---

**Article:**
Khadem, H. orcid.org/0000-0002-6878-875X, Eissa, M.R., Nemat, H. et al. (2 more authors) (2020) Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy. Talanta, 211. 120740. ISSN 0039-9140

https://doi.org/10.1016/j.talanta.2020.120740

---

# Classification before Regression for Improving the Accuracy of Glucose Quantification using Absorption Spectroscopy

Heydar Khadem [a,*], Mohammad R. Eissa [a], Hoda Nemat [a], Osamah Alrezj [a], and Mohammed Benaissa [a]

E-mail addresses: h.khdem@sheffield.au.uk, m.eissa@sheffield.ac.uk, hoda.nemat@sheffield.ac.uk, oamousa1@sheffield.ac.uk, m.benaissa@sheffield.ac.uk

[a] Department of Electronic and Electrical Engineering, University of Sheffield, S1 4DE, UK.
* Corresponding author

## Abstract

This work contributes to the improvement of glucose quantification using near-infrared (NIR), mid-infrared (MIR), and combination of NIR and MIR absorbance spectroscopy by classifying the spectral data prior to the application of regression models. Both manual and automated classification are presented based on three homogeneous classes defined following the clinical definition of the glycaemic ranges (hypoglycaemia, euglycaemia, and hyperglycaemia). For the manual classification, partial least squares and principal component regressions are applied to each class separately and shown to lead to improved quantification results compared to when applying the same regression models for the whole dataset. For the automatic classification, linear discriminant analysis coupled with principal component analysis is deployed, and regressions are applied to each class separately. The results obtained are shown to outperform those of regressions for the entire dataset.

**Keywords**: Glucose; Non-invasive; Near-infrared; Mid-infrared; Spectroscopy

## 1. Introduction

The importance of the development of non-invasive glucose monitoring in diabetes management has spurred research into the quantification of glucose through in vivo and in vitro experiments [1], [2]. The underlying modalities pursued in these studies can be listed as; near-infrared (NIR), mid-infrared (MIR), Raman and bio-impedance spectroscopy, electromagnetic sensing, fluorescence technology, optical coherence tomography, optical polarimetry, reverse iontophoresis, and ultrasound technology [3]. Of all the techniques mentioned above, NIR and MIR spectroscopy are promising and commonly used methods [4], [5].

NIR and MIR spectroscopy use light beams in the wavelength range of 750–2500nm (13333–4000$cm^{-1}$) and 2500–10000nm (4000–1000$cm^{-1}$), respectively [6]. These technologies are not expensive for frequent measurements as it does not use any specific reagent [7]. One advantage of the MIR method is decreased scattering phenomena and increased absorption because of higher wavelengths compared with NIR spectroscopy [8]. Moreover, the peaks of glucose are sharper in the MIR region [9]. NIR light, on the other

hand, possesses a deep penetration length [10] and could traverse through different skin layers to reach the subcutaneous area [11].

While travelling through a sample, some frequencies of NIR/MIR light are absorbed and scattered because of the interaction with the physiological compounds of the skin [12]. These absorption and scattering measures are used for the quantification analysis of the glucose or other chromophores in the sample [6].

For extracting glucose-related information from the NIR and MIR spectra, multivariate calibration methods such as partial least squares regression (PLSR), principal component regression (PCR), multiple linear regression, artificial neural networks, and support vector machine regression are typically applied to the recorded signals [13]. For improving the accuracy of the analyses, many pre-processing methods have also been proposed [14], such as multivariate scatter correction (MSC), smoothing, and digital band-pass filtering. However, accurate quantification results remain a challenge [15].

This paper proposes a classification-before-regression methodology to improve glucose measurement using NIR, MIR and a combination of NIR and MIR (hereafter referred to as NIR-NIR) spectroscopy. Both manual and automatic classification are carried out by classifying the dataset into three more homogeneous groups following the clinical definition of the glycaemic ranges (hypoglycaemia, euglycaemia, and hyperglycaemia). Partial least squares and principal component regressions are applied with the manual classification; for the automatic classification, linear discriminant analysis coupled with principal component analysis in both cases are deployed, and regressions are created for each class. The results obtained using the same data for both cases are shown to outperform the results obtained when no classification-before-regression is used.

Classification of spectral data before regressions has been previously used in other research areas, such as rapid analysis of coal properties using NIR spectroscopy [16], [17]. However, this is the first paper to our knowledge that correlates spectral data using a pre-classification approach in order to improve the accuracy of glucose measurement.

## 2. Data collection

### 2.1. Sample preparation

Two aqueous solutions (A and B), both with a volume of 0.5 litres, were prepared. Solution A contained glucose, human serum albumin, and phosphate with concentrations of 500mg/dl, 5g/dl, 0.01M/dl, respectively, and had a pH of 7.4. Solution B had the same properties as solution A, but without glucose.

5 ml solution A was removed and stored in a tube as the first sample. The removed amount from solution A was then replaced with 5 ml solution B so that the glucose concentration of solution A is lowered to 495mg/dl. Then, 5 ml solution A was again collected as the second sample, and it was replaced again with 5 ml solution B. Repeating the same steps, 100 samples were obtained with the glucose concentrations ranging from 5mg/dl to 500mg/dl, at intervals of 5mg/dl. The samples were prepared in the laboratories of the Department of Chemistry, University of Sheffield, Sheffield, UK.

### 2.2. Spectra acquisition

Spectra collection was performed under uncontrolled environmental conditions in the laboratories of the Department of Materials Science and Engineering, University of Sheffield, Sheffield, UK. A Fourier

transform infrared (FTIR) spectrometer (PerkinElmer Inc., USA) was utilised to collect the absorption spectra of the samples by attenuated total reflection technique.

The spectrometer sensing lens was cleaned using ethanol before placing each sample to avoid inaccurate readings. A few drops of the sample were then added so that it covered the whole lens surface. For obtaining more accurate spectral data, the spectrum of each sample was constructed by averaging four scans of the spectrometer readings [18]. The recorded spectra covered wavelength range of 2100–8000nm (4761–1250cm$^{-1}$) with a resolution of 1.7nm. Wavelengths from 2100 to 2500nm (4761–4000cm$^{-1}$) of the collected spectra belonged to NIR region, and from 2500 to 8000nm (4000–1250cm$^{-1}$) to MIR. Fig. 1 shows all the raw spectra observed.

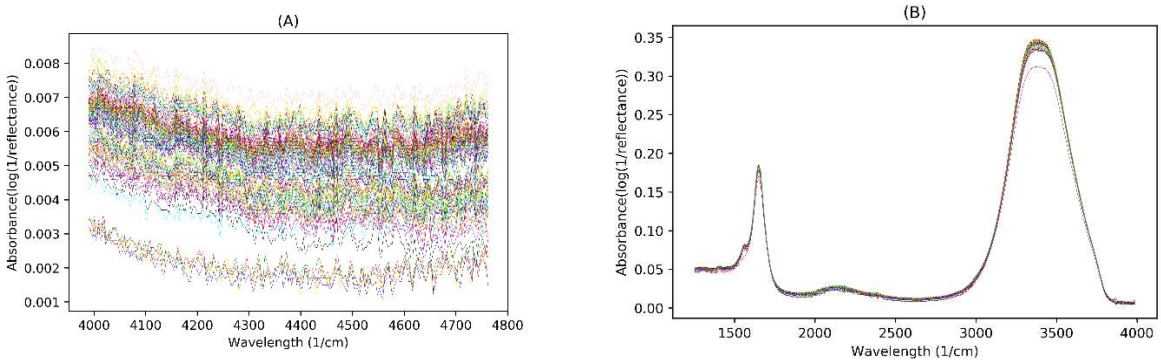


*Fig. 1. Original spectra collected from chemical samples (A) NIR spectra, (B) MIR spectra*

# 3. Methods

## 3.1. Quantification methods

As shown in Fig. 2, three methods were developed in this work for glucose quantification from the collected NIR, MIR, and NIR-NIR spectral data. For data analysis in these quantification methods, we used Python (3.6.7), scikit-learn (0.15.2), and SciPy (0.12.0).

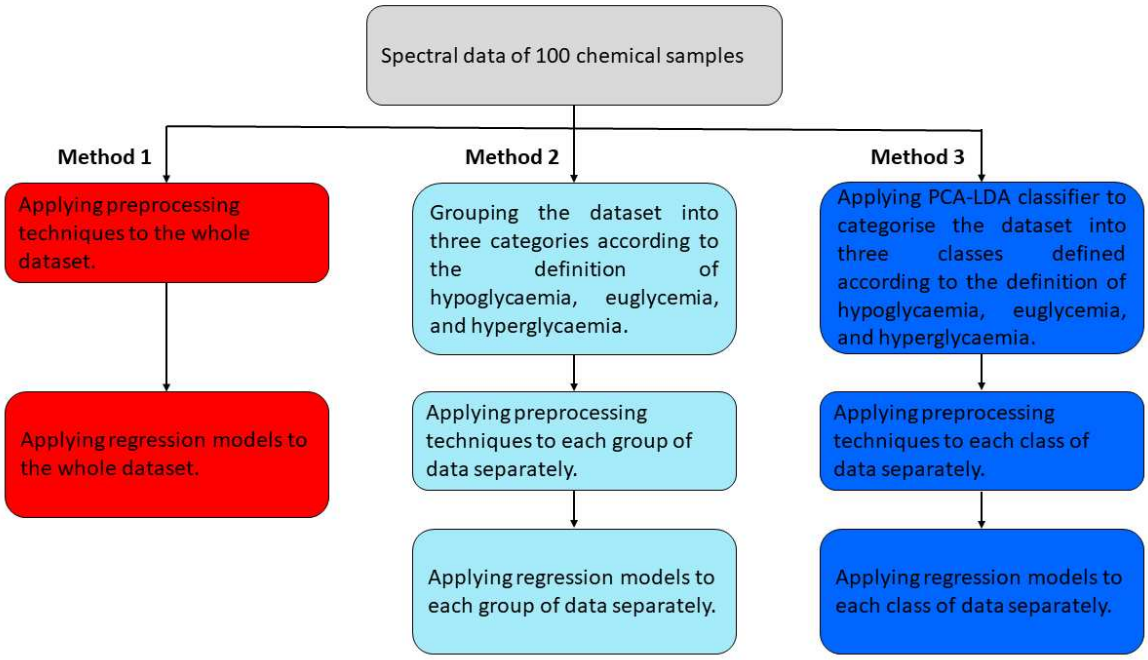


*Fig. 2. Quantification methods applied in this paper for glucose measurement using NIR, MIR, and NIR-MIR spectroscopy*

### Method 1

In this method, regression models were implemented for the raw and pre-processed spectra of the whole sample set to predict the relevant glucose concentrations.

### Method 2

In this method, the spectral data were divided into the three groups shown in Table 1. Pre-processing and regression methods were then implemented for each group individually. By performing this method, the idea was to investigate the effect of categorising the dataset as being in the hypoglycaemic range (≤70mg/dl), euglycaemic range (70-180mg/dl), and hyperglycaemic range (180mg/dl≤) [19] on the accuracy of the measurements.

As shown later in the paper, Method 2 improved the measurement results as compared to Method 1. This improvement led us to proceed to Method 3, which as Method 1 used only the spectral data without further information (Method 2 used the labels in addition to the spectral data).

*Table 1. Division of the dataset into three groups following the clinical definition of the glycaemic ranges*

| | Label | | |
|---|---|---|---|
| | Class 1 | Class 2 | Class 3 |
| Glucose concentration range | 5–70 mg/dl | 75–180 mg/dl | 185–500 mg/dl |
| Corresponding glycaemic range | Hypoglycaemia | Euglycaemia | Hyperglycaemia |
| Quantity of data in the class | 14 | 22 | 64 |

### Method 3

This method had the same principles as Method 2 but automated the grouping process using PCA-LDA classifier.

### 3.2. Pre-processing methods

In spectroscopic analyses, pre-processing techniques are generally applied to the raw spectra to curtail adverse effects from elements other than the analyte of interest and environmental conditions [20]. In this work, three pre-processing techniques applied to assess the effectiveness of the proposed pre-classification approach when it is coupled with conventional pre-processing methods.

#### Smoothing (S)

Savitzky-Golay smoothing is a pre-processing method to diminish the effect of noise on the raw spectra [17]. The method is an averaging algorithm that fits a polynomial with successive subsets of adjacent data points based on the least-squares [21]. For the Savitzky-Golay filter in this work, a five-point window and a second-order polynomial were implemented.

#### Multivariate scatter correction (MSC)

The scattering phenomenon is the most significant obstacle when attempting quantitative measurements using NIR spectroscopy [22]. MSC is a useful pre-processing technique to eliminate the effect of light scattering [23]. In this method, the scattering of each spectrum is estimated relative to a reference spectrum [14]. In this work, the average of the all signals in the calibration set was considered as the reference spectrum; each signal was then adjusted using the reference so that they all had the same scatter level [17].

#### Smoothing coupled with MSC (S-MSC)

Applying different pre-processing techniques together is a common approach to deal with spectroscopic data [15]. In this paper, a combination of the smoothing and MSC methods was also applied as a third pre-processing technique [24].

### 3.3. Regression methods

For constructing predictive models for selected analytes such as glucose, multivariate calibration methods are applied to spectroscopic data [12], [25]. Linear models such as PLSR and PCR are generally preferred since they are easy to apply and amenable to Physico-chemical interpretation [26]. Likewise, in this paper, PLSR and PCR are selected for glucose quantification using absorbance spectroscopy.

For optimising the PLS and PCR components quantity, different numbers ranging from 1 to 10 were examined; and each time, the sum of squares of differences between reference and predicted glucose levels, based on ten-fold cross-validation analysis, was calculated to form the predicted residual sum of squares (*PRESS*). The model minimising the value of $PRESS/(N-A-1)$ is then selected; where $N$ is the size of the calibration set and $A$ is the number of components [27].

### 3.4. Classification method

Principal component analysis (PCA) reduces the dimensionality of data while retaining most of the information present in the dataset [28]. Linear discriminant analysis (LDA) is a technique that maximises the variance between groups while minimising the variance within groups based on the determination of linear discriminant functions [29]. In this work, PCA coupled with LDA (PCA-LDA) is employed to classify the dataset, a method which was shown to be useful in this regard [30]. Different values from 1 to

10 were considered as the number of PCA components and the model resulting in the best classification accuracy, based on ten-fold cross-validation results, was then selected [17].

### 3.5. Evaluation method

The maximisation of the training data size is a basic approach for dealing with small datasets, and cross-validation is suitable for this purpose [31]. Ten-folds cross-validation was applied in this work to evaluate the regression and classification models [32].

### 3.6. Evaluation Metrics

**Root mean square error of prediction (RMSEP)**

In this work, RMSE was calculated as follows to measure the actual error of quantifications [33].

$$RMSEP = \sqrt{(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2) / N}$$

$N$ : the size of the calibration set

$y_i$ : reference value

$\hat{y}_i$ : predicted value

**Percentage error around the mean (PEM)**

PEM was used to analyse the performance of the quantification methods for each class of data [17] ("quantification methods" and "classes" are discussed in section 3.2).

$$PEM = (RMSECV / \bar{Y}) \times 100$$

$\bar{Y}$ : the average of reference values

**Correlation coefficients (r)**

r is a statistical measure indicating correlations between the reference and predicted glucose concentrations [34].

$$r = Cov\ (Y,\ \hat{Y}) / \sigma_Y \sigma_{\hat{Y}}$$

$Y$ : reference values

$\hat{Y}$ : predicted values

$\mathrm{Cov}(\mathrm{Y}, \hat{\mathrm{Y}})$ : covariance between $\mathrm{Y}$ and $\hat{Y}$

$\sigma_{\mathrm{Y}}$ : standard deviation of $\mathrm{Y}$

$\sigma_{\hat{\mathrm{Y}}}$ : standard deviation of $\hat{Y}$

**Clarke error grid analysis (EGA)**

EGA considers the relative difference between reference and predicted glucose level and the clinical significance of this difference [35]. In this paper, EGA was performed to assess the clinical accuracy of the measurements, a method which can be used to evaluate in vitro quantitative analysis of glucose[36].

# 4. Results

## 4.1. Classification results

It was mentioned earlier that the proper number of PCA components in the PCA-LDA classifier, implemented in the third quantification method, was chosen based on the examination of varying values. Classification results based on ten-fold cross-validation for a different number of PCA components ranging from 1 to 10 is illustrated in Fig. 3. As the figure shows, in the NIR region, the classification accuracy improved significantly when the number of components rose from 1 to 5 but remained steady afterwards. Therefore, we set the number of PCA component at 5 for this region. Similarly, the number of PCA elements for when using MIR and IR spectral data were both set at 4.

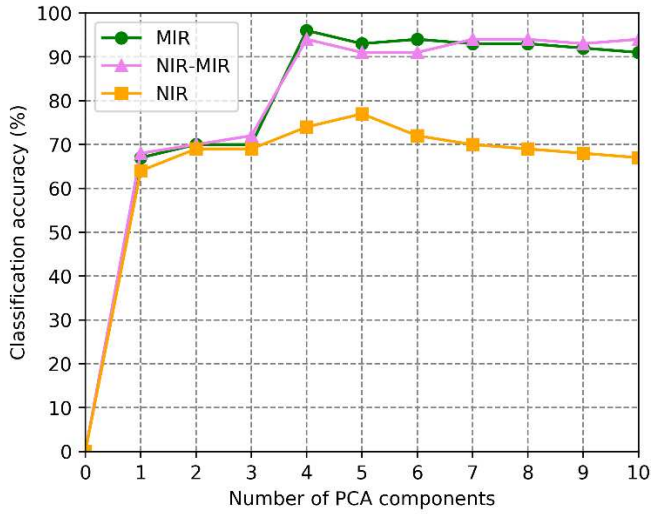


*Fig. 3. The Accuracy of the PCA-LDA classifier for different number of PCA components*

The detailed classification results after setting the number of PCA component at the values mentioned above are presented in Table 2. Overall, the best and the lowest classification accuracy were found when using the MIR and NIR spectral data, respectively.

*Table 2. The PCA-LDA classification results based on ten-fold cross-validation in different spectral regions*

| Spectral region | Class 1 data (Hypoglycaemia range) | | Class 2 data (Euglycaemia range) | | Class 3 data (Hyperglycaemia range) | | All data together (whole glycaemic range) | |
|---|---|---|---|---|---|---|---|---|
| | No. of Errors | Classification Accuracy (%) | No. of Errors | Classification Accuracy (%) | No. of Errors | Classification Accuracy (%) | No. of Errors | Classification Accuracy (%) |
| NIR | 3 | 78.5 | 15 | 31.8 | 5 | 92.1 | 23 | 77 |
| MIR | 0 | 100 | 4 | 81.8 | 0 | 100 | 4 | 96 |
| NIR-MIR | 0 | 100 | 4 | 81.8 | 2 | 96.8 | 6 | 94 |

## 4.2. Quantification results

This section is partitioned into three parts, each of which reports the quantification results belonged to the analyses in one of the spectral regions (NIR, MIR, and NIR-MIR region). In this way, the capability of the proposed method to improve the analysis precision in either of the three spectral regions could be shown more effectively.

### 4.2.1. Quantification results in the NIR region

Table 3 lists the results of RMSE, PEM and correlation coefficient (r) of the three quantification methods in the NIR region, and in addition, the improvement of RMSE for Methods 2 and 3 compared to Method 1. The values in bold indicate the best results of each quantification method based on the lowest RMSE of the whole dataset; these results are considered for EGA analysis, as discussed later. For comparison purposes, the results of each class and also for that of all classes together are presented separately.

Methods 2 and 3 possessed smaller calibration sets than Method 1, a characteristic that might have harmed the results of these methods. However, they provided more accurate quantification results than Method 1. The accuracy of predictions obtained by Method 2 outweighed those of Method 3. The reason is that the weak classification accuracy in the NIR region (discussed in section 4.1) negatively affected the measurements of Method 3. For Method 2, the improvements in the quantification results in comparison to Method 1 were more pronounced for data in Classes 1 and 2; for Method 3, it happened for data in Class 1.

*Table 3. Results of ten-fold cross-validation for the quantification methods in the NIR region*

| PM | RM | QM | Class 1 data (Hypoglycaemia range) | | | Class 2 data (Euglycaemia range) | | | Class 3 data (Hyperglycaemia range) | | | All data together (Whole glycaemic range) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | r |
| NP | PLSR | 1 | 63.5 | — | 181.4 | 100.0 | — | 80.0 | 76.6 | — | 22.3 | 81.1 | — | 0.82 |
| | | 2 | 11.8 | +81.4 | 31.5 | 19.4 | +80.6 | 15.2 | 53.2 | +30.5 | 15.5 | 43.7 | +46.1 | 0.95 |
| | | 3 | 39.5 | +37.7 | 113.4 | 97.0 | +3.0 | 77.6 | 69.7 | +9.0 | 20.3 | 74.0 | +8.7 | 0.85 |
| | PCR | 1 | 64.3 | — | 183.9 | 98.4 | — | 78.7 | 77.7 | — | 22.6 | 81.4 | — | 0.82 |
| | | **2** | **12.4** | **+80.7** | **33.0** | **19.3** | **+80.3** | **15.1** | **53.0** | **+31.7** | **15.4** | **43.6** | **+46.4** | **0.95** |
| | | 3 | 42.4 | +34.0 | 121.3 | 98.1 | +0.3 | 78.5 | 69.2 | +10.9 | 20.2 | 74.2 | +8.8 | 0.85 |
| S | PLSR | 1 | 60.5 | — | 173.0 | 100.8 | — | 80.7 | 76.3 | — | 22.2 | 80.9 | — | 0.82 |
| | | 2 | 11.9 | +80.3 | 31.8 | 19.4 | +80.7 | 15.2 | 53.0 | +30.5 | 15.4 | 43.6 | +46.1 | 0.95 |
| | | **3** | **39.1** | **+35.3** | **11.9** | **97.1** | **+3.6** | **77.6** | **68.8** | **+9.8** | **20.0** | **73.4** | **+9.2** | **0.83** |
| | PCR | **1** | **60.8** | **—** | **173.9** | **97.9** | **—** | **78.3** | **76.3** | **—** | **22.2** | **80.0** | **—** | **0.83** |
| | | 2 | 12.2 | +79.9 | 32.6 | 19.3 | +80.2 | 15.1 | 52.8 | +30.7 | 15.4 | 43.4 | 45.7 | 0.95 |
| | | 3 | 42.1 | +30.7 | 120.4 | 98.1 | -0.2 | 78.5 | 68.7 | +9.9 | 20.6 | 73.9 | +7.6 | 0.85 |
| MSC | PLSR | 1 | 63.5 | — | 181.4 | 100.0 | — | 80.0 | 76.6 | — | 22.3 | 81.1 | — | 0.82 |
| | | 2 | 11.8 | +81.4 | 31.5 | 19.4 | +80.6 | 15.2 | 53.2 | +30.5 | 15.5 | 43.7 | +46.1 | 0.95 |
| | | 3 | 39.5 | +37.7 | 113.0 | 97.0 | +3.0 | 77.6 | 69.7 | +9.0 | 20.3 | 74.0 | +8.7 | 0.85 |
| | PCR | 1 | 64.3 | — | 183.9 | 98.4 | — | 78.7 | 77.7 | — | 22.6 | 81.4 | — | 0.82 |
| | | 2 | 12.4 | +80.7 | 33.0 | 19.3 | +80.3 | 15.1 | 53.0 | +31.7 | 15.4 | 43.6 | +46.4 | 0.95 |
| | | 3 | 42.4 | +34.0 | 121.3 | 98.1 | +0.3 | 78.5 | 69.2 | +10.9 | 20.3 | 74.2 | +8.8 | 0.85 |
| S-MSC | PLSR | 1 | 175.0 | — | 500.0 | 136.2 | — | 109.0 | 95.4 | — | 27.8 | 118.6 | — | 0.57 |
| | | 2 | 12.1 | +93.0 | 32.3 | 32.6 | +76.0 | 25.6 | 56.2 | +41.0 | 16.4 | 47.7 | +59.7 | 0.94 |
| | | 3 | 79.3 | +54.6 | 226.6 | 115.4 | +15.2 | 92.3 | 71.0 | +25.5 | 20.7 | 84.3 | +28.9 | 0.81 |
| | PCR | 1 | 179.5 | — | 512.9 | 137.7 | — | 110.2 | 96.2 | — | 28.1 | 120.3 | — | 0.55 |
| | | 2 | 11.3 | +93.7 | 30.2 | 38.8 | +71.8 | 30.4 | 55.7 | +42.0 | 16.2 | 48.3 | +59.8 | 0.94 |
| | | 3 | 78.1 | +56.4 | 223.3 | 114.8 | +16.6 | 91.8 | 71.0 | +26.1 | 20.7 | 83.9 | +30.2 | 0.81 |

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

As mentioned earlier, EGA was performed to evaluate the accuracy of the measurements further. The EGA comparison between the best prediction results of the three quantification methods in the NIR region (results in bold in Table 3) is shown in Fig. 4(A); and the percentage of predictions located in Zone A—the most clinically desired measurement—is presented in Fig. 4(B). As shown in the figures, using Methods 2 and 3, a higher ratio of predictions located in zone A, especially for data in Classes 1 and 2.

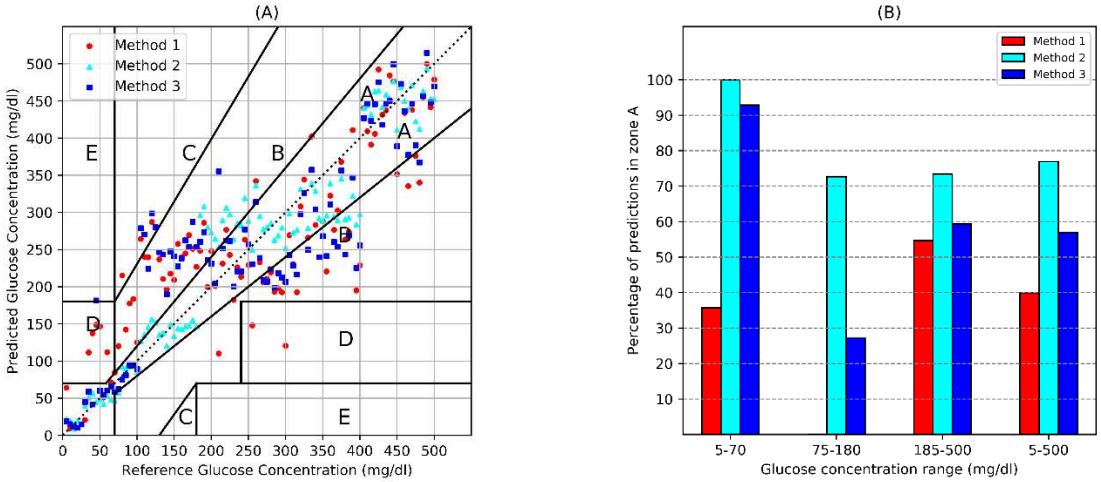


*Fig. 4. (A) EGA of the quantification methods in the NIR region (Some predictions of the first calibration method had negative values, so have not appeared in the graph.), (B) the statistics of the EGA graph*

### 4.2.2. Quantification results in the MIR region

Table 4 presents the quantification results of the three methods in the MIR region. Methods 2 and 3 were more accurate than Method 1, especially for data with lower glucose concentrations (Classes 1 and 2 data). The results of Methods 2 and 3 were comparable in the MIR region, which was due to the acceptable classification accuracy in this region.

*Table 4. Results of ten-fold cross-validation for the quantification methods in the MIR region*

| PM | RM | QM | Class 1 data (Hypoglycaemia range) | | | Class 2 data (Euglycaemia range) | | | Class 3 data (Hyperglycaemia range) | | | All data together (Whole glycaemic range) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | r |
| NP | PLSR | 1 | 73.8 | — | 211.0 | 38.2 | — | 30.5 | 32.5 | — | 9.5 | 41.5 | — | 0.95 |
| | | 2 | 17.0 | +76.9 | 45.4 | 14.4 | +62.3 | 11.3 | 31.5 | +3.0 | 9.2 | 26.6 | +35.9 | 0.98 |
| | | 3 | 15.0 | 79.6 | 42.9 | 23.3 | +39.0 | 18.6 | 31.2 | +4.0 | 9.1 | 27.9 | +32.7 | 0.98 |
| | PCR | 1 | 81.1 | — | 231.9 | 33.2 | — | 26.6 | 32.3 | — | 9.4 | 42.1 | — | 0.95 |
| | | 2 | 7.1 | +91.2 | 19.1 | 13.1 | +60.5 | 10.3 | 27.2 | +18.3 | 7.9 | 22.8 | +45.8 | 0.98 |
| | | 3 | 16.2 | +80.0 | 46.3 | 21.4 | +35.5 | 17.19 | 28.5 | +14.4 | 8.3 | 25.7 | +38.9 | 0.98 |
| S | PLSR | 1 | 73.9 | — | 211.2 | 37.9 | — | 30.3 | 32.5 | — | 9.5 | 41.4 | — | 0.95 |
| | | 2 | 17.0 | +76.9 | 45.4 | 14.3 | +62.2 | 11.2 | 31.4 | +3.3 | 9.1 | 26.8 | +35.2 | 0.98 |
| | | 3 | 15.0 | +79.7 | 42.9 | 23.2 | +38.7 | 18.5 | 31.2 | +4.0 | 9.1 | 27.9 | +32.6 | 0.98 |
| | PCR | 1 | 81.2 | — | 232.1 | 33.2 | — | 26.6 | 32.3 | — | 9.4 | 42.2 | — | 0.95 |
| | | 2 | 7.1 | +91.2 | 19.0 | 13.1 | +60.5 | 10.2 | 27.2 | +15.7 | 7.9 | 22.8 | +45.9 | 0.98 |
| | | 3 | 16.2 | +80.0 | 46.3 | 21.4 | +35.5 | 17.1 | 28.5 | +11.7 | 8.3 | 25.7 | +39.0 | 0.98 |
| MSC | PLSR | 1 | 73.8 | — | 211.0 | 38.2 | — | 30.5 | 32.5 | — | 9.5 | 41.5 | — | 0.95 |
| | | 2 | 17.0 | +76.9 | 45.4 | 14.4 | +62.3 | 11.3 | 31.5 | +3.0 | 9.2 | 26.9 | +35.1 | 0.98 |
| | | 3 | 15.0 | +79.6 | 42.9 | 23.2 | +39.2 | 18.6 | 31.2 | +4.0 | 9.1 | 27.9 | +32.7 | 0.98 |
| | PCR | 1 | 81.1 | — | 231.9 | 33.2 | — | 26.6 | 32.3 | — | 9.4 | 42.1 | — | 0.95 |
| | | **2** | **7.1** | **+91.2** | **19.1** | **13.1** | **+60.5** | **10.3** | **27.2** | **+15.7** | **7.9** | **22.8** | **+45.8** | **0.98** |
| | | 3 | 16.2 | +80.0 | 46.3 | 21.4 | +35.5 | 17.1 | 28.5 | +11.7 | 8.3 | 25.7 | +38.9 | 0.98 |
| S-MSC | PLSR | **1** | **65.9** | **—** | **188.3** | **37.2** | **—** | **29.8** | **32.5** | **—** | **9.4** | **39.5** | **—** | **0.96** |
| | | 2 | 9.3 | +85.8 | 24.9 | 11.9 | +68.0 | 9.4 | 27.8 | +14.4 | 8.1 | 23.2 | +41.2 | 0.98 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **3** | **10.2** | **+84.5** | **29.2** | **17.9** | **+51.8** | **14.3** | **26.6** | **+18.1** | **7.7** | **23.2** | **+41.2** | **0.98** |
| | PCR | 1 | 77.0 | — | 220.0 | 36.3 | — | 29.0 | 30.6 | — | 8.9 | 40.9 | — | 0.95 |
| | | 2 | 8.2 | +89.3 | 21.9 | 9.5 | +73.8 | 7.4 | 28.7 | +6.2 | 8.3 | 23.6 | +42.2 | 0.98 |
| | | 3 | 13.1 | +82.9 | 37.5 | 16.8 | +53.7 | 13.4 | 26.8 | +12.4 | 7.8 | 23.4 | +42.7 | 0.98 |

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

EGA for the best result of the quantification methods in the MIR region and the percentage of measurements distributed in Zone A for each method are displayed in Fig. 5. As the figures show, more accurate prediction results were obtained using Methods 2 and 3 rather than Method 1, notably for lower glucose levels.

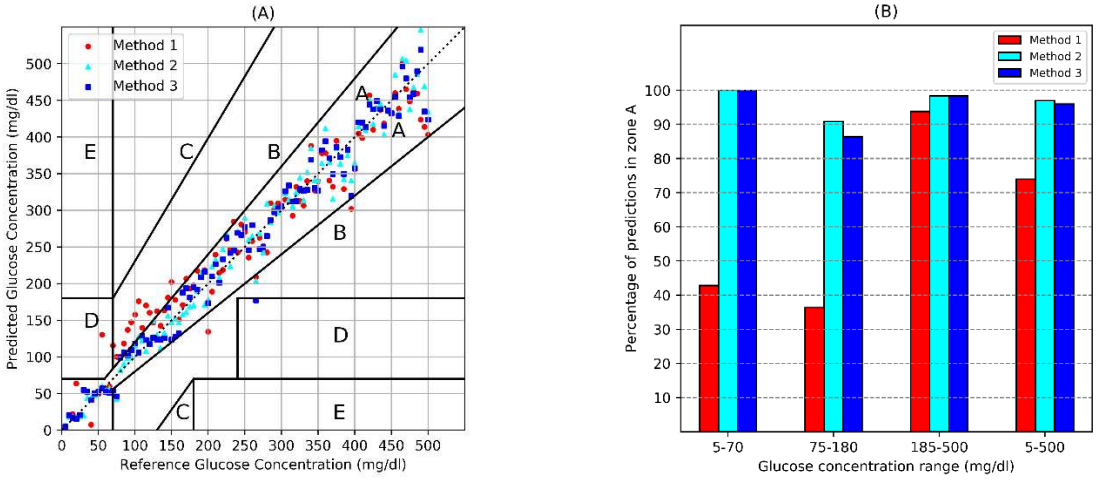


*Fig. 5. (A) EGA of the quantification methods in the MIR region, (B) the statistics of the EGA graph*

### 4.2.3. Quantification results in the NIR-MIR region

Table 5 reports the prediction results in the NIR-MIR region for all the quantification methods. Overall, in all cases, Methods 2 and 3 provided more accurate prediction results. All results obtained in the NIR-MIR region were generally comparable to those of the MIR region, which indicates that our proposed classification-before-regression methodology, still maintains its effectiveness over a wider range of spectra.

*Table 5. Results of ten-fold cross-validation for the quantification methods in the NIR-MIR region*

| PM | RM | QM | Class 1 data (Hypoglycaemia range) | | | Class 2 data (Euglycaemia range) | | | Class 3 data (Hyperglycaemia range) | | | All data together (Whole glycaemic range) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | r |
| NP | PLSR | 1 | 67.5 | — | 193.1 | 40.9 | — | 32.7 | 32.0 | — | 9.3 | 40.4 | — | 0.95 |
| | | 2 | 18.0 | +73.3 | 48.1 | 13.2 | +67.7 | 10.3 | 28.3 | +11.5 | 8.2 | 24.4 | +39.6 | 0.98 |
| | | 3 | 16.5 | +75.5 | 47.2 | 24.6 | +39.8 | 19.7 | 28.2 | +11.8 | 8.2 | 26.2 | +35.1 | 0.98 |
| | PCR | 1 | 82.1 | — | 234.8 | 37.6 | — | 30.0 | 33.1 | — | 9.6 | 43.6 | — | 0.95 |
| | | 2 | 8.1 | +90.1 | 21.6 | 12.3 | +67.2 | 9.7 | 26.2 | +20.8 | 7.6 | 21.9 | +49.7 | 0.98 |
| | | 3 | 14.1 | +82.8 | 40.5 | 14.7 | +60.9 | 11.7 | 24.9 | +24.7 | 7.2 | 21.7 | +50.2 | 0.98 |
| S | PLSR | 1 | 67.6 | — | 193.2 | 41.0 | — | 32.8 | 32.1 | — | 9.3 | 40.5 | — | 0.95 |
| | | 2 | 18.0 | +73.3 | 48.1 | 13.0 | +68.2 | 10.2 | 28.2 | +12.1 | 8.2 | 24.3 | +40.0 | 0.98 |
| | | 3 | 16.5 | +75.5 | 47.3 | 24.5 | +40.2 | 19.6 | 28.1 | +12.4 | 8.2 | 26.1 | +35.5 | 0.98 |
| | PCR | 1 | 82.0 | — | 234.3 | 37.6 | — | 30.0 | 33.1 | — | 9.6 | 43.6 | — | 0.95 |
| | | 2 | 8.0 | +90.2 | 21.3 | 12.3 | +67.2 | 9.6 | 26.2 | +20.8 | 7.6 | 21.9 | +49.7 | 0.98 |
| | | **3** | **14.2** | **+82.6** | **40.5** | **14.6** | **+61.1** | **11.6** | **24.9** | **+24.7** | **7.2** | **21.7** | **+50.2** | **0.98** |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MSC | PLSR | **1** | **67.5** | **—** | **193.1** | **40.9** | **—** | **32.7** | **32.0** | **—** | **9.3** | **40.4** | **—** | **0.95** |
| | | 2 | 18.0 | +73.3 | 48.1 | 13.2 | +67.7 | 10.3 | 28.3 | +11.5 | 8.2 | 24.4 | +39.6 | 0.98 |
| | | 3 | 16.5 | +75.5 | 47.2 | 24.6 | +39.8 | 19.7 | 28.2 | +11.5 | 8.2 | 26.2 | +35.1 | 0.98 |
| | PCR | 1 | 82.1 | — | 234.8 | 37.6 | — | 30.0 | 33.1 | — | 9.6 | 43.6 | — | 0.95 |
| | | 2 | 8.1 | +90.1 | 21.6 | 12.3 | +67.2 | 9.7 | 26.2 | +20.8 | 7.6 | 21.9 | +49.7 | 0.98 |
| | | 3 | 14.1 | +82.8 | 40.5 | 14.7 | +60.9 | 11.7 | 24.9 | +24.7 | 7.2 | 21.7 | +50.2 | 0.98 |
| S-MSC | PLSR | 1 | 75.6 | — | 216.0 | 37.1 | — | 29.7 | 32.8 | — | 9.5 | 41.8 | — | 0.95 |
| | | **2** | **9.7** | **+87.1** | **26.0** | **11.5** | **+69.0** | **9.0** | **26.0** | **+20.7** | **7.6** | **21.8** | **+47.8** | **0.98** |
| | | 3 | 10.6 | +85.9 | 30.4 | 16.5 | +55.5 | 13.2 | 26.9 | +19.9 | 7.8 | 23.28 | +44.3 | 0.98 |
| | PCR | 1 | 76.6 | — | 218.8 | 37.8 | — | 30.2 | 34.6 | — | 10.1 | 43.1 | — | 0.95 |
| | | 2 | 8.5 | +88.9 | 22.8 | 9.1 | +75.9 | 7.1 | 26.4 | +28.9 | 7.7 | 21.8 | +49.4 | 0.98 |
| | | 3 | 16.0 | +79.1 | 45.8 | 13.2 | +65.0 | 70.6 | 26.4 | +28.9 | 7.7 | 22.8 | +47.0 | 0.98 |

Abbreviations: PM = pre-processing method; RM = regression model; QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient; NP = no pre-processing; S = smoothing; MSC = multivariate scatter correction, S-MSC = smoothing couples with multivariate scatter correction.

The EGA graph for the best result of the quantification methods in the NIR-MIR region, a comparison of the predictions occurred in Zone A for each method are presented in Fig. 6. Based on the figure, it is clear that Methods 2 and 3 are more accurate than Method 1, also in the NIR-MIR region.

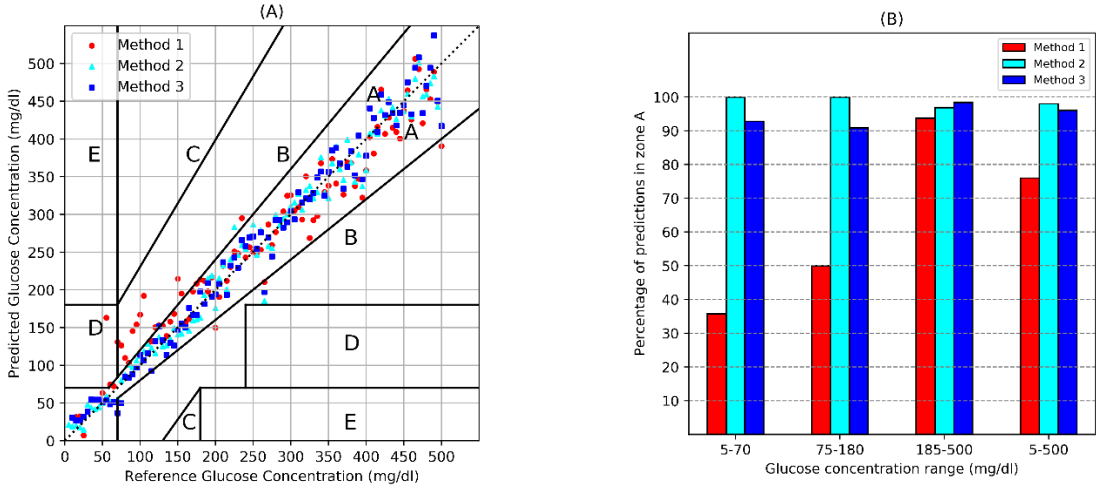


*Fig. 6. (A) EGA of the quantification methods in the NIR_MIR region, (B) the statistics of the EGA graph*

## 5. Discussion

All the quantification methods, as well as the PCA-LDA classifier applied in the third quantification method, showed a significantly better performance in the MIR and NIR-MIR regions than the NIR region. The reason is that the MIR and NIR-MIR signals contained a more extensive range of wavelengths; these can possess more informative wavebands for glucose monitoring compared to the NIR spectra in the experiment.

Methods 2 and 3 resulted in more accurate measurements than Method 1. Riley et al. showed that informative wavebands for the quantitative analysis of four chemical components, including glucose, becomes narrower by a decrease in the concentration range of these analytes [37]. It can be inferred that, in our dataset, informative wavebands for glucose measurement are possibly homogenous for signals in each class, and are different from the optimal spectral range of data in other classes. These similarities between spectra in each calibration set, using Methods 2 and 3, could improve the accuracy of the regression analyses.

As shown in Table 1, in our dataset, the 3 classes occupy different glycemic ranges dominated by samples within the hyperglycaemia range. However, datasets from in vivo experiments on humans generally tend to have the majority of data in the euglycaemia range; to show that our proposed methodology still applies in this case, the same analysis is repeated after adapting the step size in the hyperglycaemia range to allow for the majority of the data to be placed in the euglycaemia range. The step size for the hyperglycaemia range was increased from 5mg/dl to 20mg/dl; this lowered the number of samples by three quarters for this class as illustrated in Table 6, the other two classes remained unchanged. The new corresponding analysis results are shown in Table 7, which confirm the effectiveness of the pre-classification methodology for this distribution too.

*Table 6. Division of the dataset after the step size adaptation*

| | Label | | |
|---|---|---|---|
| | Class 1 | Class 2 | Class 3 |
| Glucose concentration range | 5–70 mg/dl | 75–180 mg/dl | 185–500 mg/dl |
| Corresponding glycaemic range | Hypoglycaemia | Euglycaemia | Hyperglycaemia |
| Quantity of data in the class | 14 | 22 | 16 |

*Table 7. Best results of ten-fold cross-validation for the quantification methods in all spectral region for modified data distribution*

| SR | QM | Class 1 data (Hypoglycaemia range) | | | Class 2 data (Euglycaemia range) | | | Class 3 data (Hyperglycaemia range) | | | All data together (whole glycaemic range) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | PEM (%) | RMSEP (mg/dl) | Im (%) | r |
| NIR | 1 | 115.1 | — | 328.9 | 72.2 | — | 57.8 | 164.8 | — | 49.1 | 118.2 | — | 0.42 |
| | 2 | 12.4 | +89.2 | 33.0 | 19.3 | +73.2 | 15.1 | 60.6 | +63.2 | 18.1 | 29.9 | +74.7 | 0.94 |
| | 3 | 54.1 | +52.9 | 154.6 | 76.1 | -5.1 | 60.9 | 129.8 | +21.2 | 38.7 | 92.1 | +4.1 | 0.72 |
| MIR | 1 | 78.1 | — | 223.1 | 34.0 | — | 27.2 | 50.8 | — | 14.7 | 53.1 | — | 0.92 |
| | 2 | 7.1 | +90.0 | 19.1 | 13.1 | +61.4 | 10.3 | 49.5 | +2.5 | 14.3 | 22.6 | +57.4 | 0.97 |
| | 3 | 11.0 | +85.9 | 31.7 | 20.6 | +39.4 | 16.5 | 74.3 | -46.2 | 21.5 | 44.6 | +16.0 | 0.94 |
| NIR-MIR | 1 | 62.8 | — | 179.5 | 46.7 | — | 37.4 | 51.1 | — | 14.5 | 52.1 | — | 0.92 |
| | 2 | 9.7 | +84.5 | 26.0 | 11.5 | +75.3 | 9.0 | 50.8 | +0.5 | 15.0 | 23.1 | +55.6 | 0.97 |
| | 3 | 15.8 | +74.8 | 45.1 | 20.2 | +56.7 | 16.2 | 53.6 | -4.8 | 15.5 | 34.1 | +34.5 | 0.96 |

Abbreviations: SR = spectral region QM = quantification method; RMSEP = root mean square error of prediction; Im = improvement of RMSEP in comparison to that of Method 1; PEM = percentage error around the mean; r = correlation coefficient

## 6. Conclusion

Glucose measurement using NIR and MIR absorbance spectroscopy improved by manually grouping the dataset into three categories according to the clinical definition of the glycaemic ranges and then applying regressions for each class separately. A PCA-LDA classifier was therefore implemented to assign each spectrum into the respective class automatically. Creation of regression models for different classes improved the results of glucose prediction as compared to regressions for the whole dataset. The improvements in the prediction results were more significant for lower glucose concentrations.

The performance of the proposed pre-classification approach was evaluated for two common regression methods, three pre-processing techniques, and also for a broader range of spectra by merging the NIR and MIR data. A primary evaluation of the proposed methodology was carried out by repeating the analysis for a modified version of the dataset to account for the distribution of data that is more representative of human in vivo experiments scenarios. For future work, determination of informative wavebands for glucose measurement could be investigated in each glycaemic range individually.

# References


[1] N. S. Oliver, C. Toumazou, A. E. G. Cass, and D. G. Johnston, 'Glucose sensors: A review of current and emerging technology', *Diabet. Med.*, vol. 26, no. 3, pp. 197–210, 2009.

[2] T. Vahlsing, G. Steiner, H. M. Heise, S. Delbeck, and S. Leonhardt, 'Non-invasive monitoring of blood glucose using optical methods for skin spectroscopy—opportunities and recent advances', *Anal. Bioanal. Chem.*, vol. 411, no. 1, pp. 63–77, 2018.

[3] J. Chung, H. So, Choi, and T. K. S. Wong, 'Recent advances in noninvasive glucose monitoring', *Med. Devices Evid. Res.*, p. 45, 2012.

[4] A. Al-Mbaideen and M. Benaissa, 'Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy', *Chemom. Intell. Lab. Syst.*, vol. 108, no. 2, pp. 112–122, 2011.

[5] J. Haas and B. Mizaikoff, 'Advances in Mid-Infrared Spectroscopy for Chemical Analysis', *Annu. Rev. Anal. Chem.*, vol. 9, no. 1, pp. 45–68, 2016.

[6] S. K. Vashist, 'Non-invasive glucose monitoring technology in diabetes management: A review', *Anal. Chim. Acta*, vol. 750, pp. 16–27, 2012.

[7] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis*. CRC press, 2007.

[8] H. von Lilienfeld-Toal, M. Weidenmüller, A. Xhelaj, and W. Mäntele, 'A novel approach to non-invasive glucose measurement by mid-infrared spectroscopy: The combination of quantum cascade lasers (QCL) and photoacoustic detection', *Vib. Spectrosc.*, vol. 38, no. 1–2, pp. 209–215, 2005.

[9] C.-F. So, K.-S. Choi, T. K. S. Wong, and J. W. Y. Chung, 'Recent advances in noninvasive glucose monitoring', *Med. Devices (Auckland, NZ)*, vol. 5, p. 45, 2012.

[10] B. Rabinovitch, W. F. March, and R. L. Adams, 'Noninvasive glucose monitoring of the aqueous humor of the eye: Part I. Measurement of very small optical rotations', *Diabetes Care*, vol. 5, no. 3, pp. 254–258, 1982.

[11] J. Yadav, A. Rani, V. Singh, and B. M. Murari, 'Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy', *Biomed. Signal Process. Control*, vol. 18, pp. 214–227, 2015.

[12] A. Tura, A. Maran, and G. Pacini, 'Non-invasive glucose monitoring: Assessment of technologies and devices according to quantitative criteria', *Diabetes Res. Clin. Pract.*, vol. 77, no. 1, pp. 16–40, 2007.

[13] J. Tenhunen, H. Kopola, and R. Myllylä, 'Non-invasive glucose measurement based on selective near infrared absorption; requirements on instrumentation and spectral range', *Meas. J. Int. Meas. Confed.*, vol. 24, no. 3, pp. 173–177, 1998.

[14] 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.

[15] K. C. 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', *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 6210–6213, 2016.

[16] J. M. Andrés and M. T. Bona, 'ASTM clustering for improving coal analysis by near-infrared spectroscopy', *Talanta*, vol. 70, no. 4, pp. 711–719, 2006.

[17] Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, and G. Li, 'Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy', *Sensors Actuators, B Chem.*, vol. 193, pp. 723–729, 2014.

[18] H. Chen, Z. Liu, J. Gu, W. Ai, J. Wen, and K. Cai, 'Quantitative analysis of soil nutrition based on FT-NIR spectroscopy integrated with BP neural deep learning', *Anal. Methods*, vol. 10, no. 41, pp. 5004–5013, 2018.

[19] J. Kropff *et al.*, 'Accuracy of two continuous glucose monitoring systems: To-Head Comparison Under Clinical Research Centre and Daily Life', *Diabetes, Obes. Metab.*, vol. 2015, no. 17, pp. 343–349, 2015.

[20] L. C. Lee, C. Y. Liong, and A. A. Jemain, 'A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum', *Chemom. Intell. Lab. Syst.*, vol. 163, no. December 2016, pp. 64–75, 2017.

[21] 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.

[22] J. Yadav, A. Rani, V. Singh, and B. M. Murari, 'Prospects and limitations of non-invasive blood glucose monitoring

using near-infrared spectroscopy', *Biomed. Signal Process. Control*, vol. 18, pp. 214–227, 2015.

[23] Y. Mou, X. You, D. Xu, L. Zhou, W. Zeng, and S. Yu, 'Regularized multivariate scatter correction', *Chemom. Intell. Lab. Syst.*, vol. 132, pp. 168–174, 2014.

[24] D. Wu, J. Chen, B. Lu, L. Xiong, Y. He, and Y. Zhang, 'Application of near infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract', *Food Chem.*, vol. 135, no. 4, pp. 2147–2156, Dec. 2012.

[25] Ian T. Jolliffe, 'A Note on the Use of Principal Components in Regression', *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 31, no. 3, pp. 300–303, 1982.

[26] G. M. Escandar, P. C. Damiani, H. C. Goicoechea, and A. C. Olivieri, 'A review of multivariate calibration methods applied to biomedical analysis', *Microchem. J.*, vol. 82, no. 1, pp. 29–42, 2006.

[27] and L. E. Wold, Svante, Michael Sjöström, 'PLS-regression: a basic tool of chemometrics.', *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[28] L. A. Berrueta, R. M. Alonso-Salces, and K. Héberger, 'Supervised pattern recognition in food analysis', *J. Chromatogr. A*, vol. 1158, no. 1–2, pp. 196–214, 2007.

[29] S. A. Drivelos and C. A. Georgiou, 'Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union', *TrAC - Trends Anal. Chem.*, vol. 40, pp. 38–51, 2012.

[30] J. Liu and S. Chen., 'Resampling LDA/QR and PCA+ LDA for face recognition', *Australas. Jt. Conf. Artif. Intell. Springer, Berlin, Heidelb.*, pp. 1221–1224, 2005.

[31] A. Pasini, 'Artificial neural networks for small dataset analysis', *J. Thorac. Dis.*, vol. 7, no. 5, pp. 953–960, 2015.

[32] J. Shao, 'Linear model selection by cross-validation', *J. Am. Stat. Assoc.*, vol. 88(422), pp. 486–492, 1993.

[33] M. L. F. Simeone, R. A. C. Parrella, R. E. Schaffert, C. M. B. Damasceno, M. C. B. Leal, and C. Pasquini, 'Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice', *Microchem. J.*, vol. 134, pp. 125–130, 2017.

[34] and W. A. N. Lee Rodgers, Joseph, 'Thirteen ways to look at the correlation coefficient.', *Am. Stat.*, vol. 42, no. 1, pp. 59–66, 1988.

[35] W. L. Clarke, 'The original Clarke error grid analysis (EGA)', *Diabetes Technol. Ther.*, vol. 7, no. 5, pp. 776–779, 2005.

[36] A. Al-Mbaideen and M. Benaissa, 'Frequency self deconvolution in the quantitative analysis of near infrared spectra', *Anal. Chim. Acta*, vol. 705, no. 1–2, pp. 135–147, 2011.

[37] M. R. Riley and H. M. Crider, 'The effect of analyte concentration range on measurement errors obtained by NIR spectroscopy', *Talanta*, vol. 52, no. 3, pp. 473–484, 2000.