Determination of Glucose Concentration from Near-Infrared Spectra Using PLSR Coupled with a Median Filter

Amneh A. Al-Mbaideen

Department of Electrical Engineering, Mu'tah University, Al-Karak, Jordan
e-mail: a.mbaideen@mutah.edu.jo

Received: September 4, 2018 *Accepted: October 28, 2018*

Abstract— This paper investigates the use of the median filter (MF) for the quantitative analysis of the NIR spectra. MF is a nonlinear low pass filter used widely in digital image processing for smoothing and at the same time the edges were preserved. MF is more robust and less sensitive to outliers compared to the Moving average filters. The use of MF for the quantitative analysis of the NIR spectra has not been previously evaluated in the field of chemometrics. In this work, MF is used as a pre-processing method to the Partial Least Squares Regression (PLSR) model. The effect of using MF has been evaluated and compared to the Moving Window Average (MWA) filter and Savitzky-Golay filter by computing the Standard Error of Prediction (SEP), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). The model is validated using different mixtures composed of glucose, urea and triacetin dissolved in a phosphate buffer solution. The results show that using MF combined with the MWA filter and PLSR improves SEP of the PLSR model from 35.6 mg/dL to 18 mg/dL.

Keywords— Glucose, median filter, NIR spectroscopy, noninvasive, PLSR.

I. INTRODUCTION

Most biological components have absorption bands in the NIR range 4000-12800 (2.5μm - 0.78 μm), where each biological component has a unique response to the NIR signals [1]-[10]. The variation of intensity of the reflected or absorbed NIR radiation is correlated with the variation of the entire components concentration. This property brought the usage of NIR spectroscopy to the forefront of the noninvasive glucose monitoring research [1]-[10].

The concentration of the Glucose is extracted from the collected NIR spectra by establishing a linear calibration model such as Principal Component Analysis (PCA) [11], Partial Least Squares (PLSR) [4] and independent component regression (ICR) [12]. However, the NIR spectra of biological components is weak in intensity, broad and overlapped [1]-[6]. Furthermore, the quality of the collected spectra is influenced by several factors, including instrument noise, baseline variations, and light scattering due to the features of NIR spectroscopy [10].

The above factors affect the capability of getting robust and stable calibration model that can be used to predict glucose concentration from the collected spectra with a clinically accepted range of error. Therefore, preprocessing methods are required to improve the quality of the collected spectra prior to calibration model implementation [13].

The objective of the preprocessing method is to remove base line variations, high frequency noise, and high background variations. It is also used to enhance appearance resolution for scattering correction. The most widely used pre-processing techniques are: smoothing [15]-[17] and band pass filters [4], [7], linear and nonlinear methods, derivative and Frequency self-deconvolution [6], and scatter-corrective pre-processing methods [16].

This paper presents the application of MF [18]-[19] in the field of quantitative analysis of the NIR spectroscopy. To the best of author's knowledge, despite the large number of researches on MF applications in the image processing [20]-[21], there are no published works

Corresponding author's e-mail: a.mbaideen@mutah.edu.jo addressing the applications of MF on the prediction of glucose concentration from the NIR spectroscopy. MF is a nonlinear filter that is used widely in image processing; it has the ability to deal with the shot or spiky noise even without any prior information about noise distribution [19]. It can deal with the multiplicative noise, which cannot be removed using ordinary linear filters.

In this work, MF is combined with PLSR regression model to enhance the quality of the collected spectra, which in turn, increase the prediction ability of glucose concentration. The grid search optimization method is used to optimize model parameters. The proposed combination of PLSR regression model with MF has been compared to the combination of PLSR regression model and the Moving average filter and Savitzky-Golay filter. The best results were obtained when both median and the MWA filters are combined with the PLSR model.

II. Materials and Methods

MF was described by Tukey [18] in 1970s as a tool for statistical data analysis. It is a nonlinear filter in which the input and output are not related by a linear function. MF is used as a preprocessing technique to remove noise from a signal or image. It replaces each point in the filtered signal by the median of a running window through the signal. In the filtering process, the entries of the input signal within each window are sorted in an ascending order. MF replaces the center point of the input data within the window with the median value of that window. The median value of the window is the middle value of the window. The median is the value separating the higher value from the lower values of the data within a selected window. If the window size (n) is odd, the median will be the center point. A one-dimensional MF of size *n=2k+1* is defined by the following input-output relation:

$$\mathrm{x_{fi}} = \mathrm{med}\{\mathrm{x_{i-k}}, \dots, \mathrm{x_i}, \dots, \mathrm{x_{i+k}}\} \tag{1}$$

Where *i* is the index of the ascended entries within the window; and *n* is the window size. x_{fi}, x_i are the filtered and input entries of the spectra, respectively. If the window size n is even, the median value will be the average value of the two middle points of the window.

Compared to other types of smoothing methods such as Savitzky-Golay filter [22] and Moving average filter [14]-[15], MF is widely used to reject the non-additive noise, spike (shot) noise. The spike noise affects a small range of the signal with a large of amount of noise. MF is not sensitive to extremely large and small values of the signal. Therefore, it can be considered as a robust method to the outliers compared to other smoothing methods [24]-[25].

MF is used widely in image processing to remove noise and at the same time to preserve the edges of the image [20]-[23], which is a critical factor in image analysis. On the other hand, conventional smoothing methods remove noise but they do not preserve the edges; and hence they cause a shift on the boundaries.

The Root Mean Square Difference (RMSD) between a filtered (y_f) signal and an unfiltered signal (y_{uf}) can be used to evaluate the effect of the filtering process on signals:

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{l}}\sum_{\mathrm{i}=1}^{\mathrm{l}}(\mathrm{y_f} - \mathrm{y_{uf}})^2} \tag{2}$$

where *l* is the number of points of the signal.

To illustrate the effect of the filtering process of MF and MWA filter on different types of signals, two forms of signals are generated using Matlab as shown in Figs. 1 and 2. The first waveform has a pulse peak shape; and the second has an exponential peak shape. A pulsed noise was added to both of the generated waveforms at different intervals with different widths. The generated waveforms were filtered using MF, MWA filter, and the coupling of median and MWA filters.

Figs. 1 and 2 show that MF is more robust to the pulsed noise than the MWA filter. The uniform and Gaussian pulsed shapes result from the difference in their mathematical definition. MF retains the original pulse shape of the original data by giving the same width and shape, particularly for the uniform pulsed shape. On the other hand, the MWA filter causes a distortion in the peaks of the filtered data, which is higher in the uniform waveforms, when there is a high pulsed noise.

The RMSD of the filtered Gaussian pulse shaped data using MWA and MF for different window sizes (n) is computed and shown in Table 1. The results show that the RMSD for both filters is decreased as the window size is increased, reaching an optimum value before being increased. MF shows a rapid reduction in the RMSD values compared to MWA. The optimum RMSD value of the MWA filter is 0.102 which is reduced to 0.060 for MF.

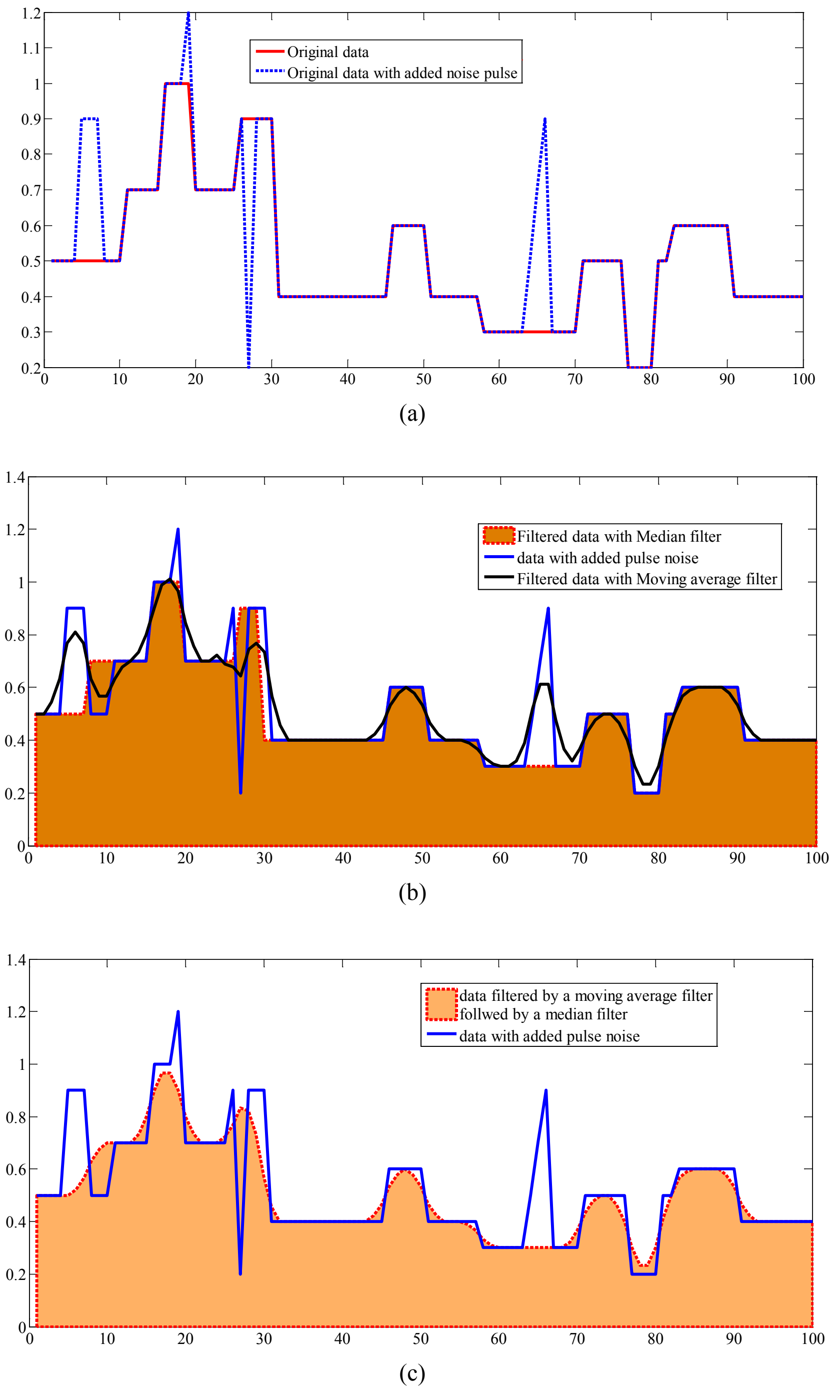
TABLE 1
THE RMSD OF THE FILTERED DATA USING MEDIAN AND MWA FILTERS WITH DIFFERENT WINDOW SIZES

Window size	MF	MWA Filter
3	0.126	0.11
5	0.122	0.102
7	0.117	0.113
9	0.079	0.142
11	0.060	0.178
13	0.076	0.216
15	0.115	0.252

III. EXPERIMENTAL DATA AND MEASUREMENT SETUP

In this work, MF is introduced to perform a quantitative analysis of the NIR spectroscopy of the Glucose. The collected NIR spectra span in the range from 2000 nm to 2500 nm (4000–5000 cm^{-1}) with a spectra resolution of 1 nm, where the most important features of glucose raw absorbance spectrum lie in this range. The measured spectra are collected using spectrophotometer Cary 5000 version 1.09. These mixtures are obtained by dissolving glucose (20 to 500) mgdL^{-1}, urea (0 to 50) mg dL^{-1}, and triacetin (10 to 190) mgdL^{-1} in a phosphate buffer solution. A quartz cuvette with a fixed path length of 1 mm is used to hold solutions during the spectra collection. The spectroscopy of each mixture is collected three times at different periods. The absorbance spectrum of the buffer is used as a reference.

The calibration phase is obtained by using 60 spectra that refer to the three replicate spectra of 20 mixtures. The testing phase is built from the triplicate spectra of 10 mixtures. All of these experiments were carried out in a non-controlled environment in order to evaluate the effect of using MF on the prediction ability of the PLSR calibration model. Matlab-13a is used to implement the proposed model.

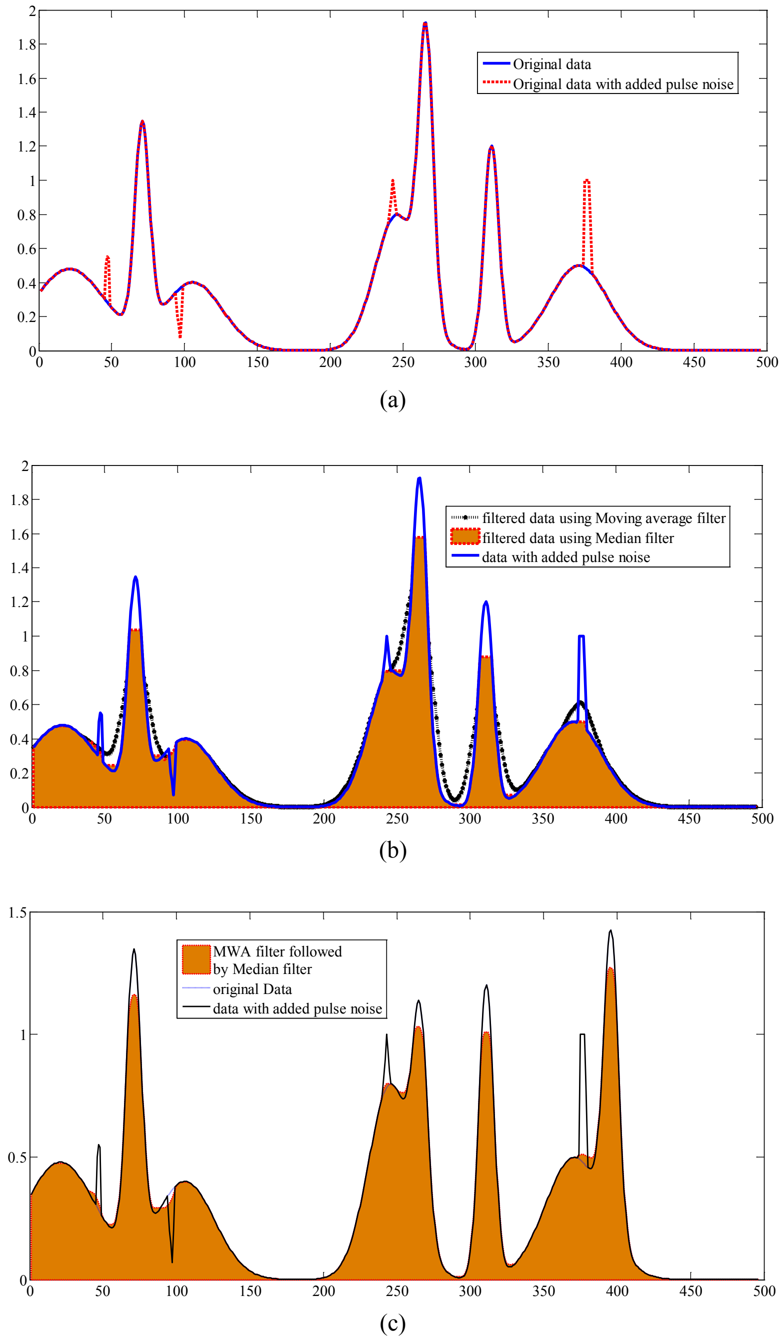


(a)

(b)

(c)

Fig. 1. The effect of filtering on a pulse peak shape: a) original data; b) smoothing by a MWA filter and MF, c) smoothing by combining median and MWA filters



(a)

(b)

(c)

Fig. 2. The effect of filtering on an exponential peak shape: a) original data, b) smoothing by the MWA and MF, c) smoothing by combining median and MWA filters

IV. Results and Discussion

In this work, various calibration models were founded by combining median, MWA and Savitzky-Golay filters with PLSR regression method. To obtain the optimal parameters of the filters and regression model, the grid search optimization method [1]-[6] was used. The NIR raw spectra were filtered with predetermined parameters of the filter. The PLSR calibration model was obtained and then used to predict glucose concentration of the testing data. SEP is computed. The whole process is then repeated several times with a different number of filter parameters. The reciprocal of the SEP of each model is plotted versus filter parameters; and a 3D plot is achieved as shown in Fig. 3. The optimum model produces the minimum SEP. Table 2 shows a summary of the optimum models obtained for each filter with their SEP, coefficient of determination (the square of the correlation coefficient) R^2, data mean percent error (MAPE) and the number of PLSR factors [24].

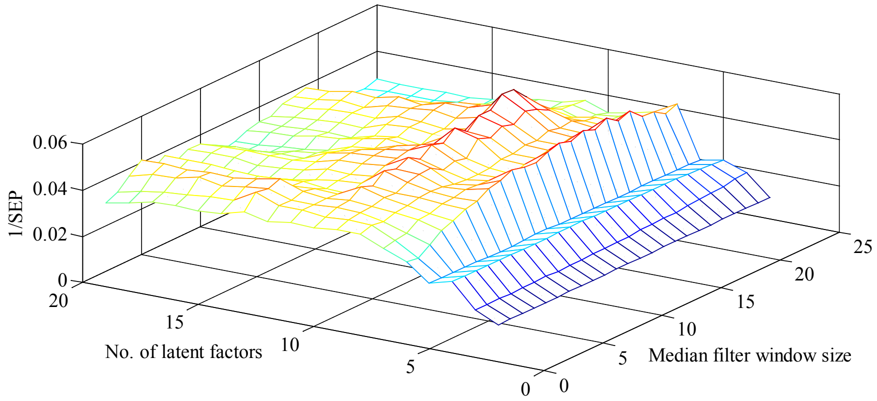


Fig. 3 The performance map of the optimal model of the PLS and MF: the reciprocal of SEP vs. the number of PLS factors and filter window size

Table 2 shows that the optimal PLSR model without preprocessing is obtained by employing 8 PLSR factors with SEP equals 35.6 mg/dL^{-1}, R^2 = 0.933 and MAPE=34.5. The collected spectra were then processed using different smoothing methods, i.e. MWA filter, MF, Savitzky-Golay filter, and finally MF followed by MWA filter. The processed spectra were used to compute the PLSR calibration model. For each method, the optimal model is obtained as explained previously. The results show that all the tested methods improve the capability of the PLSR model to predict glucose concentration. Table 2 shows that the MWA filter improves the SEP of the PLSR model by 41.6 %; MF by 37.8 %; Savitzky-Golay filter by 44.2%; and median-MWA filter by 49.4%. It is clear that the best model was obtained by combining median and MWA filters, which increased the correlation coefficient R^2 from 0.933 to 0.983. MF is a type of time domain filtering. Therefore, it is clear in Fig. 3 that MF has the effect of smoothing the signal. In Fig. 4, the third sample that corresponds to the sample with glucose concentration=99 mg/dL is displayed. MF suppresses most of the spectra variations that are not related to the chemical analyses which lower the capability of the PLSR model to predict glucose concentration.

TABLE 2
SUMMARY OF OPTIMAL FILTERING RESULTS

Smoothing method	Parameter and its optimum value	Optimum number of PLSR Factors	SEP mg/dL	MAPE	R^2
PLSR (without preprocessing)		8	35.6	34.5	0.933
PLSR-MWA filter	Window size=20	9	20.8	18.3	0.977
PLSR-MF	Window size=8	7	22.15	17.9	0.974
PLSR-Savitzky-Golay filter	Order=3 Window size=27	11	19.88	16.4	0.979
PLSR-MWA-MF	MWA=5 Median 18	11	18	14.1	0.983

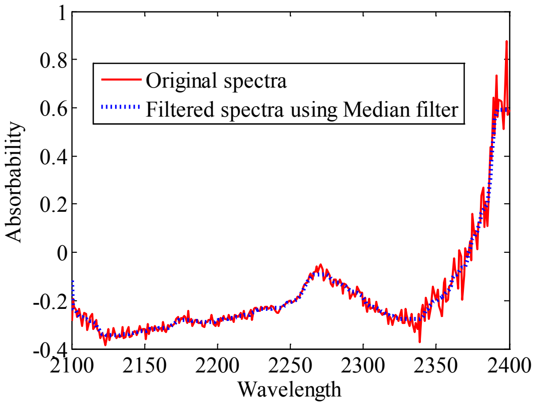


Fig. 4. The effect of using the MF on the raw spectra of the same sample (Wavelength in nm)

The model that combines median, MWA and Savitzky-Golay filters which PLSR is identified by the number of latent factors and window width of the filter. Therefore, it is necessary to study the effect of these factors on the optimal model of each preprocessing method. Keeping filter window width fixed, the PLSR calibration models were generated for each method by employing 3-20 latent factors as shown in Table 2. The generated PLSR calibration models were then used to predict glucose concentration of the testing data sets. The capability of the calibration models was evaluated by computing the SEP for each model. The computed SEP was plotted against the number of loading factors to determine the optimum number of loading factors that produces the minimum SEP as shown in Fig. 5.

Fig. 5 emphasizes the fact that the smoothing methods improve the prediction capability of the PLSR model. For all of the models, the SEP is decreased as the number of latent factors is increased. When the minimum SEP is produced, it increases again. Median-MWA filter has the best performance. The MWA and Savitzky-Golay filters have approximately the same performance.

To study the effect of the window width on MF, the optimum number of latent factors of the MF is obtained from Table 2. The same process is used to plot Fig. 6 when fixing the number of latent factors and changing the window width of MF from 3-25 instead. The results are plotted in Fig. 6 which illustrates that the window width of the filter has a significant effect on median-PLSR model prediction capability. The SEP has a clear minimum value at window width= 8. Therefore, there is a need for an optimization method to select the optimal value of the window width.

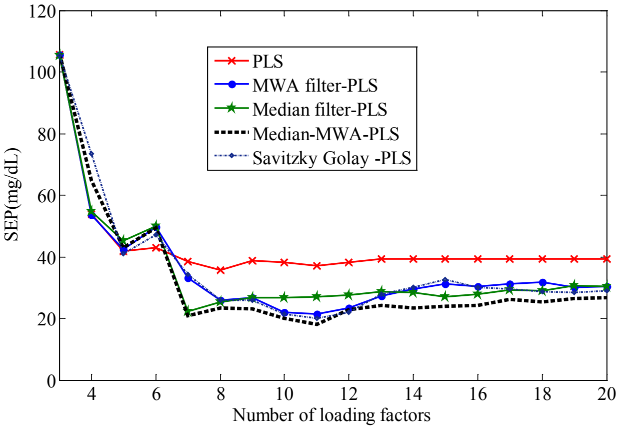


Fig. 5. Standard error of prediction SEP versus number of components for PLSR, MWA-PLSR, median-PLSR, Savitzky-Golay- PLSR, and median-MWA-PLSR models

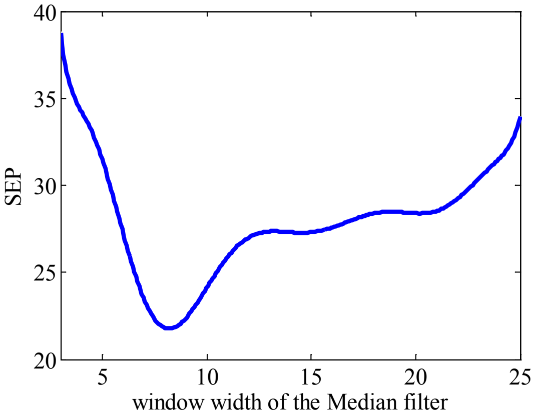


Fig. 6. Standard error of prediction SEP (mg/dL) versus window width median-PLSR models with number of latent factors= 7

Fig. 7 shows the predicted glucose concentration versus the reference glucose concentration of the training and testing data for the optimal PLSR and median-PLSR models. Fig. 7a indicates that the predicted concentrations have a high scattering around the perfect estimation line (reference line). Fig. 7b shows that MF improves the prediction capability of the PLSR model. This improvement in prediction can be clarified by plotting the Clarke error plot of median-PLSR model as shown in Fig. 8. Fig. 8b shows that most of the predicted values were in region A of the Clarke plot compared to Fig. 8a. Additionally, Fig. 8b shows that the prediction capability of median-PLSR model was consistent for low and high concentration levels.

The results in Table 2 show that median-MWA-PLSR model has the best prediction capability. The Clarke plot obtained for this model is shown in Fig. 9 which indicates that most of the points are within region A as compared to the PLSR and median-PLSR model. It also there are no points in region D.

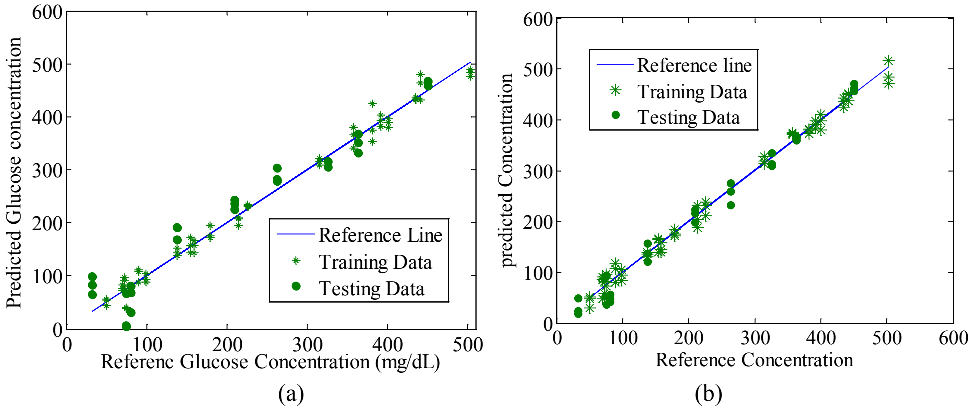
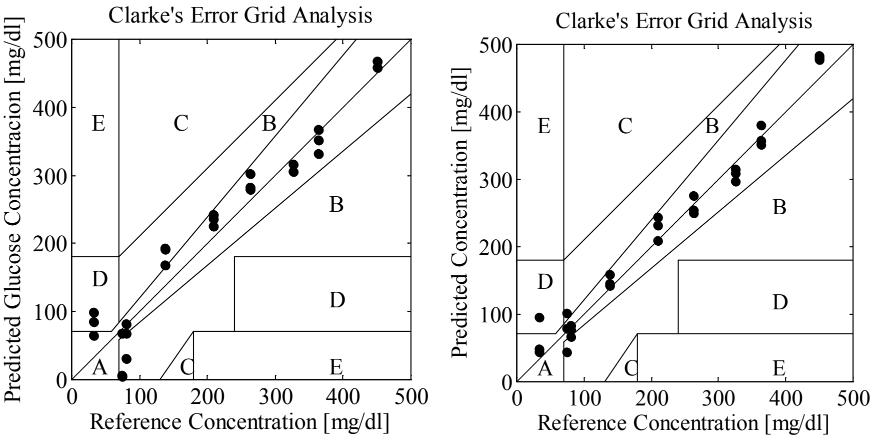


Fig. 7. The predicted glucose concentration (mg/dL) versus the actual concentration of the testing and training data that result from the: a) optimal PLSR model, b) optimal median-PLSR model



(a) (b)

Fig. 8. Clarke plot for the: a) optimal PLSR model, b) optimal median-PLSR

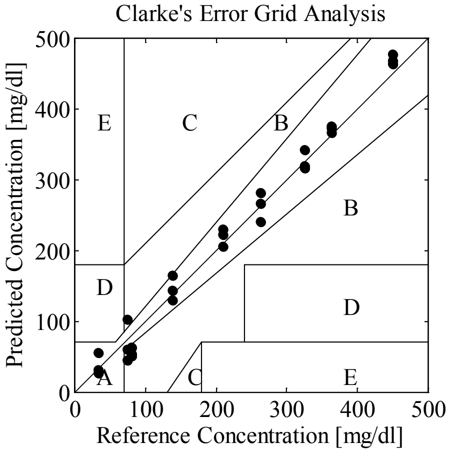


Fig. 9. Clarke plot for the optimal median-MWA-PLSR

V. CONCLUSIONS

This paper presents the application of MF in the field of the quantitative analysis of the NIR spectroscopy. The paper investigates the effect of using MF on the prediction capability of the PLSR calibration model that is constructed from NIR spectra. Various calibration PLSR models were developed by using different pre-processing techniques, including MF, MWA filter, Savitzky-Golay filter, and combination of the median and MWA filters. The optimum calibration model is selected using grid search optimization methods. The prediction capability of each model is validated using different mixtures composed of glucose, urea and triacetin dissolved in a phosphate buffer solution. It was found that the best model is obtained by using the median-MWA filters combined with the PLSR regression model. It produced the minimum value of SEP (18 mg/dL), the highest correlation coefficients (0.983) and the lowest MAPE (14.1) when 11 PLSR factors were used for the PLSR model.

REFERENCES

[1] M. Arnold and G. Small, "Noninvasive glucose sensing," *Analytical Chemestry*, vol. 77, no. 17, pp. 5429-5439, 2005.

[2] J. Burmeister and M. Arnold, "Evaluation of measurement sites for noninvasive blood glucose sensing with near-infrared transmission spectroscopy," *Clinical Chemistry*, vol. 45, no. 9, pp. 1621–1627, 1999.

[3] J. Chen, M. Arnold, and G. Small, "Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions," *Analytical Chemestry*, vol. 76, no. 18, pp. 5405-5413, 2004.

[4] F. Ham, I. Kostanic, G. Cohn, and B. Gooch, "Determination of glucose concentrations in an aqueous matrix from NIR spectra using optimal time domain filtering and partial least squares regression," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 6, pp.75-485, 1997.

[5] N. Cingo, G. Small, and M. Arnold, "Determination of glucose in a synthetic biological matrix with decimated time domain filtered Near-infrared interferogram data," *Vibrattional Spectroscopy*, vol. 23, no. 1, pp.103-117, 2000.

[6] A. Al-Mbaideen and M. Benaissa, "Frequency self deconvolution in the quantitative analysis of near infrared spectra," *Analytica Chimica Acta*, vol. 705, no. 1-2, pp. 135-147, 2011.

[7] A. Al-Mbaideen and M. Benaissa, "Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 108, no. 2, pp. 112-122, 2011.

[8] K. Patchava, M. Benaissa, and H. Behairy, "Partial least squares regression coupled with hilbert haung transformation pre-processing for the quantitative analysis of glucose in near infrared spectra," *Proceedings of the IEEE International Workshop on Signal Processing Symposium*, 2017.

[9] J. Garcia, D. Guaita, J. Gayete, S. Garrigues, and M. Guardiaa, "Determination of biochemical parameters in human serum by near-infrared spectroscopy," *Analytical Methods*, vol. 12, no. 6, pp. 39-82, 2014.

[10] A. Amerov, J. Chen, and M. Arnold, "Molar absorptivities of glucose and other biological molecules in aqueous solutions over the first overtone and combination regions of the near-infrared spectrum," Applied Spectroscopy, vol. 58, no. 10, pp. 1195-1204, 2004.

[11] A. Al-Mbaideen, T. Rahman, and M. Benaissa, "Determination of glucose concentration from near-infrared spectra using principle component regression coupled with digital bandpass filter," *Proceedings of the Signal Processing Systems IEEE Workshop*, pp. 243-248, 2010.

[12] A. Al-Mbaideen and M. Benaissa, "Determination of glucose concentration from NIR spectra using independent component regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 105, no. 1, pp 131-135, 2011.

[13] S. Rinnan, F. Berg, and S. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *Trends in Analytical Chemistry*, vol. 28, no. 10, 2009.

[14] A. Al-Mbaideen, "The application of moving average filter for the quantitative analysis of the NIR spectra," *Analytical Chemistry,* Accepted for publication in, 2018.

[15] A. Al-Mbaideen, M. Al-Soub, "Improving the prediction accuracy of the independent component regression calibration model for the quantitative analysis of the near infrared spectroscopy," *Natural and Applied Science Series*, Accepted for publication, 2018.

[16] K. Patchava, O. Alrezj, M. Benaissa, and H. Behairy, "Savitzky-golay coupled with digital bandpass filtering as a pre-processing technique in the quantitative analysis of glucose from

near infrared spectra," *Proceedings of the IEEE Engineering in Medicine and Biology Society Annual International Conference*, pp. 6210-6213, 2016.

[17] O. Alrezj, K. Patchava, M. Benaissa, and S. Alshebeili, "Scatter correction methods coupled with bandpass filtering as a pre-processing technique in the quantitative analysis of glucose in near infrared spectra," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1800-1803, 2017.

[18] J. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

[19] D. Stone, "Application of MFing to noisy data," *Canadian Journal of Chemistry*, vol. 73, no. 10, pp. 1573-1581, 1995.

[20] K. Verma, B. Singh, and A. Thoke, "An enhancement in adaptive median filter for edge preservation," *Procedia Computer Science*, vol. 48, pp. 29-36, 2015.

[21] T. Huang, *Two-Dimensional Digital Signal Processing: Transforms and MFs*, Springer-Verlag, 1981.

[22] A. Savitzky and Marcel Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemestry*, vol. 36, pp. 1627-1639, 1964.

[23] C. Lin, "An approach to improve the quality of infrared images of vein-patterns," *Sensors*, vol. 11, no. 12, pp. 11447-11463, 2011.

[24] R. Bakhshian, B. Emadi, M. Khojastehpour, M. Golzarian, and A. Sazgarnia, "Non-destructive evaluation of maturity and quality parameters of pomegranate fruit by visible/near infrared spectroscopy," *International Journal of Food Properties*, vol. 20, no. 1, pp. 41-52, 2017.

[25] A. Moghimi, M. Aghkhani, A. Sazgarnia, M. Sarmad, "Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit," *Biosystems Engineering*, vol. 106, no. 3, pp. 295-302, 2010.