data Acquisition

The laboratory data obtained will be used as validation for NIR measurement data. In the case of leukemia patient, the measurement of TW, HB, and LDH which include blood drawing is a daily routine procedure to monitor the patient condition. The laboratory result obtained from the patient chart is used in data validation. However, in normal case which involved blood donor, extra procedure needs to be performed as it is not a routine procedure. The methodology for laboratory/reference data acquisition starts with the research explanation towards blood donor followed by asking for their consent to be involved in the study. Normal blood drawing procedure is performed with the help from the nurse on duty where extra blood is taken and put into two Ethylenediamine Tetraacetic Acid (EDTA) bottles as sample for lab test. One bottle is sent to the pathology chemistry lab for LDH test to get the value of total LDH and the other one is sent to the hematology lab for FBC test to get the value of total Hb and TW. This procedure is done without involving extra blood draw process. FBC is performed using Hematology Analyzer Machine and LFT is performed using Continuous Flow Analyzer to obtain the LDH value. The request for both tests must be done using the online application system before sending the samples to the specified laboratory. Then the laboratory test result is obtained and recorded by checking the online system. This took quite some time for the result to be available online.

2.3 Prediction Model

Linear calibration techniques with cross-validation technique is applied for an early screening of leukemia based on the biomarkers (LDH, TW, and Hb) properties from near infrared spectra. Partial Least Square Regression (PLSR) analysis was performed using MATLAB. Firstly, the absorption spectral data acquired need to be pre-processed before continue with the prediction modeling. This involved the process of model design, model calibration and model validation. The result accuracy will determine the efficiency of the non-invasive method used in the actual practice. The holdout validation is combined with Leave-one-out cross validation (LOOCV) to verify the prediction model. In this analysis, holdout validation is used to create randomness in the calibration and prediction set split when applied repeatedly. During LOOCV, when one sample was removed from the total for validation (prediction), and then the training (calibration) model was built with the remaining samples. The procedure was repeated for all samples and the RMSECV was calculated. The selection of data set formed using holdout is based on the results of RMSECV.

3. Result and Discussion

There will be 30 spectral data and 30 laboratory data acquired using Reflectance NIR spectroscopy (900nm-2600nm) and blood drawing procedure respectively for each normal and leukemia group of data.

3.1 Spectral Data Pre-processing

The absorbance spectral was pre-processed to enhance the quality of the acquired data by eliminating or minimizing the effect of unwanted signal. The raw spectral were smoothing by applying zero derivative first polynomial order Savitzky-Golay (SG) Smoothing. SG-Smoothing with zero derivative order with first polynomial order is used for all biomarker set for both normal and leukemia as the accuracy using 1st and 2nd derivative order showed the same result.

In SG-smoothing, the optimum filter length (FL) is required to avoid over-fitting and under-fitting of both leukemia and normal spectrum. The optimal FL is determined at the highest RCV and lowest RMSECV in previous Principal Component Regression (PCR) analysis. FL from 5nm to 31 nm is tested with predictor numbers varied from 1 to 15. The optimal FL for normal-HB is determined by first plotting the graph of RMSECV against number of PC as shown in Figure 2 (a) Then, PC4 with the lowest RMSECV is selected and graph of RMSECV against FL of PC4 is plotted as shown in Figure 2 (b). Five is the optimal FL selected for normal-HB data as it has the lowest RMSECV. These processes are repeated for each blood parameter of both normal and leukemia data.

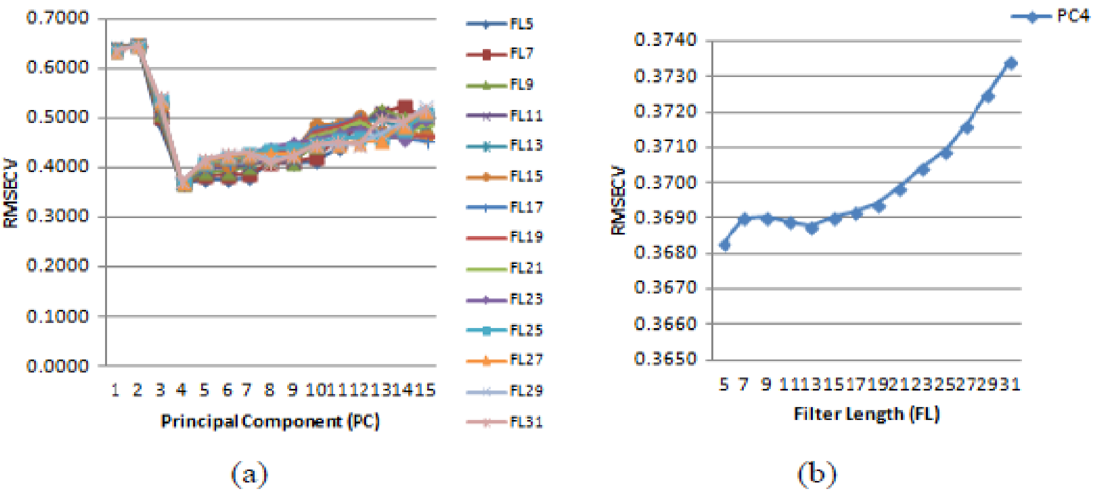


Figure. 2. Graph of RMSECV of zero order derivatives NIR spectral versus (a) number of principal component and (b) number of filter length for normal HB

The optimum FL of each blood parameter for normal and leukemia spectral was tabulated in the Table 1 and Table 2, respectively. The optimum FL for normal spectral is lower than that of leukemia for LDH parameter which means it contains less effect of unwanted signal.

Table 1. Optimum FL for HB, TW, and LDH of Normal Data

[a] BP	[b] DO	[c] PO	[d] PC	[e] FL	[f] RCV	[g] RMSECV
HB	zero	1	4	5	0.9493	0.3683
TW	zero	1	13	29	0.8930	0.6250
LDH	zero	1	8	29	0.9164	27.3645

Table 2. Optimum FL for HB, TW, and LDH of Leukemia Data

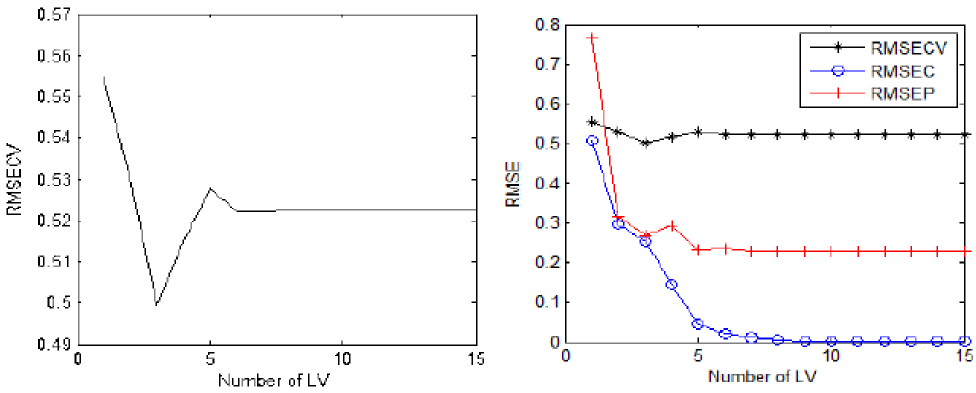
[a] BP	[b] DO	[c] PO	[d] PC	[e] FL	[f] RCV	[g] RMSECV
HB	zero	1	7	5	0.6738	1.8148
TW	zero	1	8	29	0.6119	4.4167
LDH	zero	1	4	11	0.8199	217.8269

[a] BP : Blood Parameter
[b] DO : Derivative order
[c] PO: Polynomial order
[d] FL: Filter length
[e] PC: Principal components
[f] RCV : Correlation coefficient of cross-validation
[g] RMSECV : Root mean square error of cross-validation

3.2 PLSR Analysis

The prediction of blood parameter HB, TW, and LDH are done using PLSR model. In PLSR, the optimum latent variables (LV) is determined by first plotting the graphs of root mean square error cross-validation (RMSECV), calibration (RMSEC), and prediction (RMSEP) with varying number of LV. Then, the optimum LV is selected based on the lowest RMSECV and met the condition where RMSEC is lower than that of RMSEP and RMSECV (RMSEC<RMSEP<RMSECV). Then, RC must be larger than RP (RC>RP) to avoid over fitting [33]. The optimum number of LV for normal-HB data is three (LV3) with the lowest RMSECV as shown in Figure 3 (a) and fulfilled the condition where RMSEC<RMSEP<RMSECV as shown in Figure 3 (b). The performance result for HB, TW, and LDH for both normal data set and leukemia data set is tabulated in Table 3 and Table 4 respectively.

The scatter plot for calibration and prediction of normal data set for HB, TW and LDH are each shown in Fig.4, Fig.5, and Fig.6. Then, the scatter plot for calibration and prediction of leukemia data set for HB, TW and LDH are each shown in Fig.7, Fig.8, and Fig.9. The performance of PLSR model for each blood parameter for both normal and leukemia data shows a satisfying result as the difference between RP and RC for each blood parameter is not huge. The RP value for normal data is higher than leukemia data for HB and TW and same for LDH.



(a) (b)

Figure. 3. (a) RMSECV with varying number of LV and (b) RMSECV, RMSEC, and RMSEP with varying numberof LV of Normal-HB

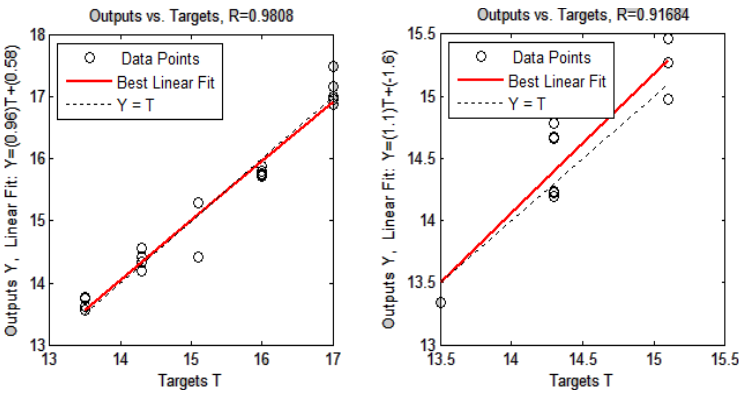
Table 3. Performance of PLSR of Normal Data Set

	[a]LV	[b]RCV	[c]RMSECV	[d]RC	[e]RMSEC	[f]RP	[g]RMSEP
HB	3	0.9243	0.4994	0.9808	0.2540	0.9168	0.2691
TW	3	0.6917	0.9737	0.9316	0.4866	0.8876	0.8615
LDH	3	0.8918	33.7133	0.9643	19.6891	0.8307	30.4161

Table 4. Performance of PLSR for Leukemia Data Set

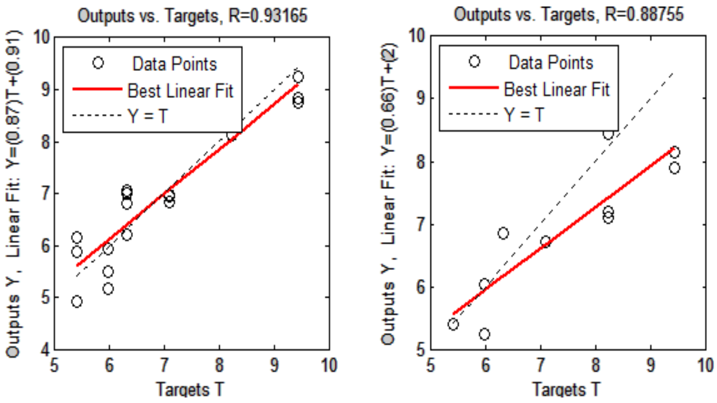
	[a]LV	[b]RCV	[c]RMSECV	[d]RC	[e]RMSEC	[f]RP	[g]RMSEP
HB	4	0.3816	2.3166	0.9897	0.3384	0.8588	1.8046
TW	5	0.2782	5.6990	0.9890	0.8252	0.8868	3.9976
LDH	4	0.6692	296.9092	0.9730	86.4395	0.8370	221.0982

[a] LV: Latent Variables
[b] RCV: Correlation coefficient of cross-validation
[c] RC : Correlation coefficient of calibration
[d] RP : Correlation coefficient of prediction
[e] RMSEC : Root mean square error of calibration
[f] RMSEP : Root mean square error of prediction
[g] RMSECV : Root mean square error of cross-validation



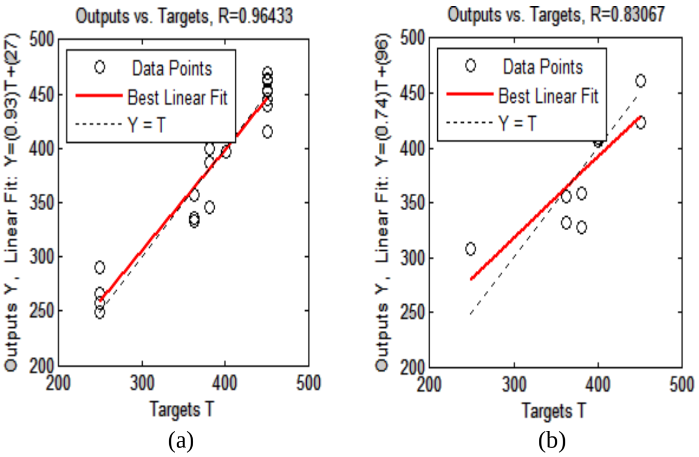
(a) (b)
(b)

Figure. 4. The best linear fit for (a) calibration and (b) prediction model of normal HB



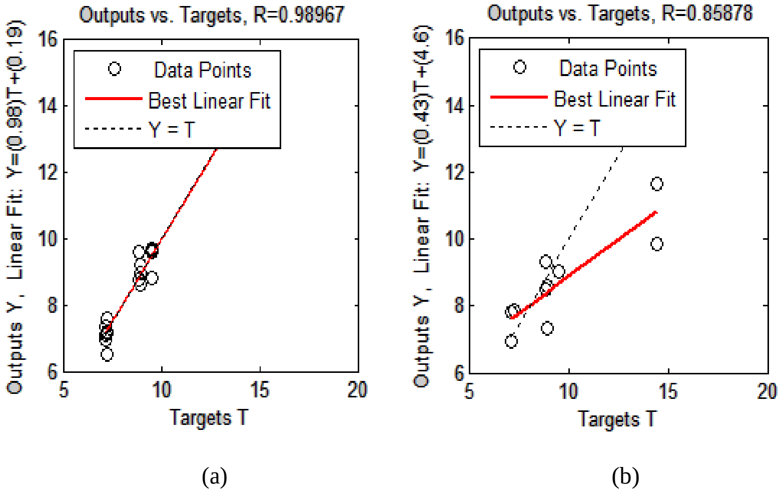
(a) (b)

Figure. 5. The best linear fit for (a) calibration and (b) prediction model of normal TW



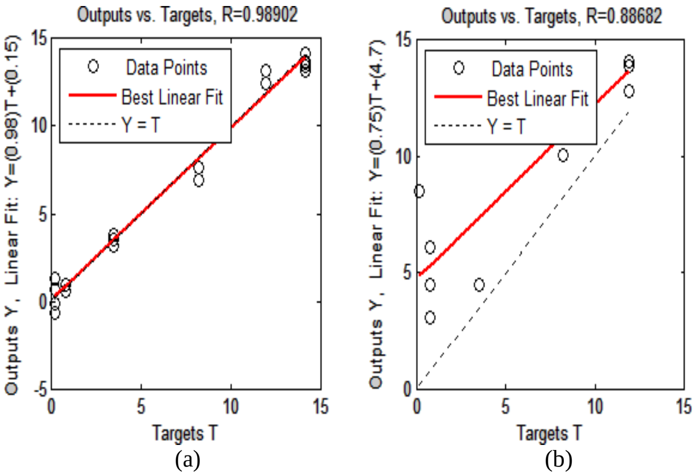
(a) (b)

Figure. 6. The best linear fit for (a) calibration and (b) prediction model of normal LDH



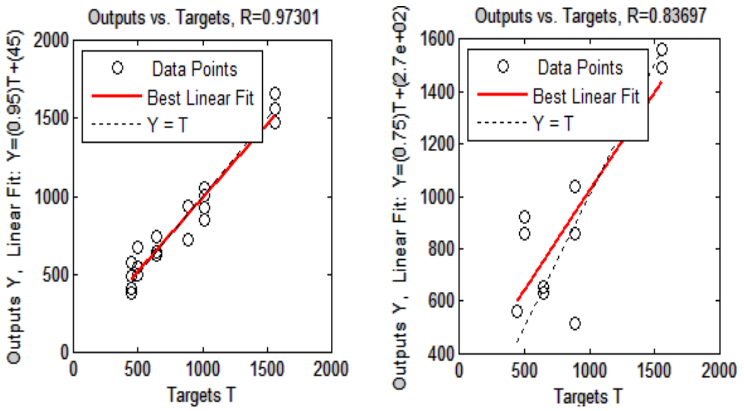
(a) (b)

Figure. 7. The best linear fit for (a) calibration and (b) prediction model of leukemia HB



(a) (b)

Figure. 8. The best linear fit for (a) calibration and (b) prediction model of leukemia TW



(a) (b)

Figure. 9. The best linear fit for (a) calibration and (b) prediction model of leukemia LDH

4. Conclusion

This study utilized full NIR spectral (900nm - 2600nm) in the prediction of HB, TW, and LDH of normal and leukemia subjects using PLSR Analysis. The value of RP for normal dataset is 0.9168, 0.8876, and 0.8307 for HB, TW, and LDH respectively. The value of RP for leukemia dataset is 0.8588, 0.8868, and 0.8307 for HB, TW, and LDH respectively. The blood parameter with the highest R and smallest difference between RMSEC and RMSEP for both normal and leukemia data sets are HB, TW, and LDH accordingly. This result shows that there is a high correlation level between the predicted values and the reference values in both normal and leukemia dataset. Therefore, the good performance of PLSR predictive model proves the potential in using NIR spectral information acquired using NIRs as a new non-invasive, simple, and faster leukemia early stage screening technique based on the blood HB, TW, and LDH.

Acknowledgement

The authors wish to thank the Ministry of Higher Education of Malaysia (MOHE) and Universiti Teknologi Malaysia (UTM) for financing this research with Vote No. (14H79). Then, the authors would like to show our gratitude to the medical staff from HUSM Kubang Kerian, Kelantan for their guide and assist while conducting the experiment. Special thanks to the patients and blood donors who enrolled in this study for their contribution.

References

[1] A. S. O. Hematology. Leukemia. 2016. Available: http://www.hematology.org/Patients/Cancers/Leukemia.aspx

[2] Mostaco-Guidolin, L.B., Murakami, L.S., et al. 2009. *Fourier transform infrared spectroscopy of skin cancer cells and tissues*. Applied Spectroscopy Reviews. **44**(5):438-455.

[3] Muda, Z. Leukemia. 2012. Available: http://www.myhealth.gov.my /index.php/en/kids/blood-disorder/leukemia

[4] Schultz, C.P., Liu, K.Z., et al. 1996. *Study of chronic lymphocytic leukemia cells by FT-IR spectroscopy and cluster analysis.* Leukemia research. **20**(8): 649-655.

[5] Schultz, C.P., Liu, K.Z., et al. 1997. *Prognosis of chronic lymphocytic leukemia from infrared spectra of lymphocyte*s. Journal of molecular structure. 408: 253-256.

[6] Benedetti, E., Bramanti, E., et al. 1997. *Determination of the relative amount of nucleic acids and proteins in leukemic and normal lymphocytes by means of Fourier transform infrared microspectroscopy*. Applied spectroscopy. **51**(6): 792-797.

[7] Mostaco-Guidolin, L.B., and Bachmann, L. 2011. *Application of FTIR spectroscopy for identification of blood and leukemia biomarkers: A review over the past 15 years*. Applied Spectroscopy Reviews. **46**(5): 388-404.

[8] Derrick, M.R., Stulik, D. and Landry, J.M. 2000. *Infrared spectroscopy in conservation science*. Getty Publications.

[9] Olsztyńska-Janus, S., Malek,K.S., et al. 2012. *Spectroscopic techniques in the study of human tissues and their components*. Part I: IR spectroscopy. Acta Bioeng Biomech. 14: 101-115.

[10] Yoshinari, H., Ishizawa, H. et al. 2012. *Non-invasive self monitoring of blood glucose system using near-infrared spectroscopy*. SICE Annual Conference 2012 Proceedings Akita. 20-23 August 2012. 1852-1854.

[11] Thomas, D. W. 2004. Advanced Biomaterials for Medical Applications. Netherland: Springer.

[12] M. W. Incorporated. Online Medical Dictionary. Available: http://c.merriam-webster.com/medlineplus/biomarker

[13] Advani, S. et al. 1999. *Acute lymphoblastic leukemia in India: an analysis of prognostic factors using a single treatment regimen*. Annals of Oncology. **10**(2): 167-176.

[14] Abu Zaid, Z. et al. 2009. *Assessing the Nutritional Status of Children with Leukemia from Hospitals in Kuala Lumpur.* Malaysian journal of nutrition. **15**(1): 45-51.

[15] O. College. 2013. *Anatomy & Physiology: Chapter 18 (The Cardiovascular System: Blood).* Available: http://cnx.org/content/col11496/1.6/

[16] Siesler, H.W., Ozaki, Y., Kawata, S. and Heise, H.M. eds. 2008. *Near-Infrared Spectroscopy: Principles, Instruments, Applications.* John Wiley & Sons.

[17] O'Brien, S., Vose, J.M. and Kantarjian, H.M. eds. 2010. Management of Hematologic Malignancies. Cambridge University Press.

[18] Nguyen, S., et al. 2002. *A white blood cell index as the main prognostic factor in t (8; 21) acute myeloid leukemia (AML): a survey of 161 cases from the French AML Intergroup*. Blood. **99**(10): 3517-3523.

[19] Aguayo, A., et al. 1999. *Cellular vascular endothelial growth factor is a predictor of outcome in patients with acute myeloid leukemia*. Blood. **94**(11): 3717-3721.

[20] Bankapur, A., Zachariah, E., et al. 2010. *Raman tweezers spectroscopy of live, single red and white blood cells.* PLoS one. **5**(4): 10427.

[21] Epstein, J. Lactate dehydrogenase test. Available: http://www.healthline.com/health/lactate-dehydrogenase-test#Overview1

[22] Bierman, H., et al. 1957. *Correlation of serum lactic dehydrogenase activity with the clinical status of patients with cancer, lymphomas, and the leukemias*. Cancer research. **17**(7): 660-667.

[23] Hafiz, M.G. and Mannan, M. 2007. *Serum lactate dehydrogenase level in childhood acute lymphoblastic leukemia*. Bangladesh Medical Research Council Bulletin. **33**(3): 88-91.

[24] Kornberg, A. and Polliack, A. 1980. *Serum lactic dehydrogenase (LDH) levels in acute leukemia: marked*. Blood. **56**(3): 351.

[25] Kornberg, A. 1991. For the love of enzymes: *The odyssey of a biochemis*t: Harvard University Press.

[26] L. Aalto Scientific. Lactate Dehydrogenase, (Lactic Dehydrogenase/LDH/ LD). Available: http://www.aaltoscientific.com/purifiedhumanproteins/LactateDehydrogenase.php

[27] Bood, J. and Carlsson, H. 1995. *Raman and Infrared Spectroscopy for Tissue Diagnostics*. Lund Reports in Atomic Physics.

[28] Chia, K.S., Abdul Rahim, H. , Abdul Rahim, R. 2013. *Evaluation of common pre-processing approaches for visible (VIS) and shortwave near infrared (SWNIR) spectroscopy in soluble solids content (SSC) assessment*, Biosystems Engineering. vol. 115, Issue **1**: 82–88.

[29] Abdi, H. 2010. *Partial least squares regression and projection on latent structure regression (PLS Regression)*. Wiley Interdisciplinary Reviews: Computational Statistics. **2**(1): 97-106.

[30] Wang, J., K. Nakano, et al. 2011. *Nondestructive evaluation of jujube quality by visible and near-infrared spectroscopy*. LWT-Food Science and Technology. **44**(4): 1119-1125.

[31] Gomez, A.H., He, y., et al. 2006. *Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques*. Journal of Food Engineering. **77**(2): 313-319.

[32] Zou, K. H., K. Tuncali, et al. 2003. *Correlation and Simple Linear Regression*. Radiology. **227**(3): 617-628.

[33] Seng, C. K. 2014. *Prediction Models and Shortwave Near Infrared Spectroscopic Analysis In Nondestructive Soluble Solids Content Assessment of Pineapples*. Ph.D. Thesis. Universiti Teknologi Malaysia.