01

# Influence of data pre-processing techniques for PLSR model to predict blood glucose by NIR spectroscopy

© S.Vasanthadev Suryakala, Shanthi Prince

Department of Electronics and Communication Engineering,
SRM Institute of Science and technology,
Kattankulathur-603203, Tamil Nadu, India

e-mail: suryakas@srmist.edu.in, shanthip@srmist.edu.in

*Received March 02, 2020*
*Revised May 30, 2020*
*Accepted September 27, 2021*

NIR diffuse reflectance spectroscopic spectra can be mathematically modelled to extract quantitative information by suitable multivariate calibration models. The analysis of spectral data becomes complex as the data is more prone to noise due to light scattering and baseline effects. These errors reduces the robustness and reliability of the developed calibration model. Hence data pre-processing becomes the most important aspect in data analysis. Different mathematical transformations are applied to remove the noise present in the data. This work focuses on the various empirical data pre-processing techniques like baseline correction, multiplicative scatter correction (MSC), robust MSC, extended multiplicative signal correction (EMSC), orthogonal signal correction (OSC) and (−log R) followed by standard normal variate (SNV) techniques for Partial Least Square Regression (PLSR) model in the prediction of blood glucose non-invasively. The performance of the PLSR model for the acquired (raw) spectral data and the same data subjected to different pre-processing techniques is analyzed. The model complexity and robustness is evaluated in terms of the number of latent variables (LVs) required to build the calibration model and obtained mean square prediction error after cross validation. This study utilizes the spectral data collected from 207 subjects from a diabetic center using Diffuse Reflectance Spectrometer (DRS). The analyzed results show that pre-processing based on (−log R) followed by SNV is found to perform well with reduced model complexity and minimum estimated mean square prediction error of 0.23 mg/dl among the other empirical pre-processing techniques.

**Keywords:** multiplicative scatter correction (MSC), orthogonal signal correction (OSC), standard normal variate (SNV), Diffuse Reflectance Spectrometer (DRS).

DOI: 10.21883/EOS.2022.05.54448.181-22

## Introduction

There are various optical measurement techniques for blood glucose detection such as Mid-infrared (MIR) spectroscopy, Raman spectroscopy, Fluorescence spectroscopy, Optical Coherence Tomography (OCT), Optical polarimetry, Near – infrared (NIR) spectroscopy [1]. The advantages and limitation of each technique is listed in Table 1. Among all these techniques NIR spectroscopy is found to be a prominent and most commonly used technique. The rationale for choosing NIR method discussed in this paper are its simplicity, increased detection sensitivity. The weak interaction of light with the tissue allows for deeper penetration up to 100mm in depth. NIR light is known to be safe for cells, does not induce auto fluorescence in cells. It has no strong cellular biological emitters, has more structural information and is inexpensive [2].

NIR spectroscopic techniques are rapid, non-destructive and non-invasive and are successfully used in the extraction of relevant information from biological samples by suitable multivariate calibration models. Partial least square regression (PLSR) method is one of the most prominent linear calibration model that is used to describe the relationship between the concentration of the chemical composition of the samples and the acquired spectra. Model development using PLSR involves pre-processing, model selection and model validation [3–5]. NIR spectroscopic measurements are highly influenced by non-linearity and baseline shift due to light scatter. The simplicity and robustness of the calibration model is highly affected by these detrimental effects accompanied in the spectral data. NIR diffuse reflectance spectroscopic measurements include both specular and diffuse reflectance. Specular reflectance is a mirror like reflection that occurs from the surface which can be avoided by proper placement of probe. Diffuse reflectance occurs when an incident light energy gets scattered at many angles from layers beneath the surface, thereby carrying the molecular information (chemical composition) in the underlying region of interest. As diffuse reflectance energy is from light scattered in different angles, the data has an inherent noise within. For biological samples, the scattering properties are excessively complex [6]. Hence pre-processing becomes the most critical part in chemometric modelling.

Pre-processing aims at reducing the unmodeled variability in the data thereby enhancing the linear correlation with the

**Table 1.** Comparison of various optical methods for blood glucose detection [1,2]

| S. No. | Optical method | Advantage | Disadvantage |
|---|---|---|---|
| 1 | Mid-infrared (MIR) | Low scattering and high absorption | Low penetration depth |
| 2 | Raman | Provides sharper and less overlapped spectra | Instability of laser wavelength and intensity, long acquisition time |
| 3 | Fluorescence | Very sensitive and can detect even single molecules | High scattering |
| 4 | Optical Coherence Tomography (OCT) | High resolution and penetration depth | Sensitive to individual's motion and temperature |
| 5 | Optical polarimetry | Glucose is the main component in aqueous humor present in the eye, hence high correlation with glucose exists in the measured signal | High errors due to eye movement and motion artefact |
| 6 | Near infrared (NIR) | High penetration depth, water is transparent to signal bandwidth of NIR, Less expensive | Weak correlation |

spectral data and the chemical composition. Therefore, various pre-processing techniques are proposed to eliminate the insignificant variations present in the data. Pre-processing technique not only improves the prediction ability but also results in a more parsimonious model that is robust and reliable. Pre-processing techniques are used to remove the variability present in the data. A complete pre-processing technique involves baseline correction, scatter correction, noise removal and scaling [7].

The present work focuses on performance of various pre-processing techniques suitable for NIR diffuse reflectance spectroscopy for blood glucose prediction using PLSR model. Model simplicity is evaluated with respect to the number of latent variables required to build the calibration model. The efficiency is determined using mean square prediction error obtained using cross validation. This study is executed using the diffuse reflectance spectral data collected from 207 subjects including both diabetic and non-diabetic subjects. NIR spectral range from 750nm -1040nm wavelength is considered for this analysis.

## Materials and methods

The physical features of spectroscopic method and experimental setup used for data acquisition and the various pre-processing techniques used for spectroscopic data are elaborated in the following sections.

### Characteristic Features of NIR Spectroscopy

NIR spectroscopy which is a subset of IR spectroscopy spans the spectral range of 750 nm to 2500 nm. The NIR spectrum is categorized into two regions according to the absorption characteristics namely long wavelength region (1300 nm to 2500 nm) and short wavelength region (700 nm to 1300 nm). NIR spectroscopy works on the basic principle of interaction of electromagnetic radiation with matter thereby producing the corresponding spectral signals. When NIR light interacts with the molecules of the given sample, exhibits a large number of weak, overlapping absorption bands in the spectral window. This proves that NIR spectra are generated by the excitation of the functional groups that have strong interatomic bonds. Absorption of energy in NIR band is dependent on the overtone and combination bands that are caused due to N−H, O−H, and C−H bonds in the long wavelength region. Hence these absorption peaks are weaker than that produced by the fundamental vibrations caused in the mid IR region. Absorptions in the long wavelength regions are stronger and sharper than the absorption corresponding to the vibration in the short wavelength region. But the shallow penetration depth of light in the long wavelength region makes it unsuitable for most of the biological applications. As water presents a relatively weak absorption signal from deep tissue in the region 700nm to 1300 nm region, this band is also called as the therapeutic optical window (identification of informative wavelength). For these reasons NIR is considered to be ideal for non-invasive measurements [8].

### Diffuse Reflectance Spectroscopic System

Diffuse reflectance is an optical phenomenon which is commonly used in NIR region to obtain the spectral characteristics of samples. When light is incident on a sample, it gets reflected in all directions. Reflection can be categorized as specular or regular reflection and diffuse reflection. Specular reflectance finds its major application in total internal spectroscopy to obtain spectroscopic information. Specular reflection can be described by Snell's law and Fresnel equations. Diffuse reflection results due to multiple reflections when light is incident on an inhomogeneous

medium. The angular distribution of the diffuse reflected signal is independent of the incident angle [9]. Moreover, diffuse reflectance measurements preserve the mechanical and biochemical degradation of the samples.

Diffuse reflectance relies on the projection of a focused beam of light in the tissue. The light that propagates through the tissue undergoes absorption and scattering events due to its constituent particle size. The light experiences multiple scattering events during propagation through the tissue and is almost isotropic when it emerges out of the tissue. Hence, it is considered to be diffuse reflected signal. The path traversed by the diffuse reflected light depends on the tissue optical properties. Thus, the diffuse reflectance spectrum contains the information pertaining to tissue composition. The backscattered light is collected by the fiber optic probe and sent to the detector. Thus NIR diffuse reflectance spectroscopy proves to be a powerful tool in the analysis of human tissue [10].

The basic building blocks of NIR instrument are the radiation source, wavelength selectors, sample presentation facility and detector. NIR radiation sources used can be categorized as thermal and non-thermal. The current study involves the use of tungsten halogen white light source (LS-1) which is a non-thermal radiation source. The tungsten halogen light source is capable of radiating a broad bandwidth light in the spectral range of 350 nm to 2500 nm. The USB 4000 spectrometer with a CCD device serves the purpose of both wavelength selection and detection. The sample presentation is done using a reflectance probe (R400) which is used as both illumination and collection probe. The light impinged on the skin undergoes multiple reflections in the area under study which provides a unique spectral signature. The acquired spectrum is extracted by interfacing the spectrometer to a computer loaded with Spectra Suite software for further analysis [11]. Figure 1 shows the experimental set up for acquiring the diffuse reflectance spectrum.

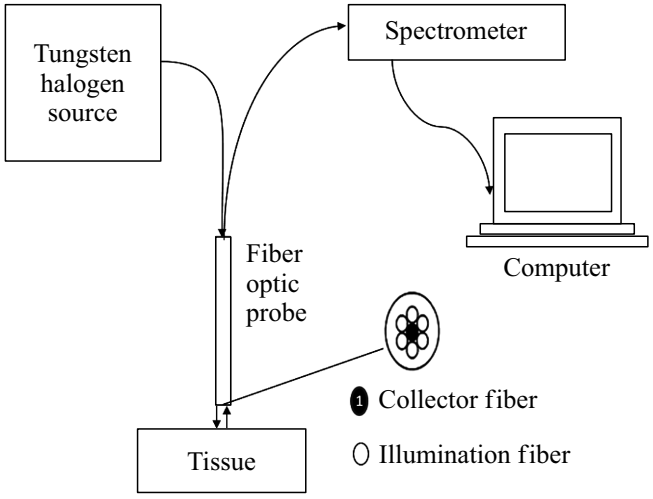


**Figure 1.** The schematic diagram of diffuse reflectance spectroscopy system.

## Interferers in the acquired spectra

The major factors that influence the accuracy of NIR based techniques are sampling, instrument setup, environment factors and modelling. The impact of these factors can be easily minimized by taking precautionary steps during the experimental setup, data acquisition and model development and validation process [12].

The USB4000 spectrometer used in this study has an optical resolution of 1.5 nm and high sensitivity detector. Data acquisition settings like integration time and number of scans can be optimized. As the current study is pertaining to blood glucose prediction in human tissue *in vivo*, the interference due to sampling like sample preparation, sample size, sample packing and packing density does not affect the obtained diffuse reflectance spectrum.

The environmental features like temperature, ambient light, vibration and scanning pressure will affect the precision of the measured blood glucose. The attenuation of light energy at each wavelength is a function of chemical composition of the tissue. Fluctuations in tissue temperature leads to increase in spectral variance thereby making it difficult to extract the blood glucose information [13]. The NIR light is focused on the measurement site for not more than 10 seconds. Thus, the temperature variations that is propagated to the skin tissue is very minimal and does not have a significant impact on the reflected signal. Moreover, the instrument is calibrated with white light standard before starting the measurement process thereby minimizing the effect of ambient conditions. The scanning probe is mounted on a holder which is then placed on the measurement site. This results in uniform pressure and reduced vibration for all measurements.

The most challenging factor in NIR region that causes interference when light interacts with human skin is the light scattering effect. Significant differences in the obtained spectra can be observed due to the complex nature of skin; variations in the thickness and composition of skin within and between individuals and the consequent difficulties with variations in light scattering; overlap of glucose absorption bands with absorption bands from water and other tissue components. Subtle effects such as changes in skin temperature and hydration further complicate the problem [14]. The uncontrollable physical variations namely the pigmentation, inhomogeneous distribution of the particles and changes in refractive index. These factors result in varying the path length thereby introducing additive, multiplicative and wavelength dependent effects. The wavelength dependent scattering effects is more pronounced in longer wavelength region of the spectrum. As the current study is concentrated towards short wavelength region (750 nm to 1100 nm), wavelength dependent scattering has minimal impact. The additive and multiplicative scattering can be minimized by the choice of suitable data pretreatment method and proper mathematical model.

### Spectral Data Acquisition

Blood glucose measurement is performed invasively using clinical measurement procedure for the same subjects and simultaneously spectral data is collected non-invasively using the Diffuse Reflectance Spectrometer. For obtaining an accurate spectral measurement and to ensure the safety of the patient from any infectious diseases, the probe head, and the measurement site is cleansed with an alcohol solution. As the instrument is calibrated in reflectance mode, the acquired spectrum gives the corresponding reflection intensities across the wavelength range of 360 nm to 1100 nm. For the current study, wavelength range of 750 nm to 1040 nm is considered. The analysis for selection of spectral range for glucose detection is carried out in detail and reported in [15]. From the analysis, it is found that the spectral range of 750.1000 nm is found to have the information related to glucose content. It is also found that the informative wavelengths corresponding to loading peaks are 763.64 nm, 970.67 nm, 982.54 nm. A reference dataset has been created by mapping the clinically obtained blood glucose values to their corresponding NIR spectral data which is collected simultaneously from the same subject. The observed NIR spectral information is disturbed by various non- linearity and baseline effects. To overcome these effects, the data are treated with different pre-processing techniques.

### Pre-processing techniques

Insignificant variation is introduced in the acquired spectra due to measurement noise and background interferences. This noise in turn destroys the information on chemical variations in the analyte and increase the complexity of the calibration model. Pre-processing techniques removes the undesired variability in the acquired spectra thereby increasing the signal to noise ratio (SNR). Hence different pre-processing methods are suggested to eliminate the irrelevant variations and background noises from NIR spectrum. Scatter correction methods and spectral derivatives are the two broad categories of pre-processing techniques commonly applied for NIR spectroscopy. Scatter-corrective pre-processing methods includes multiplicative scatter correction (MSC), Inverse multiplicative scatter correction (IMSC), extended multiplicative scatter correction (EMSC), extended inverse multiplicative scatter correction, detrending, standard normal variate (SNV), and normalization. Norris-Williams derivatives and Savitzky-Golay polynomial derivative falls under spectral derivative method [6].

Pre-processing of spectral data comprises of numerous steps each one correcting a particular artifact. So, application of consecutive pre-processing technique helps to improve the SNR and removes the outliers present in the data [16]. Selection of appropriate pre-processing technique should be considered with respect to the calibration model to be developed and is difficult to evaluate prior to model validation. Hence, the identification of suitable pre-processing technique relies on model validation. The selected pre-processing technique should be such that it reduces the model complexity with improved prediction accuracy.

### Reduction of nonlinearity

A simple pre-processing technique for NIR diffuse reflectance spectroscopy that corrects for the non-linearity effects in the measured spectral data is to show a linear relationship between the spectra and the concentration of the constituents as shown in equation (1).

$$A_\lambda = -\log_{10}(R) = \varepsilon_\lambda l c, \tag{1}$$

where $A_\lambda$ is the wavelength dependent absorbance, $R$ is the reflectance, $\varepsilon_\lambda$ is the wavelength dependent molar absorptivity, $l$ is the path length and $c$ is the concentration of the constituent of interest. If the acquired spectral data does not obey Beer Lambert law, then the non-linearity effect might be compensated by increased number of latent variables thus increasing the model complexity [6].

### Multiplicative scatter correction (MSC)

MSC is the most commonly applied pre-processing technique for NIR spectroscopy. MSC compensates for different scatter and particle sizes within the acquired spectra. Every individual spectrum is corrected to have the same scatter level as the reference spectrum which is the mean spectrum of the representative data set [17]. The fit for the individual and the average spectrum is obtained by least square regression.

$$x_i = a_i + b_i \bar{x}_j + e_i, \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, p, \tag{2}$$

where $x_i$ is the individual spectrum, $\bar{x}_j$ is the average spectrum of the data set, $e_i$ is the residual spectrum, $n$ is the number of samples, $p$ is the number of wavelengths, $a_i$ is the offset and $b_i$ is the slope. The corrected spectrum $(X_{i,\mathrm{MSC}})$ is given by,

$$X_{ij,\mathrm{MSC}} = \big((X_{ij} - a_i)|b_i\big), \tag{3}$$

where $X_{ij}$ is the intensity of the $i^{\mathrm{th}}$ spectrum and $j^{\mathrm{th}}$ wavelength of the spectral data. The region which represents the baseline with no chemical information is identified to determine $a_i$ and $b_i$ which is used to determine the corrected spectra. Thus all samples of the MSC corrected spectra appears to have same scatter level.

### Robust MSC

Robust MSC considers the median spectrum as the reference spectrum and the robust least trimmed square regression. The regression model is given by

$$x_{i,j} = \alpha_i + \beta_i(\mathrm{med}(x)_j) + e_{i,j},$$

$$i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, 2, \ldots, p, \tag{4}$$

where $\beta_i$ the slope of the regression model, $\alpha_i$ is the intercept of the regression model, $n$ is the number of samples and $p$ is the number of wavelengths. The slope and the intercept in the regression line for every sample are estimated by the robust least trimmed square (LTS) estimator. The corrected spectral data $x^*_{i,j}$ of the corrected spectrum $x^*_i$ is given by,

$$X^*_{i,j} = ((X_{i,j} - \hat{\alpha}_i)|\hat{\beta}_i), \tag{5}$$

where $\hat{\alpha}, \hat{\beta}$ are robust estimates of the regression coefficient. Thus, robust MSC shifts each spectrum vertically and scales proportionally to make it close to the reference spectrum [5,18].

### Extended multiplicative scatter correction (EMSC)

The EMSC is the extended form of MSC and it includes both polynomial fitting to the reference spectrum, baseline fitting and uses apriori knowledge of the spectra of interest or spectral interferes. The first order polynomial is commonly used for reference correction [19,20].

### Orthogonal signal correction (OSC)

The deviation present in the spectral data that are orthogonal to the response can be eliminated by OSC. OSC also reduces the light scatter effects and other interferences that has zero correlation with the reference value of the response variable. OSC retains all the information pertaining to the response variable. OSC computes the loading weights such that the score vector explains the maximum variance in the data set. Then it is subtracted before computing the new component. Thus OSC requires minimum components for correction. The residuals after OSC are used for model calibration [21].

### Baseline offset

The vertical offset or slope observed in the acquired spectra due to low frequency detector variation leads to baseline effects. Initially, the baseline is estimated, and is then, subtracted from the measured spectrum. This process assists in removing the baseline offsets. Estimation of baseline can be done by detrending or asymmetric least squares smoothing. The derivatives of the input signal also help in eliminating the baseline effects [22,23].

### Standard normal variate (SNV)

In addition to MSC technique, the SNV technique is also commonly adopted for scatter correction for NIR data. The SNV technique eliminates the slope variation in the spectra and incorporates the corrections of the scattering effects. Each individual spectrum is transformed independently in order to eliminate the slope variations in the spectra by applying equation (6), which is given as

$$x_{ij(SNV)} = ((x_{ij} - \overline{x_i})|SD),$$
$$i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, 2, \ldots, p, \tag{6}$$

where $x_{ij(SNV)}$ denote transformed spectra $x_{ij}$ is the original spectra, $\overline{x_i}$ is the mean spectrum of $i^{\text{th}}$ sample, $SD$ is the standard deviation, $n$ denotes the number of samples and $p$ denotes the number of wavelengths. Thus SNV corrects individual spectrum by initially mean centering the spectral data. Thereafter, the centered spectrum is scaled down by the $SD$ of the individual spectral values [6,17].

### Spectral Derivatives

Derivatives removes both additive and multiplicative effects present in the spectra and increases the spectral resolution by removing the background variation. The baseline effects and the linear background variation can be eliminated by I (D1) and II (D2) order derivatives respectively. The II order derivative is the most commonly used technique as it is easier for data interpretation. As the derivatives amplify the noise present in the spectral data proper smoothing has to be performed before finding the spectral derivatives. Commonly used spectral derivatives are Norris-Williams derivation and Savitzky-Golay derivation [6,17].

## Comparative results of pre-processing techniques

This section deals with the investigation of the effectiveness of various empirical pre-processing techniques applied on the VIS-NIR diffuse reflectance spectral data in terms of the ability to predict blood glucose using partial least square regression model. Diffuse reflectance spectral data of 207 subjects including both diabetic and non-diabetic are used in this study. The study involves the spectral data collected from the subjects under fasting and post prandial state. The subjects include both male and female within the age group of 30–70 years. This study deals with the spectral pretreatments like baseline shift, MSC using mean scaling, robust MSC, EMSC, OSC, $(-\log R)$ followed by SNV. All preprocessing techniques are executed using Unscrambler X 10.3.

NIR spectroscopy in biological samples is highly influenced by light scattering due to comparable size of the particle and the wavelength band. Thus, NIR spectroscopy is more prone to non-linearities generated due to scattering effects [6]. Thus, there is a need to apply suitable pre-processing technique to remove these non-linearities. After the suitable pre-processing is done, variation present in the raw data is removed which is the reason for large variations between raw data and pre-processed data.

The variation in acquired raw spectral data and the pre-processed data are analyzed in terms of variance, standard

**Table 2.** Variance, Standard deviation and coefficient of variance in raw data

| Pre-Processing Technique | Mean | Variance | Standard deviation (SD) | CV = SD/Mean |
|---|---|---|---|---|
| Raw data | 26.37 | 519.81 | 22.79 | 0.86 |
| Baseline | 22.68 | 514.18 | 22.67 | 0.99 |
| MSC mean scaling | 0.07 | 0.88 | 0.94 | 13.4 |
| Robust MSC | 0.24 | 1.88 | 1.37 | 5.71 |
| EMSC | 11.87 | 50.40 | 7.09 | 0.59 |
| OSC | 9.93 | 0.17 | 0.41 | 0.04 |
| $(-\log R)$ followed by SNV (Proposed pre-processing technique) | 9.83 | 1.00 | 1.00 | 0.10 |

deviation (SD) and coefficient of variance (CV) and are tabulated in Table 2.

The reason for large variation between raw data and $(-\log R)$ followed by SNV preprocessing technique is that in SNV technique, each spectral data is mean centered and scaled by the standard deviation [6,17]. This results in unity variance, thereby reducing the noise components present in the data and resulting in variations between the raw data and data after pre-processing.

Figure 2 demonstrates the plot of raw and pre-processed spectral data for all the 207 subjects. It is found from Fig. 2, *a*, that the measured reflection intensity is found to have large variations with respect to wavelength. However after applying pre-processing techniques such as baseline correction and EMSC the variations are found to still remain the same as shown in Fig. 2, *b* and 2, *e* respectively. Figure 2, *c, d, f* shows spectral overlapping at some regions, though variability is reduced. This might tend to lower the prediction performance. It can be inferred form Figure 2, *g* that the variability is present in the data is highly reduced and the spectral data is found to be more linear throughout the spectral interval.

## odel complexity evaluation

The spectral data after transformation using the pre-processing techniques mentioned above are used in model development using PLS regression model. The model complexity and prediction accuracy of the various pre-treated data for the PLS regression model is analyzed in this section.

The model complexity mainly depends on the number of latent variables that are needed to develop the calibration model. Too many latent variables result in overfitting while less number of latent variables can cause loss of information [7,24–26]. Hence, the choice of selecting the optimal number of latent variables plays an important role in model development. As a first step, PLS model is applied to the spectral data without applying any preprocessing technique. Then the model complexity is evaluated with respect to the number of latent variables (LVs) that maximizes the variance in the data corresponding to the response (glucose) for various preprocessing techniques listed above. Once the required number of latent variables are determined, blood glucose is predicted using the regression model for all the above mentioned preprocessing techniques. Figure 3 illustrates the scatter plot of the clinically calculated blood glucose concentration against the predicted blood glucose concentration for the various preprocessing techniques.

From Figure 3, it is noted that the scatter plot of the predicted glucose concentration is poor for raw data, baseline corrected and OSC spectra respectively as most of the scatter points doesn.t lie close to the fitted response. An improvement in prediction is observed for MSC mean centered and robust MSC pre-processed spectral data as shown in Fig. 3, *c, d*. The scatter plots lie closer to the fitted response curve for EMSC as shown in Fig. 3, *e*. A more linear relationship is exhibited for $(-\log R)$ followed by SNV technique as most of the scatter plots lie on or very close to the fitted response as shown in Fig. 3, *g*. This concludes that $(-\log R)$ followed by SNV technique results in a good fit.

## Model efficacy estimation

The prediction accuracy is further evaluated using 35 fold cross validation and the estimated mean square prediction error (MSPECV) is determined. The number of latent variables required for model development without preprocessing and by the application of various preprocessing techniques and their corresponding range of MSPECV are summarized in Table 3.

From Table 3, it is evident that the number of latent variables required to build the calibration model is greater for raw data, baseline and OSC and is less for other pre-processing techniques. MSC mean scaling and $(-\log R)$ followed by SNV pre-processing methods produces 28 latent variables. Thus it is evident from the above table that linearization of the spectral data prior to the application of pre-processing technique tends to reduce the model complexity. Also the minimum estimated mean square prediction error is 0.18 mg/dl for $(-\log R)$ followed by SNV preprocessing technique for 36 latent variables. Moreover, it is observed that MSC mean centering and SNV techniques behave similarly with minimum difference in their prediction ability. Figure 4 shows the estimated mean square prediction error plotted across the number of folds for various preprocessing techniques. It is clear from

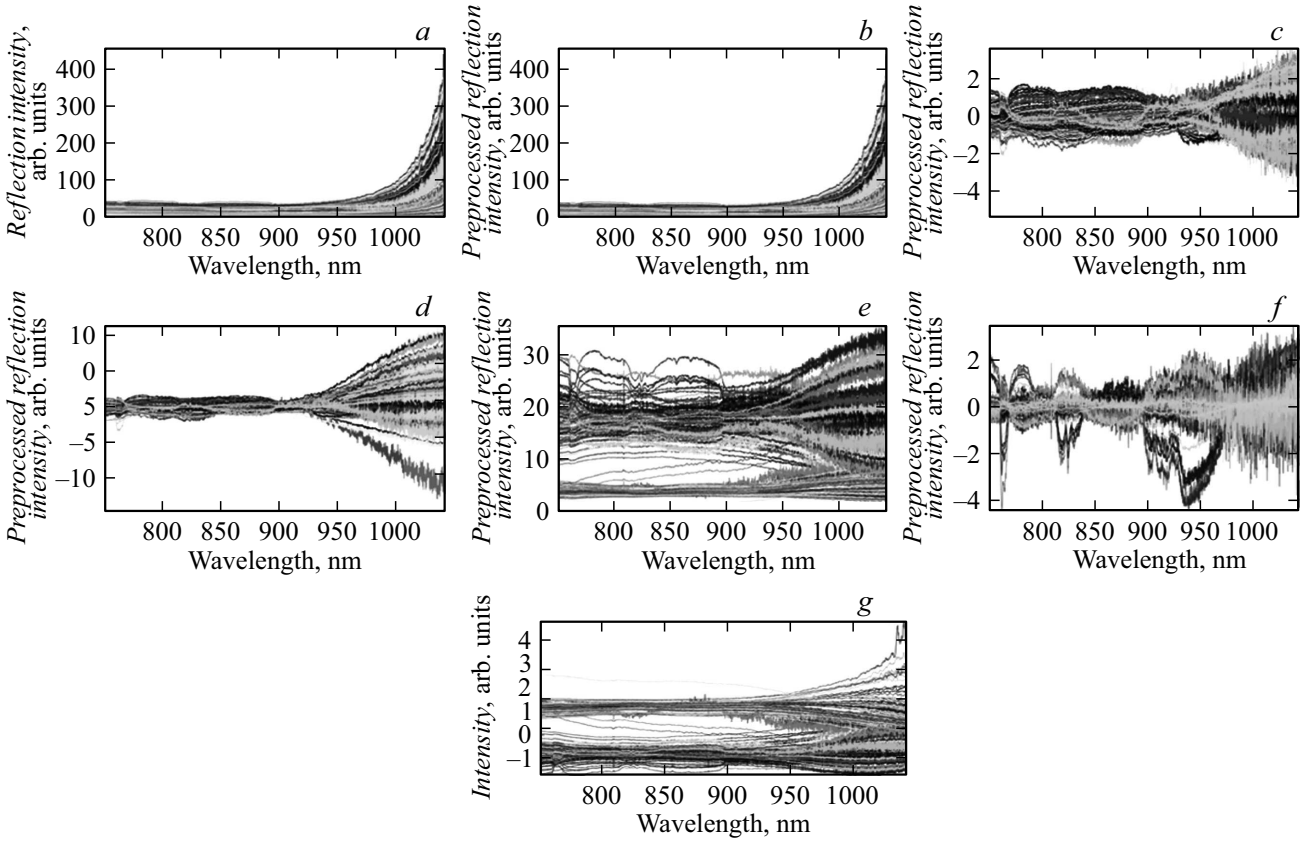


**Figure 2.** Raw and transformed spectra using different pre-processing techniques: (*a*) Raw spectra, (*b*) baseline correction, (*c*) MSC mean scaling, (*d*) robust MSC, (*e*) EMSC, (*f*) OSC, (*g*) $(-\log R)$ followed by SNV.

**Table 3.** Statistical parameters for PLSR model with various pre-processing techniques

| Pre-processing technique | Number of latent variables | Range of mean square prediction error for cross validation (mg/dl) |
|---|---|---|
| Raw data | 58 | [1.33.5.29] |
| Baseline | 56 | [1.10.4.88] |
| MSC mean scaling | 28 | [0.40.5.29] |
| Robust MSC | 30 | [0.31.5.29] |
| EMSC | 31 | [0.34.4.88] |
| OSC | 54 | [0.55.4.88] |
| $(-\log R)$ followed by SNV (Proposed pre-processing technique) | 28 | [0.18.4.88] |

Figure 4 that the mean square prediction error tends to decrease with increase in the number of folds.

It is also inferred form Fig. 4 that the mean square prediction error is minimum for $(-\log R)$ followed by SNV pre-processing technique whereas for the raw data, error is maximum. Thus, it is concluded form this study that, the proposed pre-processing technique i. e. $(-\log R)$ followed by SNV technique is the best pre-processing technique amongst the various other techniques discussed in this paper for the PLSR model development in the prediction of blood glucose.

## Discussion

Oliver Devos et al [27] in his study has proposed the parallel genetic algorithm co-optimization technique to simultaneously optimize spectral pre-processing and variable selection using PLSR model. Three near infrared spectroscopic data sets namely corn data, tecator data and sugar beet data are subjected to 31 pre-processing functions. The authors have proved that genetic algorithm performs better for simultaneous optimization with improved prediction ability. Yiming Bi et al [28] has proposed a localized version of SNV pre-processing technique. The prediction ability was compared with three different data sets namely meat data, pharmaceutical data and wheat data. The data sets were subjected to pre-processing techniques like SNV, MSC, Extended Inverted Signal Correction (EISC), EMSC, first order derivative (D1), second order derivative (D2), localized SNV (LSNV) and proved that LSNV gives good prediction with minimum RMSECV. S.N. Thennadil, E.B. Martin [29] in their study analyzed the performance of empirical preprocessing methods to extract information

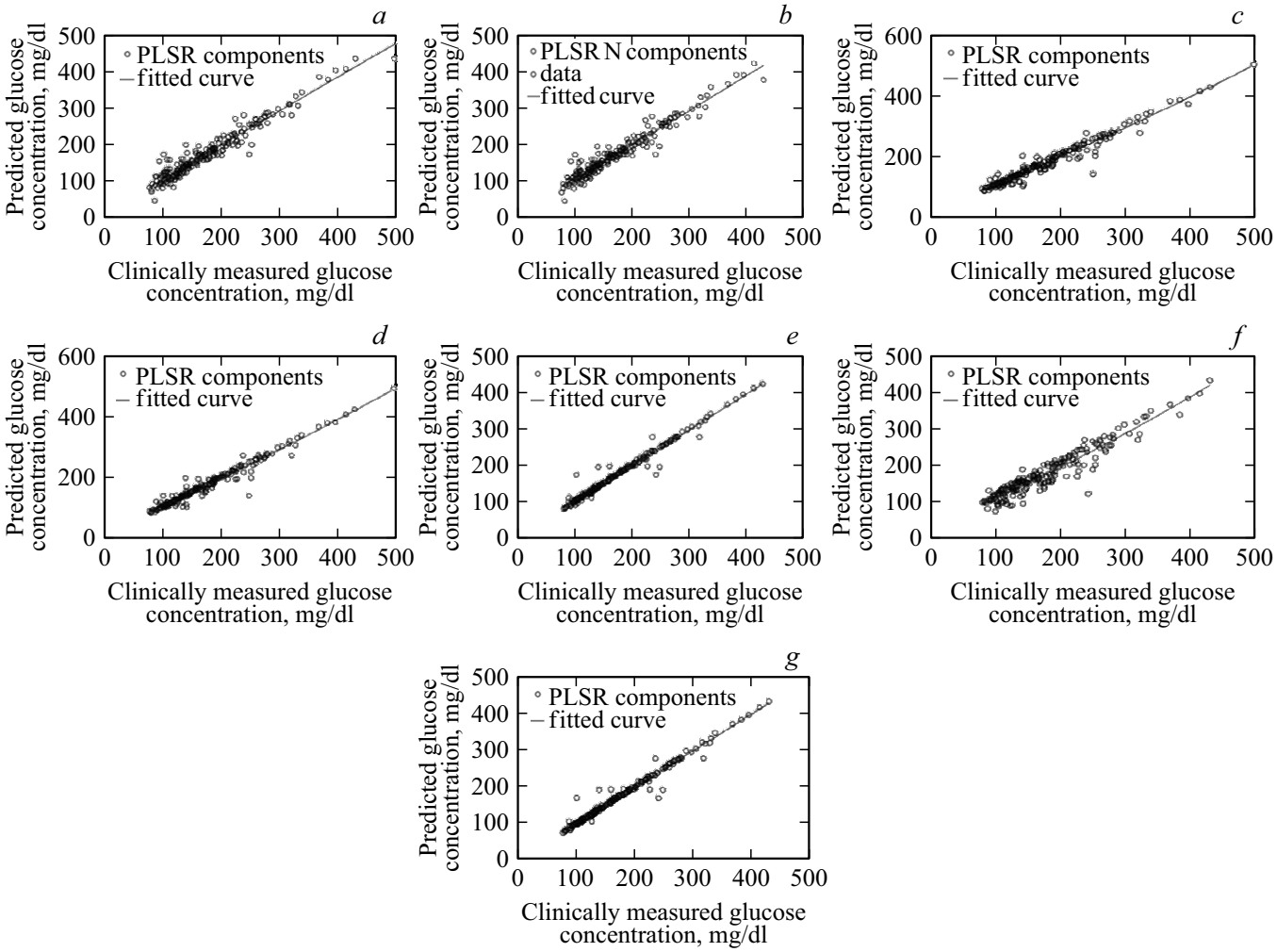


**Figure 3.** Scatter plot of predicted glucose Vs clinically measured glucose for raw and transformed spectra using different pre-processing techniques: (*a*) Raw spectra, (*b*) baseline, (*c*) MSC mean scaling, (*d*) robust MSC, (*e*) EMSC, (*f*) OSC, (*g*) $(-\log R)$ followed by SNV.

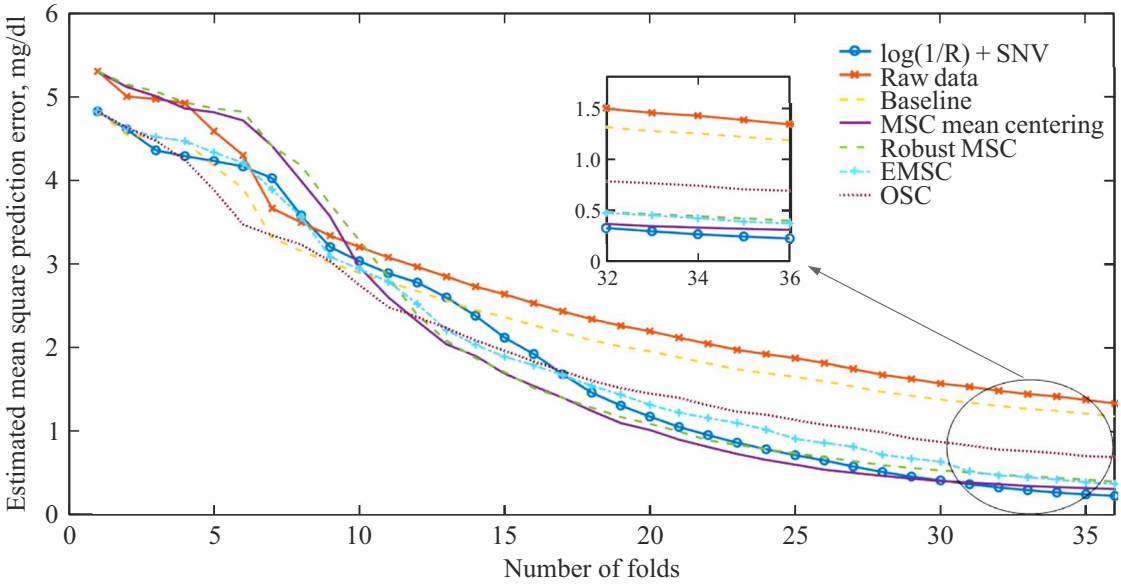


**Figure 4.** Comparative prediction analysis of PLSR models for various data pre-processing methods.

about chemical species in a system. They inferred that EMSCL, a form of EMSC that uses log term for wavelength dependencies was found to be statistically significant than other methods. The proposed study focusses on the performance of various pre-processing methods by implementing PLSR model for predicting the blood glucose. It is found that the proposed pre-processing technique based on logarithmic linearization of data subjected to classical SNV technique performs better than other techniques with minimum mean square error for prediction.

## Conclusion

The current research work presents various empirical pre-processing methods that are suitable for NIR spectroscopic study in order to predict the blood glucose concentration using PLSR calibration model. This paper has shown that the complicated NIR spectra can yield good prediction results with suitable pre-processing technique. It also emphasises on the importance of pre-processing techniques in NIR study.

The suggested pre-processing technique of linearizing the spectral data $(-\log R)$ prior to SNV aided in a PLSR model with better prediction of blood glucose from the data acquired using the diffuse reflectance spectroscopic setup discussed in this paper. Although the current research work provides promising results for the proposed pre-processing technique, the choice of adopting an optimal pre-processing technique is challenging as it depends on the instrumental setup, data acquisition and application based prediction model development.

### Acknowledgments

This research work has been carried out with the dataset that has been collected from Hariharan Institute of Diabetic and Research Centre, Nanganallur, Chennai, Tamil Nadu. The NIR spectrometer used for the data acquisition is procured through the Funding obtained from SRM Institute of Science and Technology.

## References

[1] I.L. Jernelv, K. Milenko, S.S. Fuglerud, D.R. Hjelme, R. Ellingsen, A. Aksnes. Appl. Spectr. Rev., **54** (7), 543–572 (2019).
[2] C.F. So, K.S. Choi, T.K. Wong, J.W. Chung. Medical Devices: Evidence and Research, **5**, 45–52 (2012).
[3] Lu Xu, Yan-Ping Zhou, Li-Juan Tang, Hai-Long Wu, Jian-Hui Jiang, Guo-Li Shen, Ru-Qin Yu. Anal. Chim. Acta, **616** (2), 138–143 (2008).
[4] O. CDevos, G. Downey, L. Duponchel. Food Chemistry, **148**, 124–130 (2014).
[5] S. Verbovena, M. Hubertb, P. Goos. J. Chemometrics, **26** (6), 282–289 (2012).
[6] A. Rinnan, F. van den Berg, S.B. Engelsen. Trends in Anal. Chem., **28** (10), 1201–1222 (2009).
[7] M. Goodarzi, S. Sharma, H. Ramon, W. Saeys. TrAC Trends in Anal. Chem., **67**, 147–158 (2015).
[8] R.J. McNichols, G.L. Cote, J. Biomed. Optics, **5** (1), 5–17 (2000).
[9] E. Hecht. Optics, 4th ed (Addison-Wesley, San Francisco, California, 2002).
[10] V.V. Tuchin. Handbook of Optical Sensing of Glucose in Biological Fluids and Tissues (CRC Press, Taylor & Francis Group, London, 2009).
[11] S. Prince, S. Malarvizhi. IFMBE Proceedings, **25** (7), 240–243 (2009).
[12] https://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/
[13] M.R. Makarewicz, M. Mattu, T.B. Blank, S.L. Monfre, T.L. Ruchti, inventors. Sensys Medical Inc., assignee. United States patent US 6, 640, 117 (2003).
[14] M. Jackson, G. Wagnieres, H.H. Mantsch. *Encyclopedia of Spectroscopy and Spectrophotometry* (2017).
[15] S.V. Suryakala, S. Prince. Optical and Quantum Electronics, **51** (8), 271 (2019).
[16] J. Engel, J. Gerretzen, E. Szymanska, J.J. Jansen, G. Downey, L. Blanchet, M.C. Lutgarde Buydens. Trends in Anal. Chem., **50**, 96–106 (2013).
[17] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, J. Pharmaceutical and Biomedical Analysis, **21**, 115–132 (1999).
[18] P.J. Rousseeuw, J. Am. Statistical Association, **79**, 871–880 (1984).
[19] H. Martens, E. Stark, J. Pharmaceutical and Biomedical Analysis, **9**(8), 625–635 (1991).
[20] Zeng-Ping Chen, Julian Morris, Elaine Martin. Analytical Chemistry, **78** (22), 7674–7681 (2006).
[21] C. Pizarro, I. Esteban-Diez, A.-J. Nistal, J.-M. Gonzalez-Saiz. Anal. Chim. Acta, **509**, 217–227 (2004).
[22] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, J. Chromatography A, **1255**, 3–11 (2012).
[23] J. Engel, J. Gerretzen, E. Szymanska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens. Trends in Anal. Chem., **50**, 96–106 (2013).
[24] N.M. Faber. J. Chemometrics, **13**, 185–192, (1999).
[25] O.E. De Noord. Chemometrics and Intelligent Laboratory Systems, **23**, 65–70 (1994).
[26] T. Naes, H. Martens. J. Chemometrics, **2**, 155–167 (1988).
[27] O. Devos, L. Duponchel. Chemometrics and Intelligent Laboratory Systems, **107**, 50–58 (2011).
[28] Y. Bi, Kailong Yuan, Weiqiang Xiao, Jizhong Wu, Chunyun Shi, Jun Xia, Guohai Chu, Guangxin Zhang, Guojun Zhou. Anal. Chim. Acta, **909**, 30–40 (2016).
[29] S.N. Thennadil, E.B. Martin. J. Chemometrics, **19**, 77–89 (2005).