**Signal Preprocessing Techniques to enhance Glucose Prediction from NIR and MIR spectra**

Thesis submitted to the University of Sheffield for the degree of

Doctor of Philosophy

By

**Osamah Abdulhameed Mousa Alrezj**

Department of Electronics and Electrical Engineering

The University of Sheffield

Mappin Street,

Sheffield, S1 3JD

United Kingdom

September 2019

**Abstract**

Diabetes is a combination of [metabolic diseases](https://en.wikipedia.org/wiki/Metabolic_disease) affecting the lives of millions of peoples around the world and growing vastly. If the pancreas don’t produce enough insulin or the body cells resist the produced insulin this will affect the glucose level in the blood. If the high glucose level left without treatment it could accumulate and cause organs failure leading in some cases to death. Insulin injections or medication can reduce the high glucose ratio effect, but also can cause low glucose levels below the normal level denying the brain from the required glucose and causing comma.

The main factor is accurate measurements of the glucose level in the blood to avoid both extreme high and low glucose levels (Hyperglycemic and hypoglycemic). Monitoring currently is done using invasive and minimal invasive methods which are painful and inconvenient. Recent studies show that blood glucose levels can be enhanced by adapting noninvasive methods.

Since the glucose signal is overlapped with other constituents in the blood a computational mathematical methods such as Principal components and Partial least squares are used to estimate and separate the glucose from other constituents after transferring the signal in terms of light absorbance using NIR and MIR spectroscopy.

The extraction of glucose signal using the spectrometer added additional noise to the glucose signal which make it more complicated for the computational methods to accurately predict the glucose signal hence, Digital signal processing step has been proposed to clean the glucose from unwanted noise before the use of the prediction models but these preprocessing step cannot reduce all the interfered signals.

This thesis proposes the use of multiple preprocessing methods before the use of the prediction model to further clean the raw data from additional noise to enhance the prediction ability of the prediction models of glucose levels in a designed chemical mixture consist of glucose and human serum albumin in a phosphate buffer solution using near and mid Infrared spectroscopy and with the help of different Signal processing techniques such as Savitzky Golay and Multiplicative scatter correction to produce Continuous Remote Noninvasive Glucose Monitoring.

**Acknowledgement**

All praises be for Allah s.w.t, the beneficent, the merciful and the lord of the whole universe along with his mercy that have been awarded to the author.

I would like to express my appreciation and ultimate gratitude to my supervisor Dr. Mohammad Benaissa for his continues support, encouragement, motivation and many helpful advices and guidance through my research work during my PhD study period and doing research in his group in the University of Sheffield. With his enthusiastic supervision, logical thinking, comments, his idea and knowledge made my whole PhD journey clear, positive and enjoyable.

I am very thankful to all my research colleagues and staff in the department of Electronics and Electrical Engineering for their help, advice and assistance, especially Krishna Patchava and all the people from C34 and B28 in Portobello Building.

Special thanks to my lovely wife, Amani Albomesheh, for being my friend, and number one fan. Without her moral support, sacrifice and understanding it would not have been possible to me to complete this project, to my dearest daughter Noorulzahraa Alrezj and my lovely sons Ali and Abbas.

My love and most sincere gratitude to my beloved family members Father, Abdulhameed Alrezj, mother Shukriyah Maalah, brothers: Ali and Jaafar, and sister Esraa, thank you for your thoughtful prayer and your countless love, support, understanding and encouragement to chase my dreams.

Financial support provided by Higher Committee of Education Development Iraq (HCED).

**Content**

Abstract…………………………………………………………………………………….II

Acknowledgment………………………………………………………………………….III

Content ……………………………………………………………………………………IV

List of figures…….....………………………………...………………………………….VIII

List of tables..……………………………………………………………………………....X

Abbreviation………….…………………………………………………………...............XI

Chapter One: Introduction…………………………………………………………………. 1

1.1 Overview…….….………...………………………………………………………..1

1.2 Research aim and motivation………………...………….……………………….....2

1.3 Key contribution.......………………...……………………………………………..3

1.4 Thesis outline……………………………………………………………………….4

1.5 Publication………………………………………………………………………….5

Chapter Two: Literature Review ………………..……………………………………...…..7

2.1 NIR spectroscopy………………………….……………………………………….. 7

2.1-A Glucose sensing using near infrared spectroscopy…...……………………….8

2.2 MIR Spectroscopy.....………..……………………………………………………..14

2.2-A Glucose sensing using mid infrared spectroscopy…………………………...15

2.3 Scatter correction..………………………………………………………………….21

2.4 Savitzky and Golay filtering..…...………………………………………………….26

Chapter Three: Background, sample preparation and spectra collection………………….34

3.1 The research gap..………………….……………………………………………….34

3.2 Multivariate calibration methods..………………………………………………….35

3.2-A Principal component analysis (PCA) ……………………………………….36

3.2-B Principal component regression (PCR)..…………………………………….38

3.2-C Partial least squares regression (PLSR)...…………………………………...39

3.3 Selecting the model parameters..…………………………………………………...41

3.4 Preprocessing..…………………………………………………………………..…42

3.4-A Data metrics mean centering...…..………………………………………….43

3.4-B Smoothing...………………………………………………………………...43

3.4-C Derivative...…………………………………………………………………44

3.4-D Digital filtering...……………………………………………………………45

3.4-D.1 Digital bandpass filtering…………………………………………...45

3.4-D.2 Digital bandstop filtering……………………………………………46

3.4-D.3 Butterworth and Chebyshev filters………………………………….46

3.5 Data preparation...………………………………………………………………….47

3.5-A Dataset 1..……………………………………………………………………47

3.5-B Dataset 2..……………………………………………………………………48

3.5-C Dataset 3..……………………………………………………………………50

3.5-D Dataset 4..……………………………………………………………………51

Chapter Four: Scatter correction techniques coupled with bandpass filtering.………..…..55

4.1 Introduction..……………………………...………………………………………..55

4.2 NIR and MIR spectroscopy...….………………………….………………………..55

4.3 Light scattering and baseline effect...…….………………………………………....57

4.3-A Multivariate scatter correction (MSC)……………………………………...58

4.3-B Standard normal variate (SNV)……………………………………………..59

4.4 Experimental data preparation..……………………………………………………59

4.5 Discussion and experimental results..……………………………………………...60

4.6 Conclusion....………………………………………………………………………62

Chapter Five: Savitzky Golay optimization coupled with bandpass filtering……………..69

5.1 Introduction…………………………………………....…………………………69

5.2 Savitzky Golay filtering ………………..………………………………………..70

5.3 Discussion and experimental results……………………………………………..73

5.4 Conclusion………………………………………………………………………..74

Chapter Six: bandstop filtering in the quantitative analysis of glucose ……………….......86

6.1 Introduction………………………………………………………………….......86

6.2 Background………………….…………………………………………………..87

6.3 Theory…………………….………….……………………………………...…..89

6.4 Multivariate calibration..…………………………………………………..….....89

6.5 Preprocessing……...………………………………………………………...…..89

6.6 Digital filtering………………………….…………...………………………….90

6.6-A Digital bandpass filtering…………….…...……………………………...90

6.6-B Digital bandstop filtering…..….……………………………………….....91

6.6-C Butterworth and Chebyshev filter….…….…………………………….....91

6.7 Data preparation..……………………….……………..………………………..91

6.7-A Chemical sample preparation………………………………………….....92

6.7-B Spectra collection………………………………………………………...92

6.8 Proposed model…..……………………………………………………….….....93

6.9 Experimental results and discussion…..………………………………………...96

6.10 Conclusion……………………………………………………………………..98

Chapter Seven: Bandstop filtering with preprocessing techniques…………………..…..107

7.1 Introduction…..………………………………………………………………..107

7.2 Proposed model.……………………………………..………………………...107

7.3 Preprocessing….………………………………………………………………108

7.4 Digital filtering.………………………………………………..………………108

7.4-A Digital bandpass filtering……………………………………………….109

7.4-B Digital bandstop filtering………………………………………………..109

7.4-C Butterworth filtering…………………………………...………………..110

7.5 Light scattering and baseline effects……………………………..……………110

7.6 Experimental data preparation…………………………………………………111

7.7 Discussion and experimental results…………………………………………...112

7.8 Conclusion ….…………………………………………………………………113

Chapter Eight: Conclusion and future work……………………………………………...120

8.1 Conclusion…………………………………………………………………….120

8.2 Future work…………………………………………………………………...122

8.3 Implementation Cost …………………………………………………….……124

8.3. A NIR ………………………………………………………..…………....124

8.3A.1 DLP2010NIR …………………………………………………..…..124

8.3A.2 DLP4500NIR ………………………………………………….…...124

8.3B MIR ………………………………………………………………………..125

8.3B.1 4100 Exo Scan Series …………………………………………………125

References ……………………………………………………………………………….126

**List of Figures**

|  |  |
| --- | --- |
| 3.1 Choosing the optimum latent value………………………..………………...… | 42 |
| 3.2 Dataset 1 plot …………..……………………………………………………… | 48 |
| 3.3 Data set 2 plot ...……………………………………………………………….. | 49 |
| 3.4 Data set 3 plot...……………………………………………………………...… | 51 |
| 3.5 Data set 4 plot……....………………………………………...………….…..… | 52 |
| 3.6 Preparation process of dataset 1……………………………….…………….…. | 53 |
| 3.7 3.7 Preparation process of dataset 2, 3 and 4……..……………….………...…. | 54 |
| 4.1 RMSEP PCR dataset 2 ….…………………….………………….…………..... | 66 |
| 4.2 RMSEP PLSR dataset 2…………………….…...…..……………………….…. | 66 |
| 4.3 RMSEP PCR dataset 3 ...…………………………………..…..……...…............ | 67 |
| 4.4 RMSEP PLSR dataset 3.…………...………..………………….………...…..…. | 67 |
| 4.5 RMSEP PCR dataset 4 .……………………..…………………..…………...….. | 68 |
| 4.6 RMSEP PLSR dataset 4….……….…………………………………………..….. | 68 |
| 5.1 PLSR vs PCR for dataset 1.…………………………….……….………….……. | 76 |
| 5.2 PLSR vs PCR for dataset 3………………..…………………………………....... | 76 |
| 5.3 PLSR vs PCR for dataset 4………………………………………..………..……. | 77 |
| 5.4 PCR dataset 1 with different window sizes …………...……………….………... | 77 |
| 5.5 PLSR dataset 1 with different window sizes ….……….………………………... | 81 |
| 5.6 PCR dataset 3 with different window sizes ……………….…………………….. | 81 |
| 5.7 PLSR dataset 3 with different window sizes………………………….…………. | 82 |
| 5.8 PCR dataset 4 with different window sizes…………………………….….…….. | 82 |
| 5.9 PLSR dataset 4 with different window sizes ……………...……………….……. | 83 |
| 5.10 PCR dataset 1 comparison between different window sizes....……….………... | 83 |
| 5.11 PLSR dataset 1 comparison between different window sizes ………………..... | 84 |
| 5.12 PCR dataset 3 comparison between different window sizes …………..……..... | 84 |
| 5.13 PLSR dataset 3 comparison between different window sizes………………….. | 85 |
| 5.14 PCR dataset 4 comparison between different window sizes…………………… | 85 |
| 5.15 PLSR dataset 4 comparison between different window sizes……...…………... | 85 |
| 6.1 Dataset 2 and 4 plot ……………………………………….……………………..  6.2 PCR variance for dataset 2 and 4 ………………………………………..…….... | 94  95 |
| 6.3 PLSR variance for dataset 2 and 4 ….……………………………..…………..... | 95 |
| 6.4 system flow chart ………………………………………...……………………… | 96 |
| 6.5 PLSR dataset 2 with different filters ……………………………...……………... | 99 |
| 6.6 PCR dataset 2 with different filters………………………………………………. | 99 |
| 6.7 PLSR dataset 4 with different filters …...………………………………………... | 100 |
| 6.8 PCR dataset 4 with different filters ………………………..……………………. | 100 |
| 6.9 RMSEP for Butterworth bandpass vs bandstop PLSR dataset 2…………………  6.10 RMSEP for Butterworth Bandpass vs Bandstop PLSR dataset 4……………….  6.11 RMSEP for Chebyshev bandpass vs bandstop PLSR dataset 2………………… | 101  101  104 |
| 6.12 RMSEP for Chebyshev bandpass vs bandstop PCR dataset 2………………….. | 104 |
| 6.13 RMSEP for Butterworth bandpass vs bandstop PCR dataset 2………………… | 105 |
| 6.14 RMSEP for Chebyshev bandpass vs bandstop PLSR dataset 4………………… | 105 |
| 6.15 RMSEP for Chebyshev bandpass vs bandstop PCR dataset 4………………………….. | 106 |
| 6.16 RMSEP for Butterworth bandpass vs bandstop PCR dataset 4………………………… | 106 |
| 7.1 Dataset 2 PLSR Butterworth with preprocessing….…………………………….. | 118 |
| 7.2 Dataset 4 PLSR Butterworth with preprocessing………………………………... | 118 |
| 7.3 Dataset 2 PLSR Chebyshev with preprocessing…………………………………. | 119 |
| 7.4 Dataset 4 PLSR Chebyshev with preprocessing………………………………… | 119 |
| **List of Tables** |  |
|  |  |
| 4.1 Results for dataset 2 .…...…..…………………………………………………… | 63 |
| 4.2 Results for dataset 3…………..…………………………………………………. | 64 |
| 4.3 Results for dataset 4.…………………………………………………………….. | 65 |
| 5.1 Results for dataset 2……………………………………………………………... | 78 |
| 5.2 Results for dataset 3…….……………………………………………………….. | 79 |
| 5.3 Results for dataset 4……………………………………………………………... | 80 |
| 6.1 Results for dataset 2…………………………………..………………………….  6.2 Results for dataset 4…………………………………..………………………….  7.1 Results for PLSR Butterworth dataset 4………………………………………… | 102  103  114 |
| 7.2 Results for PLSR Butterworth dataset 2…………………………………………  7.3 Results for PLSR Chebyshev dataset 2…………………………………………. | 115  116 |
| 7.4 Results for PLSR Chebyshev dataset 4…………………………………………. | 117 |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **Abbreviation** |  |
| |  |  | | --- | --- | | ACLS | Augmented Classical Least Square | | BSA | Bovine Serum Albumin | | CC | Carbon Carbon bond | | CDBPF | Chebyshev Digital Band Pass Filter | | CGM | Continues Glucose Monitoring | | CH | Carbon Hydrogen Bond | | CO | Carbon Oxygen Bond | | CV | Cross Validation | | EC-QCL | Quantum Cascade Laser with External Cavity | | EIMSC | Extended Inverse Extended Multiplicative Scatter Correction | | EMSC | Extended Multiplicative Scatter Correction | | FSDRPLS | Frequency Self-Deconvolution Regression PLS | | FTIR | Fourier Transform Infrared | | GA | Genetic Algorithm | | GDBPF | Gaussian Digital Band Pass Filter | | HSA | Human Serum Albumin | | ICA | Independent Component Regression | | IMSC | Inverse Extended Multiplicative Scatter Correction | | IPLS | Interval PLS | | IR | Infrared Radiation | | KS | Kennard-Stone | | LLER | Local Linear Embedded Regression | | LOOC | Leave One Out Cross Validation | | LS-SVR | Least Square SVR | | LV | Latent Value | | LWPLS | Locally Weighted PLS | | MC | Multivariate Calibration | | MIR | Mid Infrared Radiation | | MLR | Multiple Linear Regression | | MSC | Multiplicative Scatter Correction | | MSEP | Mean Square Error of Prediction | | MWPLS | Moving Window Partial Least Square | | NIR | Near Infrared Radiation | | PC | Principal Component | | PCA | Principal Component Analysis | | PCR | Principal Component Regression | | PLS | Partial Least Square | | PLSR | Partial Least Square Regression | | QCL | Quantum Cascade Laser | | R | Determination Coefficient | | RBF | Radial Basis Function | | RMSEP | Root Mean Square Error of Prediction | | RPLS | Recursive Partial Least Square | | RreliefF | Regressional Relief-F | | SCMPWPLS | Searching Combination MPWPLS | | SDICA | Subband Decomposition PCA | | SEP | Standard Error of Prediction | | SG | Savitzky-Golay | | SNV | Standard Normal Variate | | SVR | Support Vector Regression | | UVE | Uninformative Variable Elimination | | VIP | Variable Importance in Projection | |  |

**Chapter One**

**Introduction**

**1.1 Overview**

Human blood is a magnificent mixture consisting of balanced chemical components; changing the concentration of any one of these components could generate a serious illness affecting the vitality of major organs in the human body such as the heart, kidneys and the eyes. One of these well-known diseases is Diabetes which is a multiple type disease caused by different reasons and require different treatments yet the effect is common. This effect is the irregularity in the glucose level of the human blood. The variation in the glucose concentration introduces two critical cases.

1. Hyperglycemia refers to the high glucose level. This additional glucose concentration in the blood will accumulate in the body organs and cause failure in their functionality. This may produce another serious illness such as Heart attack or kidney failure.
2. Hypoglycemic refers to very low glucose concentrating in the blood. This could cause brain failure and immediate comma.

Treating diabetic patients requires continues measurement of the glucose level before any treatment should be given. Current measurements are based on finger bricking for collecting blood samples which are painful, inconvenient and invasive. In addition the accessories are expensive, about 800£ per year per patient. Many studies show that diabetes can be controlled if a continuous monitoring process is provided to diabetic patients. The invasive technique prevents patients from keeping up with the monitoring process, therefore; adapting a noninvasive techniques is important to monitor the glucose level and enhance the health of the diabetic patients.

Near and Mid Infrared Spectroscopy suggests perfect solution for none invasive glucose level monitoring due to low sample preparation time, wide absorbance range and none destructive nature. Yet the collected spectra contains information about all the chemical components not only Glucose, hence comes the necessity to use complicated computational algorithms to extract the desired signal from the correlated and overlapped mixture. This research shows combining signal processing techniques with the computational algorithms to help improving the Glucose level prediction by eliminating the additional noise and suppressing the unwanted signals.

## **1.2 Research Aims and motivation**

This Thesis introduces an engineering solution to a problem in the field of chemometrics by integrating signal preprocessing techniques with regression computational algorithms to enhance the prediction of the glucose level from near and mid-infrared spectra extracted from a designed chemical mixture.

The objectives of the current research are listed below:

1. To introduce accurate and full-bodied prediction model that can be integrated into a practical system for none – invasive continuous glucose level monitoring.
2. To experimentally test all suggested signal pre-processing techniques and regression models using more realistic data such as blood serum.
3. To measure the collected data of interest with different machines (spectrometers) to ensure the accuracy of the introduced prediction model.
4. To explore and implement new prediction models then compare their accuracy, efficiency, robustness and complexity.

## **1.3 Key contribution**

The thesis mainly focuses on enhancing the prediction ability of linear regression techniques like the well-known principal component and the partial least squares to predict glucose concentration from a chemical mixture consist of glucose and HSA in a phosphate buffer. The prediction process is being facilitated by measuring the light intensity absorbance of the mixture samples using NIR and MIR spectroscopy. The data generated from the spectroscopic process is then being treated using different preprocessing techniques to clean the noisy data from instrumental noise and interference with other mixture components. The thesis main contributions are listed below.

1. Reducing the light scattering effects using light scattering techniques combined with bandpass filtering.
2. Reducing the effects of the background noise and the ambient noise using the combination of Savitzky Golay smoothing filter with bandpass filtering.
3. Reducing the overlapping effects between the glucose and other mixture contents using bandstop filtering methodology.
4. Reducing the regression models complexity by reducing the required regression factors the principal components and the latent values.
5. Exploring the effects of different proposed models on different spectral ranges such as NIR and MIR.
6. Exploring the effects of wide and narrow band filters such as Chebyshev and Butterworth on the glucose prediction.

## **1.4 Thesis outline**

The thesis consist of eight chapters, the organization of these chapter is as follow.

1. Chapter one introduces the effects of diabetes on the human body and the necessity of noninvasive monitoring which glucose prediction plays a great role in it. Thesis motivation and main contribution is also introduced in this chapter in addition to the resulting publication.
2. Chapter two introduces a detailed literature review listing the research articles and projects which addressed the use of NIR and MIR spectroscopy in glucose prediction with the help of different preprocessing methods.
3. Chapter three addresses the background section and the tools required for fulfilling the objectives of this thesis in addition to the chemical mixture preparation, sample acquisition and spectra collection, to provide the necessary data for our analysis.
4. Chapter four introduced a novel model based on the combination of scatter correction techniques with bandpass filtering to enhance the prediction ability of the linear calibration models PCR and PLSR. The chapter starts with a brief introduction about the scatter correction techniques and its importance to the prediction of glucose concentration, a brief of which data being used in this chapter and experimental results with a detailed discussion.
5. Chapter five introduces a novel model based on the optimization of SG smoothing filter parameters combined with the bandpass filtering to enhance the prediction ability of the linear calibration models PCR and PLSR. This chapter starts with a brief introduction about the Savitzky Golay filtering and its importance to the prediction of glucose concentration, a brief of which data being used in this chapter and experimental results with a detailed discussion.
6. Chapter six introduces a novel model based on the use of bandstop filtering to enhance the prediction ability of the linear calibration models PCR and PLSR as an alternative to the bandpass filtering method. A data preparation section and experimental results with a detailed discussion has been provided.
7. Chapter seven introduces a novel model based on the combination of bandstop filtering with different preprocessing methods used in chapters, 4, 5 and 6 such as SNV and Savitzky Golay filter to enhance the prediction ability of the linear calibration models PCR and PLSR as an alternative to the bandpass filtering method combined with the same preprocessing methods. A data preparation section and experimental results with a detailed discussion has been provided.
8. Chapter eight summarizes the conclusion and the future work proposed to complete the research in the future.

## **1.5 Publications**

1. Krishna Chaitanya, Osamah Alrezj and Mohammed Benaissa “Savitzky-Golay coupled with digital bandpass filtering as a pre-processing technique in the quantitative analysis of glucose from near infrared spectra” Published by the *IEEE* *EMBC 2016, Orlando, FL, USA, Augest 17-20, 2016*.
2. Osamah Alrezj, Krishna Chaitanya and Mohammed Benaissa “Scatter correction techniques coupled with band pass filtering to predict glucose from near and mid-infrared spectroscopy using human serum albumin data” Published by the *IEEE* *EMBC 2017, JeJu Iseland, South Korea, July 17-21, 2017*.
3. O.A. Alrezj, K. Patchava M. Benaissa and S. A. Alshebeili “Pre-processing to Enhance the Quantitative Analysis of Glucose from NIR and MIR Spectra” Published by the *EMBEC 2017, Tampere, Finland, June 10-14, 2017.*
4. O Alrezj, M Benaissa and S Alshebeili “Digital Bandstop Filtering in the Quantitative Analysis of Glucose from Near-Infrared and Mid Infrared Spectra” Published in the Journal of Chemometrics 25 December 2019.
5. O Alrezj and M Benaissa “Digital Bandstop Filtering with Preprocessing in the Quantitative Analysis of Glucose from Near-Infrared and Mid Infrared Spectra” submitted to the ICASSP international conference Barcelona 4-8 May 2020.
6. 6- Heydar Khadem, Mohammad R. Eissa, Hoda Nemat, Osamah Alrezj and Mohammed Benaissa “Classification before Regression for Improving the Accuracy of Glucose Quantification using Absorption Spectroscopy” published in Talanta Journal, 13 Jan 2020.

**Chapter Two**

**Literature review**

This chapter introduces the previous work done in Glucose prediction from NIR and MIR spectra. The reader can identify the phases that Glucose prediction went through starting from simple MLR and ending with coupling different preprocessing techniques with different regression models in addition to the different datasets and wavebands used in every research.

**2.1 NIR spectroscopy**

Recent studies give most attention to near-infrared (NIR) spectroscopy because it has the following advantages.

1. The ability to measure spectra samples for different states such as a liquid or solid without any manipulation.
2. Affordable equipment cost.
3. Availability of portable equipment.

In addition to that, the radiation of NIR can penetrate the human skin in depth up to few millimeters through serum, tissue and fluids [1]. NIR light primarily is absorbed by glucose in two observed regions [1].

1. The overtone absorption region which varies from 1500 nm–1800 nm.
2. The combination band region which varies from 2050 nm– 2300 nm.

The NIR spectra of components can be collected by using a Fourier Transform Infrared FTIR spectrometer that measures the absorbed or transmitted NIR radiation as a function of the wavenumber or wavelength [2]. FTIR spectrometers have found their way into several applications in fields such as environmental applications, industry, medicine, biochemistry and chemistry due to the following advantages.

1. FTIR spectrometers tend to provide high signal to noise ratio.
2. FTIR spectrometers provide high resolution and precision.
3. FTIR spectrometers provide high optical throughout with a short scanning time.

**2.1-A Glucose sensing using near infrared spectroscopy**

Arnold and his group have devoted special attention to measuring glucose form NIR spectra in [3], they used digital Fourier filter to measure glucose from phosphate buffer with pH of 7.2 and wavelength range from 2000 nm - 3000 nm. This work used only 17 samples and the data was very simple compared to human blood. In [4], Arnold introduced the idea of coupling digital band pass filtering with the multivariate regression model PLS to predict the glucose from Bovine serum albumin in a phosphate buffer and narrowed the waveband to be from 2000 nm to 2500 nm in this work they identify the protein as a major potential interference with the NIR measurements and used 14 LV which considers high and could lead to model overfitting. The model overfitting was not introduced yet and background variation yet to be tested and the calibration model was not able to predict glucose from different data sets also the parameters of the preprocessing technique required further tuning. In [5], the same group mentioned above increased their data complexity by using bovine plasma to predict the glucose from NIR spectra and developed the calibration model to achieve better SEP as well. The same waveband used. In this work a relation between the bandpass filtering and the enhancement of the calibration model is established by reducing the optimum LV number required for glucose prediction but the work was not conclusive about the required optimization degree of the filter parameters. Also, this work drew attention to the importance of SG as a derivation filter with a smoothing technique.

The effect of change in temperature on the glucose measurements accuracy was investigated by several groups. In [6], a PLS model coupled with digital Fourier filter was implemented to predict glucose form phosphate buffer with a range of temperature from 32o to 41o C. The research explained that it is possible to minimize the effect of change in temperature on the glucose prediction and an SEP of 2.52 mg/dL was obtained. In [7], two different data sets used by Arnold and his associates to use NIR spectroscopy in glucose level. In the first dataset Triacetin has been used to provide chemical samples with triglyceride levels similar to the levels in human blood whereas in the second dataset the BSA were used to provide chemical samples with protein levels similar to the human blood. The wave band range is 2000 nm to 2500 nm for both data sets and PLS is coupled with digital Fourier filter as a preprocessing technique. This work presented the importance of choosing the optimum number of LV to achieve maximum SEP reduction. Also, it explained in the case of overlapped absorption bands such as the case of BAS, glucose and the triacetin, it will be difficult for Fourier filtering to extract the information related to a specific element in the mixture without being affected by the rest. Whereas Fourier filtering increase the independency of the regression model on the required number of regression factors and on the required spectral range to reach desirable prediction model.

PLSR has been used in [8] [9], to predict glucose from NIR and raised the Glucose concentration to higher levels up to 100 mM with different data sets using Ammonia, Lactate and Glutamate. This work has explained that the optimum number of LV is unique for each data set and the best performance can be achieved by combining Fourier filtering and PLS regression. The first introduction to GA was explained in [10], to apply the procedure of selecting the wavelength automatically to generate a regression models based on PLS without pre-processing **.** In [11], Shaffer has integrated the method in [8], by using GA to implement coupling of Bandpass filtering with PLS regression to predict the glucose from three datasets where GA was used to optimize the regression model spectral range, the number latent values used to implement the regression model and the position and width of the bandpass digital filter.Two critical drawbacks to the optimization success were found in this study: the development of an appropriate fitness function and GA configuration. In [12], GA has been employed to decrease the required number wavelengths to build a PLS regression model with the use of lower spectral resolution. On the other hand reducing the resolution of the spectra will not give accurate information about the glucose. The integrating of PLS with RBF networks has been explained in [13], to predict glucose from NIR spectra and has shown that even extremely small deviations of the spectra cost an increment in the prediction errors. This work has introduced RMSP as a prediction evaluation factor and also has introduced the LOOCV as a calibration model. In this work, several samples have been removed from the experiment in order to enhance the RMSP which may remove additional information about the analyte of interest. The signal complication was increased in [14], to predict Glucose from NIR spectra by using Human serum albumin data set for the first time and also the effect of different hardware on the prediction stability by using different spectrometer than the one used in the previous researches.

In [15], Arimoto et al have described a linear relationship between the temperature and the glucose absorption. The peaks wavelength of the water absorption are shifted to a shorter wavelength when raising the temperature while reducing the temperature leads to shifting the peaks to a longer wavelength due to the change in the hydrogen bonding. This work has produced temperature – insensitive PLS regression model but no pre-processing is applied to the NIR spectra. MSC and Savitzky-Golay Derivation has been introduced in [16], as pre-processing techniques for NIR spectra yet they were not applied on glucose signal. This work has used less complicated signals and smaller sample number. In the other hand, the authors of [17] have introduced scattering effects in glucose prediction for the first time. In this work, a strong dependency has been established between the concentration level of glucose and thescattering properties of the whole blood matrix. This reliance has been described in terms of differences in the refractive divergence index between the red cells and the surrounding medium. The work has indicated that the divergence index is related to the glucose infiltration to the red cells and has demonstrated that amplified scattering relate to the acceptance of glucose by the red blood cells. This work has introduced the scattering problem but has not proposed any solution to treat the scattering effects to improve the prediction of glucose using NIR spectra. In [18], the bar has been raised into different levels by introducing new regression model MWPLS, RMSEP instead of MSEP as error prediction parameter, the correlation coefficient R and using a data set consist of HSA, Glucose and γ-globulin in a phosphate buffer solutions. This work has not mentioned any pre-processing technique on the raw spectra. In [19], a new chemometrics algorithm called SCMWPLS has been introduced to enhance the glucose prediction ability of PLS by allowing the choice of search for a specific enhanced spectral region or for an enhanced selected group of spectral regions with informative peaks. The selection of these peaks is performed by a selection method of the wavelength intervals called MWPLSR. The research has suggested that using SCMWPLS will facilitate locating informative peaks for designing accurate regression model to predict glucose from the samples of bovine serum and human skin. This method has evaluated the behaviour of SCMWPLS two datasets, the first dataset used chemical samples of bovine serum in the NIR spectral region whereas the second dataset used human skin samples in the same spectral region, yet again no pre-processing techniques employed in this work. In [20], a comparison between Raman and NIR has been introduced to predict glucose from different sixty samples of a chemical mixture consist of glucose, lactate and urea in aqueous phosphate buffer using PLS. This work has proved that PLS can produce SEP from NIR lower than Raman which is better prediction yet no pre-processing what so ever was employed in this work. While in [21], Chen et al have used PLSR to compare between the combination and the first overtone absorption regions to predict glucose from NIR spectra. The chemical mixture consists of glucose, lactate, urea, ascorbate, triacetin and alanine in aqueous solutions. This work has not performed any pre-processing technique to the raw spectra. In [22], a new multivariate calibration model called ACLS has been suggested based on the combination of obvious linear preservative model with the analytical power of reverse models, such as the PCR and PLS. The work has been done on two experiments and has shown prediction enhancement to the regression models yet no pre-processing was administrated and the glucose experiment used powders instead of liquids. In [23], the authors has introduced the coupling of PCR with two band pass filters. In this experiment, 90 data samples extracted from a mixture of glucose, urea and triacetin in a phosphate buffer. The work has been implemented in a non-controlled environment or sample conditions. The filters were implemented as Chebyshev filter for different orders (1st, 2nd and 3rd) and as Gaussian bandpass filter. This work has used a high number of PC up to 17 for Gaussian filter and 22 for Chebyshev filter which could lead into model overfitting. Only three samples were used for validating the model which is very low number leading to poor prediction accuracy. In [24], a new technique called ICR has been suggested to enhance the glucose prediction by combining ICA, MLR and PCA to predict the glucose concentration levels in the NIR spectral region. The ICA algorithm is performed using the scores of the PCA model then the calibration model is built by providing the MLR with the estimated mixing matrix. There is no need to calculate the ICA and PCA using the test dataset to predict the glucose in this model which gather all the mentioned algorithms advantages. Still the number of PC, IC and LV are high and leading to model overfitting also no pre-processing has been employed in this work. The authors have managed to add a Pre-processing technique in [25], by proposing a new method called SDICA based on coupling digital bandpass filtering with the regression model ICR which was introduced previously to improve the glucose prediction from NIR spectra. In this work, a 35% reduction in the SEP has been obtained using the proposed model with 10 subbands. In [26], the FSD-PLS model has been introduced which sums the advantages of the both the FSD where the generation of the principal spectral components and the PLSR where the calibration model is built to predict glucose concentration from new datasets. The proposed model optimum parameters were obtained using the grid search optimization algorithm. This model has significantly enhanced the prediction accuracy of the regression model not only for original data but also for new datasets even if the new dataset has slight difference from the training dataset. The proposed model also has been modified to remove the variation in the baseline from NIR spectra with a discussed comparison to the PLSR model with 2nd derivative. A new regression model called LW-PLSR has been suggested in [27], to use as an alternative multivariate calibration model instead of PLS and PCR to enhance the glucose prediction from NIR spectra using the same data in [21]. The proposed model used 1st derivative as a pre-processing technique yet no mention for the number of LV and PC used to determine the calibration model. In [28], another model has been introduced called LS-SVR by coupling the least square with support vector machine regression. This work has used the same pre-processing techniques in [25] and also no mention of the number of the optimum factors employed to design the calibration model. The effects of three pre-processing techniques have been discussed in [29]. These techniques are 1st derivative, GDBPF and CDBPF and have been tested with different regression models such as LLER, PCR, PLSR and SVR to enhance the prediction of glucose from the same dataset used in [21] which is not complex enough. This work has shown the superiority of LLER coupled with CDBPF to give the glucose best prediction. A novel technique has been introduced in [30], called FSDR-PLSR based on coupling FSD with RreliefF algorithm to improve the prediction performance of the PLSR model using NIR spectra. The FSD is well known for attenuating the variation in the baseline and the noise from high frequency components whereas the RReliefF is proposed as a weighting algorithm for the features. The proposed model has been tested to predict glucose concentration from a chemical mixture consist of glucose, urea and triacetin dissolved in phosphate buffer solution. The chemical components levels were chosen to be similar to the levels of the human blood. The proposed model predicted results shows that it has better higher concentration levels of glucose which indicate higher prediction accuracy.

**2.2 MIR spectroscopy**

MIR spectroscopy has been introduced as a sensitive and very selective technique to obtain the levels of concentration for an identified material, and the representative immersion band in the MIR region facilitate the distinctive identification of a targeted material, such as glucose. In the existence of several chemical substances glucose has a strong and sharp fundamental frequency vibration the MIR spectrum, analysis of this band has been vastly investigated for prediction of glucose [31]*.* The MIR region wavelengths start from 2500 nm to 16000 nm. Below are some advantages for MIR spectroscopy [31]:

1. Data calibration in the MIR is more general than what in the NIR and hence is more easily transferable between instruments.
2. Organic groups have well observed absorption band characteristics in this spectral region.

MIR is suitable for monitoring process using the increase or decrease of a functional absorption band, determination of a particular component in a mixture such as the trans fatty acid concentration in vegetable oil, level of alcohol in beverages; general gas analysis [31].

**2.2-A Glucose sensing using mid infrared spectroscopy**

In [32], MIR has been used to predict the concentration of glucose from a chemical mixture consist of serum solutions with glucose levels varies from 50 to 500 mg/dL. The author has used PLSR to build the prediction model using a part of the dataset as training set and the rest as testing set. In this study accurate results has been obtained up to 30 mg/dL no matter what the buffer solution was. The author in [33], has indicated that few groups have demonstrated that FT-IR spectroscopy can be used in the MIR range for glucose detection. The work done in [34], has clarified that due to viscosity and high particle content, and high water background absorption of blood, MIR transmission measurement with FT-IR spectroscopy has been adapted by few research groups. Mid infrared spectroscopy has been chosen in [1], over several techniques such as near infrared, optical rotation of polarized light, near infrared spectroscopy and radio wave impedance due to that it does not require any external energy source to the human body. In [35], the glucose concentration is obtained as a result of converting the collision of glucose in the blood with the applied radiation or thermal emissions in the case of mid infrared spectroscopy. The search mentioned above explains that glucose has strong absorption in the mid infrared region without a noticeable interference from other absorbing compounds except water which also has a very strong absorption in the same region, the research also recommends mid infrared spectroscopy due to its cost efficient, non-invasive specific and rapid. The reason of choosing mid infrared in this work was due the excellent emission of mid infrared light from the human body at the desired spectral region and that means the required radiation exists within the body itself and adjacent to the surface approximately 10’s of microns therefore it is possible to measure glucose from its emitted spectrum. The work in [35], used few number of samples and did not discuss the light scattering effects also no Preprocessing techniques has been implemented. In [36], David C. Klonoff et al used the mid infrared region to predict glucose from several mixtures including but not limited to aqueous glucose solutions and whole blood using the wavelength range from 9564 nm to 9694 nm, this experiment detected an optimum wavelength for glucose at 9676 nm using a aqueous glucose solution range from 149 to 670 mg/dL and with glucose concentration up to 600 mg/dL were a linear dependency has been shown. Despite measuring blood glucose using photoacoustic noninvasively in several mixtures such as gelatine phantom, plasma, aqueous solution and whole blood, yet the overlapping between other components especially water still present and affecting the results and no mention to pre-treating the collected data from all associated noises. In [37], A study have been done using MIR spectroscopy to investigate the possibility of basis formation to quantify the levels of glucose and urea in whole blood using dry films of whole blood diluted in a solution of potassium thiocyanate. This work used repeated measurements for a small number of samples and did not include the effects of light scattering or its correction never to mention that no pre-processing has been applied to the collected spectra which means lower signal to noise ratio also there was no mention to the levels of glucose concentration for the used samples. Another study discussed the use of MIR spectroscopy in [38], and introduced the range from 2500 nm to 10000 nm as the best region for predicting glucose because it includes a highly selective fundamental absorption of glucose compared to the invisible light or the NIR region. The experiment used 78 samples of pooled blood with a glucose range varies from 17 to 458 mg/dl and implemented Partial least squares Regression to predict the glucose from the MIR spectra and it concentrated on the effect of haemoglobin on the prediction process of glucose. The work done by Yoen-Joo Kim in [38], final selected a specific wavelength region that represent glucose concentration and in the same time has lowest interference with other compounds such as haemoglobin and water, the chosen region in this work lays between 8936.55 nm to 9784.74 nm, despite the author mentioned the use of pre-processing techniques in his literature yet no use of any Preprocessing techniques in the experiment especially light scattering techniques and there was no mention to the calibration techniques employed in this research. Omar S Khalil has published a review paper in 2004 [39], discussing the none invasive glucose monitoring from 1999 till 2004 the paper discussed the mid infrared spectroscopy in detail and clarified the difference between mid-infrared and near infrared spectroscopy, the paper stated that the wavelength range between 2500 nm and 10000 nm mainly represent frequencies of fundamental molecular vibrations, these fundamentals are characteristic of the specific chemical bonds, unlike the near infrared spectral range which include broad overtone bands and weak combinations the mid infrared encompasses sharper bands and have coefficients with higher absorptivity. The bands of some carbohydrate and glucose in the mid infrared spectral region are dominated by bending vibrations and stretching of OH, CH and CC bonds. The spectral region between 8000 nm and 12500 nm representing glucose bands in the following wavelengths (8000, 9293, 9551, 9801, 10976, 11961) nm, these bands has been allocated to C-H bending vibrations, while the wavelength band of 9746 nm represent the C-O-H bending vibrations these spectral ranges has been used to obtain spectral measurements of glucose in blood and serum. In the other hand Khalil [39], discussed the importance of the Light Scattering effects in the mid infrared region from 2500 nm to 10000 nm on the tissue and how it differs from the near infrared region, the relationship between light scattering and wavelength length is reversed so the longer the wavelength the scattering is lowered leading to attenuation in light by the dominating absorbed spectra of proteins, water and fats. A new method to predict glucose from mid infrared spectra using photoacoustic detection of absorption combined with derived radiation of Quantum Cascade laser QCL has been introduces by Hermann von Lilienfeld-Toal et al in [40], this experiment has been done both in vivo and in vitro and gave some detailed information about the structure of the outer layers of the human skin and the absorption ability of each one, there are 3 main layers of the human skin, the first outer layer called stratum corneum which has a thickness of 80 to 100 micro meter, the second layer called stratum granulosum and the third layer known as stratum spinosum. The paper argued about the glucose concentration contained in each layer and its proximity to the glucose levels of the actual blood. The paper managed to reduce the number of required wavelengths to predict glucose noninvasively yet it did not discuss the associated noises affecting the collected glucose information especially scatter correction techniques. Another study in [41], introduces the use of Mid infrared region to predict glucose using Quantum Cascade laser but this time combined with an optical silver halide fiber, the experiment allowed the spectroscopy transmission of aqueous glucose solutions by implementing a gap in the fiber with a diameter of 30 mm, the used mid infrared wavelength was cantered at 9690 nm due to its relativity to the glucose absorption vibration. In this experiment a low concentration of noise equivalent of 4 mg/dL has been achieved with 50 second integration time and at 1.8 mW of average power. There were 10 samples of different glucose concentrations obtained from dissolving phosphate buffer in deionized water, these concentration levels are 40, 70,100,150,200,250,300,400,500 and 600 mg/dL with a fixed pH level at 7.4. The study identified the drift in baseline caused by other compounds variation as a major interference with the glucose signal and suggested an extra QCL to correct it, yet no mention to the light scattering effects and the number of test samples were too low and require additional coverage to the glucose levels in order to enhance the accuracy of obtained levels of the glucose concentration. In [42], the mid infrared fingerprint has been used to predict glucose using an quantum cascade laser with external cavity ( EC-QCL) to produce high pulse energy tuned in the fingerprint region of the glucose (8196-10000) nm, the laser was focused on the skin, then combining this method with a photoacoustic spectroscopy in the ultrasound region to detect the absorption process with the help of a gas cell attached to the skin. The combination of these two techniques enabled the quantitative analysis techniques such as PCA and PLS to detect glucose from the human skin in the range between 50 to 300 mg/dL, which consider an acceptable range for the monitoring of diabetic patients. In the experiment above the authors used interstitial fluids to predict the glucose levels due its representation in the blood and low delay in time required to follow it about 10 minutes, hence the assumption made that it is possible to achieve the noninvasive glucose prediction which the diabetic patients are anguish to for a long time. Also the authors mentioned the need to operate some calibration on each patient due to the different in skin characteristics between the patients, yet no mention to scattering effects has been given in this study. The university of Uppsala Sweden published a work for a PhD student named Louise Nybacka in [43], the work introduces the use of mid infrared spectroscopy to predict glucose using a portable hand held spectrometer, the work identified the absorption bands for each fundamental bond of the glucose in the mid infrared region. The stretch of OH bond absorbed at 2985 nm, the asymmetric vibration of CH is absorbed at 3424 nm, the asymmetric vibration of CH8 is absorbed at 3508 nm, the bending vibration of CH is absorbed at 6896 nm and CO is absorbed at 9661 nm. The paper discussed the shift in baseline as an effect on the absorption spectra and managed to sweep the range of glucose concentration between 100 to 5000 mg/dL yet no mention to any Preprocessing techniques or complication has been discussed in this research. Another study by Jonas Kottmann et al [44], introduces a two techniques using mid infrared spectra to predict glucose from human skin with the help of the test of oral glucose tolerance. The first technique was using a tuneable quantum cascade laser QCL with fiber coupled photoacoustic (PA) while the second technique was using two quantum cascade lasers at a different wavelength and then combined them with a PA detection, both techniques used direct contact between the skin and the PA. The study results was that the second technique provides better glucose prediction with improved stability and uncertainty with about ±30 mg/dL and with a level of confidence up to 90% without the use of PCR or PLS yet no mention of the overlapping or the baseline shift or their treatments. In [45], a study by Kasahara et al used mid infrared spectroscopy to predict glucose none invasively using spectral characteristics of oral mucosa from human volunteers. Two data sets has been used in this study, the first data set consist of 131 samples from 13 series of measurements obtained over a period of five months from one healthy adult volunteer who has been given different meals prior the experiment, while The second data set consist of 414 samples collected from 18 measurement series collected from five healthy adult volunteers over a time period of 15 month, the test participants has been given either glucose solution consist of 75 g of glucose powder dissolved in 150 ml of water or a variety of meals. This study used the first dataset as calibration set and the second dataset as testing set to ensure the occurrence of correlation between calibration and testing set from different subjects and has different acquiring conditions. This study proved that using multivariate linear regression using three wavelengths can give a similar correlation coefficients to the ones obtained by using PLS with higher number of latent values, the multiple linear regression model used only 3 wavelengths 9090, 9523 and 9345 nm with a time delay of 20 min, these results were inaccurate compared to previous results and also the data set require expansion. The future work proposed using machine learning algorithms to enhance the results but did not discuss importance of the baseline shift or other effects and their treatments. The use of bandpass filtering has been introduced to reduce the noise in [46], and pre-process the data in order to predict glucose from human skin using mid infrared spectra in the spectral range from 8064 to 11111 nm, the study used both PCR and PLSR at the beginning then PLSR proved to its ability to predict glucose slightly better than PCR, thus only PLSR has used for the rest of the experiment. The data used in this study consist of 76 samples collected from 5 healthy and diabetic volunteers, the use of PLS cross validation has been introduced in this experiment to validate the obtained results. The idea of this experiment was based on scanning non secretion areas in the volunteers hand instead of secretion areas with maintaining constant temperature at 24₀, the identification of the skin condition has been achieved with the help of spatial information collected by a microscope before the spectroscopic measurements using the same laser used in the spectroscopy itself. The study proved that the prediction of glucose from none secreting areas is better than the obtained glucose from ca secretin areas and provided better prospective regard reducing the effects of variation from day to day or person to yet perform worse than other techniques such as Raman and photoacoustic spectroscopy and still no detailed discussion about the associated noises and their treatment.

**2.3 Scatter correction**

Paul Geladi, D MacDougall and Harald Martens introduced in their experiments in [47] [48], the use of multiplicative scatter correction as a pretreatment method with the help of partial least squares to predict fat from pork meat by isolating the chemical light absorption from the physical scattered light. This experiment were implemented on a linearized spectral data of meat fat with a wide concentration range, the obtained coefficients of MSC indicates that fat introduce to additional components to the meat near infrared spectra the first component is the absorption components and the second one is the multiplicative scatter component. In [49], five different NIR data sets were used to predict fat, protein, moisture and sucrose in different wheat and biscuit powders. The aim of the experiment was to compare between using MSC with and without linearized datasets, in this study the used models did not improve the prediction as much as it gave a deep insight of the importance of the use of normalization with MSC to provide a simplified model by revealing important secondary correlations to the model. The use of MSC and principal component has been introduced in [50], for near infrared spectroscopy using two and three principal components. The focus of the paper was on outlier treatment and the difference between linear on none liner approaches using MSC as a pretreatment method, three data sets has been used in this experiment and the results indicated that MSC plays an important role in enhancing the spectra before using the calibration model and that leads to a better prediction in general, also the experiment concluded that between linear and none linear approaches there is no preferred one. In [51], MSC has been combined with partial least squares to predict protein, moisture, water absorbance, falling number and wet gluten of wheat flours collected from commercial and roller wind mills. Two data sets has been used in this experiment the first consist of forty samples from nine different Finnish wheat breeds and the second one consist of forty five wheat flours from Greece. The link between multiplicative scatter correction MSC set-independent and standard normal variate SNV set-dependent has been introduced in [52], by M Dhanoa et al, in this paper Dhanoa practically proved that MSC and SNV are related linearly to each other as expected theoretically. A comparison between MSC and other scattering methods such as SNV, ITMSC and ISC has been introduced in [53], to predict fat from three meat NIR data sets and the results indicated that although they are all provide relatively similar prediction but MSC is the most preferred choice to eliminate the effects of light scattering in general. Light scattering of near-infrared spectra has been compared to orthogonal signal correction in [54], to predict measured viscosity from three cellulose NIR data sets and the forth data set was for pulp NIR spectra. MSC has been used to preprocess 96 samples of pork longissimus muscles to predict shear force and drip loss with the help of PLSR in [55], this experiment used the spectral range from 1000 nm to 2500 nm and the data also smoothed using 19 point window Savitzky Golay filter. Light scatter correction methods has been discussed also in [56], to predict soil characteristics from MIR spectra with the help of PLSR. In [57], the light scattering techniques has been compared to a new proposed algorithm named optical path-length estimation and correction OPLEC to predict characteristics from two NIR data sets, the first data set consist of 415 samples of wheat kernels and the second data set consist of twenty replicate samples from five mixtures of starch and gluten powder. In [58], the light physical characteristics were used to perform a transformation on the light scattering techniques to enhance the performance of PLS using the NIR gluten starch data sets from [59]. The online use of MSC has been proposed in [60], to pretreat and assist in online sensing of 300 NIR soil samples. A detailed explanation of the most common light scattering methods the MSC and SNV were discussed by Tom Fearn in [61], to describe their similarities and characteristics, the paper concluded that despite MSC and SNV gave the same results yet they reach these results via completely different approaches. The experiment applied both scatter correction methods to NIR imaging data and then compared the output mean spectrum resulting from each case. Achim Kohler et al has published a full chapter in [62], detailing the importance of light scattering correction methods such as MSC and SNV in treating several imperfections affecting the spectroscopic data and their combinations like baseline, scaling, additive, multiplication and instrument induced unwanted channel shift. In [63], a comparison study between several preprocessing methods has been done by Fei Liu, SNV was one of these preprocessing methods to determine the content of soluble solids of beer using 360 NIR samples, the samples were divided into three data sets, 180 samples as calibration set, 90 samples as validation set and 90 for prediction set. The aim of this experiment was to compare between linear and nonlinear calibrations using three wavelength selection methods combined with partial least square regression. Asmund Rinnan introduced a very detailed theoretical and practical discussion about several preprocessing methods to treat the NIR spectral data [64], these methods were divided into two groups according to their approach of treating the spectra, the first group is the scatter correction which include MSC, SNV, Extended MSC, Inverse MSC and de-trending, the second group is the spectral derivation group which consist of two filters Savitzky Golay and Norris Williams. The authors used three NIR data sets, the first one consist of 13 samples of sucrose powder and the second data set consist of 32 marzipan samples measured with six different spectrometers and the third consist of 7 pectin samples with a spectral range varies from 1100 nm to 2500 nm. The use of MSC and SNV to predict the acidity and the content of soluble solids from kiwi fruits has been introduced in [65], by Ali Moghimi, Mohammad Aghkhani, Majid Sarmad and Ameneh Sazgarnia, the light scatter correction methods were combined with a median filter and quantitative analysis algorithms PCR and PLSR in the spectral range from 400 nm to 1000 nm and the root mean square error of prediction RMSEP was the factor identifying the level of enhancement in regression model prediction performance. Standard normal variate SNV were used to preprocess NIR spectroscopic data in [66], after been combined with PLS for purpose of online monitoring of the soybean oil reaction progress of catalyzed transesterification that used in biodiesel production, RMSEP was also used to determine the prediction ability of the used regression model. In [67], SNV and MSC has been used to preprocess NIR data study the soluble solid contents of Valencia oranges and its relation to the test of these oranges, the test used PLSR and PCR regression models and the determinant value was the RMSEP. The Savitzky Golay smoothing filter has been used in this experiment as well, the data set used in this paper consists of 120 orange samples from Valencia purchased on different days for 10 days and stored in controlled temperature of 25 degrees and the spectroscopic range varies from 350 nm to 1050 nm. A review of the most common preprocessing approaches to pretreat IR spectra has been widely discussed by Jasper Engel et al in [68], this review presented three preprocessing approaches to pretreat the infrared data in general. The first approach is trial and error where different approaches of preprocessing are applied and then the one with best performance is being chosen according to data analysis requirements, the second approach is preprocessed data assessment by the parameters of data quality which targets the measuring of artefact presence in the data, finally the third approach is based on visual inspection which is done by visually inspecting each desired preprocessing method to identify any presence of artefact within it. In addition to the mentioned above the authors explained the most common factors associated with IR spectra extraction which strongly affects the acquired spectral data, these factors are noise, baseline slope and offset, spectral and temporal misalignment and finally light scatter. the data used in this experiment explained in details in [69]. Huazhou Chen et al proposed a new preprocessing approach in [70], based on two preprocessing steps, this approach suggested the joint combination between Savitzky Golay SG smoothing filter and light scattering method multiplicative scatter correction MSC to enhance partial least squares prediction performance to predict soil characteristics from 135 NIR soil samples from china, the measurement parameter is RMSEP and the spectral range varied from 1000 nm to 2500 nm. In [71], Willem Windig and his coauthors has proposed a preprocessing method based on multiple repetition of MSC to treat NIR data collected from [59] the experiment used two data sets, the first set is NIR data for wheat gluten and the second data set is octane in toluene Raman spectra. The authors performed a loop on MSC where the output spectrum is being fed again to the MSC function and hence come the name loopy MSC also the authors concluded that a two-step repetition of MSC is the best choice which provided the best results. The use of scatter correction methods such as standard normal variate SNV and multiplicative scatter correction MSC has been discussed in a review paper about vibrational spectroscopy and its application in food industry by Santosh Lohumi and his colleagues in [72], the review paper outlined the use of MSC and SNV as one of the most common preprocessing techniques used in the NIR and MIR spectroscopy. Another review paper discussing the NIR spectroscopy online applications in food industry published in 2015 by Jan Porep, Dietmar Kammerer and Reinhold Carle [73], this paper presented MSC and SNV as an important preprocessing techniques used with PLSR to pretreat different NIR spectral data sets and enhanced the prediction ability of the regression model each time. The use of a localized version of SNV has been proposed by Yiming Bi et al. In [74], to preprocess four NIR spectral data, first data set has 215 samples of meat with a spectral range varies from 850 nm to 1050 nm and two nm of separation between samples the calibration set contains 129 samples and the test set contains 86 samples, the second data set has 310 samples of pharmaceutical tablet with a spectral range varies from 700 nm to 2500 nm with a spectral resolution of 16 cm-1 the data separation has been done using Kennard stone algorithm where 217 samples (70%) of the data set were assigned to the calibration set and 93 samples (30%) of the data set were assigned to the test set, the third data set consist of 100 samples of wheat with a specified level of moisture and the spectral range varied from 1100 nm to 2500 nm and 2 nm separation between samples and the fourth data set consist of 141 samples of NIR spectra for fescue powdered samples of grass with specified ranges of nitrogen and carbon contents, the prediction set consist of 43 samples and the rest assigned to the calibration set. The combination between MSC and first derivative has been proposed in [75], to enhance the prediction ability of support vector regression SVM to predict nitrogen content in apple tree canopy, two spectral ranges were used in this experiment with different sampling intervals and resolution the first range varies from 350 nm to 1000 nm with spectral resolution of 3 nm and sampling interval of 1.4 nm while the second spectral range varies from 1000 nm to 2500 nm with a spectral resolution of 8 nm and sampling interval of 2 nm. The latest most general review paper about NIR spectroscopy published in 2018 by Celio Pasquini in [76], introduced the light scatter correction techniques MSC and SNV as most popular and effective pretreatment methods in NIR spectroscopy.

**2.4 Savitzky and Golay filtering**

In 1964, Savitzky and Golay introduced the procedures of a simplified least squares as a vital tool for differentiation and smoothing of data for the first time in [77], which is known now as Savitzky Golay Filter (SG), It has been introduced to overcome noise and it was more efficient compared to other known techniques such as moving average due to the peak intensity degradation of the signal despite the effectiveness of moving average in noise removal. Despite the enormous number of citation in several journals the Savitzky and Golay paper has some numerical errors, these errors were mainly in the published tables which has been addressed and almost corrected completely in [78], by Steiner, Deltour and Termonia in 1972. The maximum number of smoothing points in the Savitzky Golay original paper was 25 points and this has been extended in 1978 by Maddin and Hannibal [79], who introduced new equations to check for the arrays errors and provided examples explains the advantages of using the smoothing techniques in information extraction in addition to height, position and width of the peak from spectral data. In 1984 Keisuke and Keiichiro discussed the effects of the polynomial degree of the Savitzky Golay filter in enhancing the noise associated with the second degree derivative in [80], they explained that the second degree derivative provide good performance yet generate additional distortion noise when using Savitzky Golay differentiation with 17 points of data, this generated noise is related to the width and high of the peaks, its trivial to the peak width and reverse to the peak height. The paper concluded that the Analogue to digital conversion process will definitely add noise to the collected data even if the original spectrum has little or none depending on the width of the peaks and the second-degree derivative can treat some of this noise. This generated noise is increased if a quantic polynomial convolution is used to enhance the shoulders of the spectral peak, unlike the cubic polynomial convolution which reduces the generated noise except for sharp peaks. In 1990 another 2 drawbacks of Savitzky Golay original paper has been introduced in [81], which is the treatment of 2m missing points for filters of 2m+1 points and the special case discussing the starting point values. The proposed method is based on replacing the least squares polynomial with the Gram polynomial recursive properties, this proposed model enables the coverage of all spectrum points with the convolution method. The use of Savitzky Golay as a smoothing digital filter has been discussed in detail in [82], by Press and Teukolsky; they established that Savitzky Golay filter is able to perform smoothing within specific limits without the loss of resolution. This has been done by assuming that the equally distributed data points preserve a usable redundancy, which is able to decrease the noise level. A special case for the underlying function of this assumed redundancy is that it requires a polynomial to fit it well locally. If this special case is confirmed, such as the case of line profile wider than the width of the filter, that’s when the Savitzky Golay filter performance become optimum, but when the line profile is not wider than the filter width, then these filters have no fascinating advantage over the smoothing filter coefficients of other classes. In [83], King and Ruffin introduced the idea of detecting the positions of band absorption using Savitzky Golay filtering in the hyperspectral reflectance domain, they discussed the importance of Savitzky Golay in smoothing and de-noising spectral domains in addition to its ability to preserve the original information which includes the desired width and height of the absorption bands. The authors also mentioned the effectiveness of the Savitzky Golay filter compared to its simplicity due to the use of a set of filter coefficients in addition to a linear convolution. They also highlighted the ability of these filters to calculate derivatives in addition to smoothing inside the same process of convolution. The authors stated that for any integer S the noise can be attenuated optimally by a 2S order Savitzky Golay filter while preserving up to the 2S+1 order for all the original spectrum moments, which considers a very interesting characteristic of Savitzky Golay filters because it ensures the removal of a maximum amount of noise while keeping the desired and important characteristics of the spectra like the area below the spectral features and the spectral features mean location across the spectrum. In [84], Savitzky Golay filter has been introduced by Chen et al as a better filter than 3 different pre-processing techniques such as the methods of threshold-based which include BISE best index slope extraction, and the fitting method of Fourier based, the methods of fitting asymmetric function which includes the fitting function of the asymmetric Gaussian approach and the approach of linear regression based on weighted least squares. All these used techniques have their own advantages which have been proved by applying them successfully to the NDVI (Normalized Difference Vegetation Index). Despite these pre-processing methods has been successful in treating the NDVI yet each one has its own drawbacks, for example when applying the method Fourier based fitting to the NDVI data whether it's symmetric or irregular it could cause problems due to their critical dependency on the symmetricity of the sine and cosine functions. In addition to that, they might produce higher vibrations to the time series of NDVI. Savitzky and Golay have been introduced in [8], as a simple yet robust solution for the noise reduction drawbacks of the three techniques previously used to treat the NDVI data as mentioned in the proposed work. The characteristics of the Savitzky Golay digital differentiation SGDD has been discussed and developed by Luo and Ying in [85], as they studied its effects on A single Gaussian noise free line and doublet Gaussian noise-free signal for first-order differentiation. The results of this study showed a dependency on the SGDD loss resolution and contrast on the length of the filter and the signal derivative width. This study also explained that with the proper choice of the filter length quantic and cubic derivative Savitzky Golay is recommended to maintain the signal derivative resolution. The use of Savitzky Golay in ECG has been discussed in details by the authors in [86], and the result was that differentiation and smoothing Savitzky Golay quartic filter with the length of 17 points is the best choice to be used in ECG signal processing. In [87], a new method for Savitzky Golay filter has been introduced by the authors, this method solves the problem of using an odd number of data subsets and enables Savitzky Golay to use an even number of data subsets. The first introduction of Savitzky Golay window size optimization has been done in [88], by Truyols and Schoenmakers who developed an experimental design to enable them to identify the retrieved signal characteristics and then studied the effects of changing the size of the smoothing window for Savitzky Golay filter. They concluded that it is very difficult to remove the noise from a smoothed signal without removing some of the required information about the interested substance, therefore a balance should be established between the noise elimination and the distortion of the signal and hence, they proposed a method based on comparing the instrument noise with the fitting residuals ( difference between the smoothed and the original signal) and the considered optimal smoothed function will be the one which generates the closest autocorrelation residual to the corresponding blank signal autocorrelation. This method has been applied to optimize the size of the Savitzky Golay smoothing window but it could be generalized to any filter yet the Savitzky Golay was used here due to its simplicity and the providing of signal derivative without any extra calculation cost. The robustness of the proposed method has been tested by the authors in [88], the performed experimental design has been done in a way that the amplitude of the noise, the sampling frequency, the width of the peak and the autocorrelation of the noise were regularly varied. The use of 2-dimensional Savitzky Golay differentiator Digital filter has been proposed by Pan et al in [89], to assist with the estimation of placement accuracy in image processing based on the fitting of a least squares principals using a polynomial of 2 dimensions. It was easy to implement the calculation process which was basically a convolution between the fields of the estimated displacement with the Savitzky Golay differentiation filter, the proposed method has been verified by using inhomogeneous and homogenous images which have been deformed. The proposed model tends to be efficient and simple. A comparison between Savitzky Golay filtering and finite impulse response filtering for reduction in Raman spectroscopy has been introduced by Clupek, Matejka and Volka in [90], they discussed the similarity between the two filtering techniques in removing the background noise and the ability of finite impulse response filters to preserve the signal original form yet the computational time is a drawback compared to the Savitzky Golay. The joint combination of Savitzky Golay differentiation and smoothing filter with partial least squares has been introduced for the first time by Delwiche and Reeves in [91], to identify the effects of different degrees of the derivative on the noise removal from NIR spectra, the study explained that the scattering medium is been treated by the second degree derivative while the sloped baseline is treated by the first degree derivative and the offset noise is dealt with by the zeroth degree derivative. After removing the noise with the help of derivation a careful choice for the regression model parameters should be taken in consideration to avoid miss interpretation for the predicted results especially for the following cases, the analyte of interest ability to be interpreted is weak, the number of calibration set samples is less than 50 and the use of the multivariate determination coefficient R2 is been used rather than the regular terms of error based residual. Another discussion about the coupling of Savitzky Golay and partial least squares and their optimization has been discussed in the same year by Jun et al in [92], to predict glucose serum from NIR spectra. The wavelength range in two bands varies from 1887 nm to 10000 nm and from 2032 nm to 2400 nm. The number of smoothing points of Savitzky Golay filter has been varied from 5 to 87 oddly spacing by 2 points and the degree of the polynomial has been extended from 2 to 6 respectively and as a result, this work calculated 582 smoothing mode using up to 14 smoothing coefficient tables and all PLS coefficients from1 to 40 related to these tables have been constructed. The parameters of the calculated optimum models by the effect of prediction were as follow the polynomial degree 3, the derivative order 1, the number of smoothing points 53, the optimum number of components 7 and the optimum RMSEP 0.376 mmol/L. In [93], the authors suggested Savitzky Golay filter as a preferable solution to treat the ECG signal because its ability to remove the noise without reshaping the original signal, they proposed a variation in the polynomial degree and the size of the frame to identify the optimum filter that can DE noise the ECG signal and they proposed an adaptive Savitzky Golay filter as a future work to adaptively select these parameters and generate better ECG signal. The proposal of a digital fractional differentiation Savitzky Golay filter has been introduced in [94], by Chen and Yang who suggested using a fractional order derivative filter instead of the integer order to treat a contaminated signal, the proposed model calculates the weights of the moving window using the method of polynomial least squares and the fractional order derivative definition of Riemann-Liouville. In [95], the authors introduced a detailed analysis of the Savitzky Golay filter in the time domain, this analysis indicated that when an increment occurs in filter length a decrement occurs in the estimated error variance, while in the case of a constant filter length the approximation error will be reduced due to the reduction of the sample period. In the case of the existence of random noise in the sampling data and a constant filter length, a very large error estimation may occur when the sampling period is very small and very sensitive when the derivative order is increasing. The introduction of comparing Savitzky Golay filter to bandpass filter has been introduced for the first time by Chakraborty and Das in [96], to reduce the noise of an ECG signal and Savitzky Golay filter demonstrates better and more effective noise cancellation the bandpass filtering method. The optimization of the Savitzky Golay filter length has been discussed in detail by Krishnan and Sekhar in [97], from real ECG signal by using the criteria of a pointwise minimum mean squared error (MMSE) and Stien’s unbiased risk estimator (SURE) to assist with the trade-off between the bias and the variance conflict and it demonstrated powerful de noising results for both real and white noise signals compared to traditional de noising algorithms in addition to reducing the required computation time. A detailed analysis of the effects of optimizing the Savitzky Golay parameters especially the size of the smoothing window has been introduced by Zimmerman and Kohler in [98], using the coefficient of determination as an evaluation factor. The authors explained the importance of each one of the Savitzky Golay filter components in the mid-infrared spectral region, differentiation, polynomial degree, and the smoothing window size; then concentrated on the effect of the size of the smoothing window and considered it as the most important feature in the Savitzky Golay filtering process and gave detailed analysis of its effect on the determination coefficient in particular. The founding of the paper was that every optimum window in Savitzky Golay smoothing filter is associated with a specific component and may not give the same results when combined with other components in the regression model, based on that the authors recommended to optimize every component separately in order to find the associated optimum window which will provide the best prediction of the regression model. A general frame for implementing and derivation of Savitzky Golay filter has been proposed by Candan and Inan in [99], the authors outlined the importance of SG filtering in several applications and suggested a unified model for SG to be used generally for any application instead of the old modules. These modules were designed separately based on the required application while the proposed modules can be used to derive any SG filter easily. In addition to the mentioned above the authors provided a Mat lab function that can be used to assist deriving any SG filter for other applications that might not be presented in their work. In 2016, an adaptive SG filter has been proposed by Acharya, Rani, Agarwal and Singh in [100], to remove the noise associated with an EEG signal, the proposed filter optimal parameters are determined by the higher correlation coefficients and have simple mathematical model representation. The proposed SG filter removes the noise from the noisy EEG signal while keeping the signal shape unchanged and also can be implemented to different signals such as EGG, EMG, ECG and EOG. The application of Nyquist Shannon theorem for smoothing Savitzky Golay has been proposed in [101], to calculate the window size in bio-optical signals such as the case of cervical spine X-ray images and mid-IR glucose absorption.

**Chapter Three**

**Background, sample preparation, spectra collection and data acquisition**

**3.1 The research gap**

The present work will focus on investigating signal processing methods for the non-invasive detection of glucose in samples from chemical mixtures. With the use of NIR and MIR spectroscopy. In the non-invasive detection of glucose there are multiple factors that may have a negative effect on the quality of the collected spectra and hence may reduce the accuracy of the calibration model. These factors are due to the instrumental variations, in addition to the sample itself:

1. The high overlapping of the spectra of other components of blood and skin (haemoglobin, protein, fat, etc.), due to the similarity of the glucose molecules and bonds structure with many constituents of tissue and blood.
2. The high background spectra of water, which has high absorbability in the IR range due to the high concentration of water in the human body and the high absorption of OH bond.
3. The low levels of glucose concentration in the blood provides low signals compared to other constituents of blood and tissue. The normal glucose concentration in blood ranges from 65 to 120 mg/dL.
4. The resulting baseline variations from the instruments and ambient variation.
5. High-frequency noise, such as the detector noise.
6. The light scattering causing from the skin and other ingredients such as red cells. This type of scattering is anisotropic, inhomogeneous, and nonlinear. The scattering reduces the SNR as a result of the degradation of the optical signal.

The previous researches have discussed different pre-processing techniques to enhance the glucose prediction from NIR and MIR spectra by reducing or eliminating one of the drawback factors mentioned above but some factors remain untreated. Bandpass filtering removes the baseline variation and the high-frequency noise but does not treat the light scattering effects. This indicates that the predicted glucose resulting from the Band pass filtering still suffering from light scattering effects. MSC and SNV correct the scattering effects but do not treat the high-frequency noise and the baseline variance. The SG filtering has been used as spectral derivation filtering technique with a smoothing window. The purpose of the smoothing window is to compensate the degradation in the signal to noise ratio caused by the derivation process but SG does not remove the high-frequency noise and baseline variance. The number of the factors of the regression model (PC and LV) chosen in the previous researches were too high.

**3.2 Multivariate calibration methods**

Chemometrics uses Multivariate Calibration models in a wide range to extract the concentration of the desired analyte from the collected spectra of the samples. The main idea of multivariate regression is to find the strength of the relation between two factors. These factors are the absorbance spectra (response) and the concentration of the analyte of interest (predictors). The other function of MC is to extract the information that is related to the analyte of interest. In order to satisfy this goal, Multivariate methods generate a calibration model that is related to the response and predictor variables. This calibration model can be used later on to estimate the response of new variables. The MC process requires dividing the data into two sets the calibration or training set and the testing or validation set.

1. The training set is used to design a calibration model regresses the spectra of the training data against their corresponding known concentration of the glucose mixture.
2. Then the testing data is used to validate the ability of the calibration model to predict the glucose concentration.
3. Finally, the concentration of unknown samples can be predicted using the calibration model. The robustness, fitting and performance of this model can be tested by using testing data or by using external data.

Generally, the performance of the calibration model can be evaluated by metrics such as:

1. The RMSEP (root mean square error of prediction)

(3.1)

1. The coefficient of determination *R*2

(3.2)

Where represents the total number of test samples, represents the actual concentration of the test data, represents the predicted concentration and represents the sample mean concentration.

In this thesis, two multivariate methods are used to generate the calibration model: principle component regression (PCR) and partial least square regression (PLSR).

**3.2-A Principal Component analysis (PCA)**

Principal components analysis (PCA) is an essential and widely used tool in analysing data for different applications such as neuroscience, chemometrics and computer vision. The reason of its wide spread use is its simplicity and it does not require parameters when its used with overlapped data sets to extract useful information [102]. Principal component analysis reduces the size of a complex data matrix into a lower order with minimal loss of useful information. Let B be our original extracted data set, where each column represents a single sample of our data set. B is an m x n matrix where m represent the rows and n represent the columns of B. Let C be another m x n matrix connected by a linear transformation A. B is the original collected data set and C represent a new transformation of that data set. Principal component analysis can be represented in equation 3.1 shown below [102].

(3.1)

Equation 3.1 can have many interpretation because it represents a change in basis.

Mathematically is a transformation matrix transforming to and geometrically it is a stretch and a rotation which transforms to .

The rows of matrix represent a vector set of new basis to express the columns of matrix.

Where

represent the rows of.

represent the columns of .

represent the columns of .

And this mean each column of C will be equal to.

It is shown that each coefficient of is the result of the dot product between and the matching row in and in more details the coefficient of is a projection on the row of. Based on the assumption of linearity to reduce the complicity of finding the desired change of basis, the elements of the row vector of the transformation will be the principal component of the matrix . The mathematical details explaining the best choice of the principal components and the best way to re-express are described in details in [102].

**3.2-B Principal component regression (PCR)**

Principal components regression has been proposed to treat the short comes of multiple linear regression MLR, this can be done by combining the positive outcomes of MLR and principal component analysis PCA together. Two steps are required to perform this combination, step one is using the PCA to calculate the scores from the data matrix and step two is regressing the acquired scores against the glucose concentration and provide the MLR calibration model instead of the original absorbed spectra. The scores and loading vectors can be explained in equation 3-2 below.

(3.2)

Where the rows of represent the scores vector, the columns of represent the loading vectors of spectra matrix.

The calibration model of PCR can be shown in equation 3.3 below.

(3.3)

Where is the glucose concentration, is the scores matrix and represent the vector of regression coefficients, and it can be explained in equation 3.4 below.

(3.4)

From equation 3.2, and by substituting in equation 3.4 we can get.

(3.5)

Where represent the new regression coefficients which will be used to predict glucose from new datasets without the need to use PCA for the testing spectra.

(3.6)

Where is the average values of the spectra matrix columns and is the average values of the training data concentrations.

**3.2-C Partial Least Square Regression (PLSR)**

Partial least squares is essential regression tool introduced in early 1960 by Herman Wold as a constructive method for model prediction when the desired variables has a have a high collinearity and consist of larger number than the observations [103]. PLS has been used widely in chemometrics applications, PLS is very similar to PCA with a major difference which is PCR uses only the input matrix to obtain the principal components while PLS involves the output matrix as well as the input matrix to extract the desired information using new factors to describe the input matrix, these factors called latent Variables (LV). The objective of PLS is to obtain most of the useful information of the variables in the input matrix to give the best prediction of the output matrix while reducing the dimensions of the input matrix by using less number of components than the original matrix [103].

Equation 3.7 and 3.8 explains how PLS decomposes matrices of concentration and absorbance to loading and scores factors.

(3.7)

(3.8)

Where represents the scores of the concentration matrix and represent the loadings, represent the number of regression factors, and and represent the residuals. There are two popular algorithms to calculate the estimators for PLS, NIPALS and SIMPLS, this work uses the NIPALS [103] and below how the estimators being calculated.

Initially the NIPALS algorithm use the spectra matrix to set an initial vector and the estimators can be represented below.

(3.9)

(3.10)

The first score of the spectra, the output matrix and the loading factor of the spectra matrix are obtained using the estimating factors as shown below.

(3.11)

(3.12)

(3.13)

(3.14)

Where is the loading vector value from the previous iteration and is the current value.

The absorbance spectra in the case of converged loading factors can be computed below.

(3.15)

(3.16)

And we can describe the value of the final scores and weighting factors as below.

(3.17)

(3.18)

**3.3 Selecting the model parameters**

Choosing the number of PC’s for PCR and LV’s for PLSR (tuning the model complexity) is a crucial part in building a biased regression model. Figure (2.1) shows that increasing the model complexity (LV) can be achieved by decreasing the model bias or by increasing the variance [104]. The best tradeoff between the bias and the variance will decide the model complexity, it is always difficult to decide the PC's or LV's because if few PC's or LV's were selected the model will generate bias in the prediction meaning under fitted model and when selecting too many PC's and LV's the designed model will be over fitted with poor prediction ability because of the incorporated superior noise. The prediction accuracy level in a model is related to the bias (how close the prediction values to the actual glucose concentration) while the variance which shows the level of variation for the predicted values is the error estimation [105]. This step in building a regression model is very important because the proper selection of the PC’s and LV’s number was a major challenge to researchers in chemometrics who are comparing existing techniques or developing new ones. Standard statistical measures should be applied in order to evaluate the quality of the calibration model, one of these standards is the Root Mean Square Error of Prediction (RMSEP) which is a measurement of the difference between the actual and predicted values (accuracy) and thus combines both bias and variance in one term. Normally the complexity of the model is decided based on the minimization of RMSEP in the validation stage, [106-107].



Figure 3.1 choosing the optimum latent value

**3.4 Preprocessing**

The quality of the collected spectral data is a very essential element during the prediction process of the analyte concentration in the calibration models and as we mention in this chapter the collected spectra suffer from several draw backs, such as light scattering, variation in the baseline and noise. These drawbacks affect the prediction performance of the regression model hence, come the need to clean the data from these unwanted drawbacks. Preprocessing techniques can be used to treat the raw data and enhance the prediction ability of the regression model, such as derivative, mean centring, bandpass filtering, smoothing, scatter correction and bandstop filtering. These Preprocessing methods are used extensively in this thesis.

**3.4-A Data matrix mean centring**

Mean centring consider one of the essential steps to perform the regression model such as PCR and PLSR. It is used mainly for model accuracy improvement to protect the first generated regression components or latent values from being dominated by the mean. It can be implemented by rationing out the average of any column from the same column.

(3.19)

Where is the column of which is centred and represent the content of the row and the column. Scaling is usually implemented after mean centring especially when different units are use in the measurements.

**3.4-B Smoothing**

Smoothing is a technique used in a wide range in chemometrics. It aims to enhance the raw spectra signal to noise ratio. It operate by attenuating the noise components with high frequencies in the raw data. Smoothing is based on the fact that the spatial spectrum mid and low band ranges contain the biological components most important information. Several methods were derived from smoothing such as Savitzky Golay filter (SG), moving window average (MWA). In this thesis we will use SG as a Preprocessing technique to predict the glucose from NIR and MIR spectra.

**3.4-C Derivative**

One of the common Preprocessing techniques is derivative, it is usually used to expose features or information in the raw data which tend to be subtle. It enhance the resolution of the spectra appearance by locating and resolving the peaks into their original components. The peaks are located using the first derivative yet the common one is the second derivative which is used to attenuate the variation in the baseline and the background noise. The second derivative also reduce the overlapping when sharpening the peaks. The first derivative of a spectrum can be simply expressed blow.

(3.20)

Where is the sampling wavelength, and is the variables number. The first derivative for equal spacing sampling can be simplified as follow.

(3.21)

Using the same definition the second derivative can be calculated by replacing with its derivative. Derivative has a main effect on the resulting spectra which a shift in the minimum and maximum peaks, in addition the differentiated spectrum require less number of points than the raw spectra and that’s why this method require high resolution data.

The Savitzky golay smoothing filter is the source of the Moving Window Polynomial Least Squares Fitting method (MWPLSF), the derivative of the order in this method is calculated using the smoothing method on the polynomial ordinary derivative of the order instead of smoothing the data using the polynomial directly. There is no shift in the peaks resulting from using MWPLSF but it will attenuate the differentiation sides of the spectrum.

**3.4-D Digital filtering**

Digital filtering corresponds to multiple mathematical steps in the frequency domain that manipulate the spectral characteristics of the data in a controlled way to provide desirable effects. Frequency spectrum encoded signals such as NIR/MIR spectra, where the required information is preserved in their Fourier transform frequency components, are examples of data that can be enhanced by digital filtering. However, the design of an optimal filter in practice is challenging; the Grid search optimization algorithm was proposed by Arnold [108] has been adopted in many reported publications to define the optimum filter parameters.

**3.4-D.1 Digital bandpass filtering**

Digital bandpass filtering has been introduced as one of the most popular pre-processing techniques in the quantitative analysis of glucose from NIR spectra [23] [108-110]. Bandpass filtering is used to eliminate the variation of the baseline and noise present in the high frequencies of the raw spectra. Center frequency and bandwidth are the main components that define the digital bandpass filter and these two factors must be optimized to find the desired band which represents the maximum information about the analyte of interest which in our case is the glucose concentration. The raw spectra are first pre-processed using the digital bandpass filtering and then the PCR and the PLSR models are coupled with bandpass filtering using a specific number of Principal Components (PC) and Latent Variables (LV) [23] [109-110].

**3.4-D.2 Digital bandstop filtering**

Bandstop filtering are very useful in removing unwanted signals. These filters are designed with methods very similar to the bandpass filtering mentioned above. The only difference is that the frequency of the poles is the same as the frequency of the zeros [111-112]. Bandstop filtering has not been used before to predict the concentration of glucose from NIR or MIR spectra. Unlike the case of bandpass filtering, the bandstop filtering targets a specific undesired band and removes it with minimum effects on the concentration of glucose. The main parameters that require optimization are cut-off bandwidth and the centre frequency that define the desirable frequency band that has optimum information about the glucose concentration. The PCR and the PLSR models are then built using the filtered spectra with a specific number of PCs and LVs [23] [109-110].

**3.4-D.3 Butterworth and Chebyshev filters**

The Butterworth filter has a monotonic characteristic of its magnitude function of the n-pole in both the stopband and passband of the filter. Alternatively, the Chebyshev filter has a ripple in the passband. The reason for flexible ripples in the passband or stopband is to gain sharper transition between the passband and stopband compared to the Butterworth filter [111]. The filters characteristics are controlled by the filter order and ripple. The filter ripple can be in both bands the passband or in the stopband, this work will concentrate on both bands to identify the difference between these bands in order to improve the prediction capability of the regression model [23] [109-110].

**3.5 Data Preparation**

In this thesis 4 datasets has been used in different chapters of the thesis and not necessarily all datasets used in every chapter these datasets were designed in the chemistry lab of Sheffield university then after preparing the chemical samples the data acquisition part were performed in the materials labs of the university of Sheffield. Using a Frontier MIR/NIR spectrometer with Two temperature-stabilized DTGS (deuterated triglyceride sulfate) detectors: one optimized for MIR with an enhanced SNR of 15,000:1 peak-to-peak for a 5 second scan, and one suited for NIR with <10 µA RMS noise over 250 nm range for a 1 minute scan. Liquid nitrogen cooled MCT (mercury cadmium telluride) detector Optical system with KBr beam splitter for MIR for a spectral range 8,300 – 350 cm-1 at a best resolution of 0.4 cm-1, and a CaF2 beam splitter for a spectral range of 14,700 – 2,000 cm-1 at a best resolution of 0.5 cm-1.

**3.5-A Dataset 1**

This data set consist of 3 replicate spectral measurements for 30 samples of a glucose chemical mixture. The mixture consist of triacetin, urea and glucose dissolved in a buffer solution. The concentration of the chemical constituents were designed to be similar to the levels in the normal human body except for the glucose which will be variable. The glucose concentration level in this mixture varies from 20 to 500 mg/dL. The urea concentration level varies from 0 to 50 mg/dL. The triacetin concentration level varies from 10 to 190 mg/dL. Three spectral measurements were taken for each sample without removing the sample from the spectrometer compartment. The wavelength spectral range spans overtone and the combination absorption bands which is from 2100 nm to 2500 nm. The spectral spacing between the measurements is 1 nm. The entire environment for this experiment was none controlled including the temperature. The collected spectra were divided using Kennard stone algorithm into two sets. The calibration dataset consist of 29 samples and the test dataset consist of 70 samples.

C:\Users\user\Desktop\Miner correction folder\chapter 3 figures\Dataset 1 plot.tiff

Figure 3.2 Represent the plot of the raw collected dataset1 which consist of 3 replicates of 30 chemical samples of glucose, urea and triacetin. The plot shows the level of absorbed spectra as a function of wavelength and glucose level concentration.

**3.5-B Dataset 2**

In this thesis, a Fourier transform NIR spectrometer was used to collect 100 NIR spectra for 100 mixtures composed of human serum albumin and glucose in a phosphate buffer solution. The preparation of buffer solution was done by adding phosphate powder into 1 liter of distilled water to prepare the buffer solution with a concentration of 0.01 M/dL and pH value was adjusted to 7.4. For each sample four NIR & four MIR spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [113]. The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose and human serum albumin concentrations of the prepared 100 samples are from 5 to 500 mg/dl and 5 g/dl respectively. The collected spectra spanned the spectral region from 2100 nm to 2500 nm for NIR and 2500 nm to 8000 nm for MIR with a spectral resolution of 1.7 nm. The collected raw spectra are divided into training and testing data subsets using Kennard-Stone (KS) algorithm such that C:\Users\user\Desktop\Miner correction folder\chapter 3 figures\dataset 2 Plot.tiffthe training set consisting of 75 spectra samples and the test set has 25 spectra samples.

Figure 3.3 Represent the plot of the raw collected dataset 2 which consist of 100 chemical samples of glucose, Human serum albumin and triacetin dissolved in a phosphate buffer. The plot shows the level of absorbed spectra as a function of wavelength and glucose level concentration.

**3.5-C Dataset 3**

In this thesis, a Fourier transform NIR spectrometer was used to collect 100 MIR spectra for 100 mixtures composed of human serum albumin and glucose in a phosphate buffer solution. The preparation of buffer solution was done by adding phosphate powder into 1 liter of distilled water to prepare the buffer solution with a concentration of 0.01 M/dl and pH value was adjusted to 7.4. For each sample four NIR & four MIR spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [113]. The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose and human serum albumin concentrations of the prepared 100 samples are from 5 to 500 mg/dl and 5 g/dl respectively. The collected spectra spanned the spectral region from 2500 nm to 8000 nm with a spectral resolution of 1.7 nm. The collected raw spectra are divided into training and testing data subsets using Kennard-Stone (KS) algorithm such that the training set consisting of 75 spectra samples and the test set has 25 spectra samples

C:\Users\user\Desktop\Miner correction folder\chapter 3 figures\dataset 3 plot.tiff Figure 3.4 Represent the plot of the raw collected dataset 3 which consist of 3 replicates of 30 chemical samples of glucose, urea and triacetin. The plot shows the level of absorbed spectra as a function of wavelength and glucose level concentration.

**3.5-D Dataset 4**

In this thesis, a Fourier transform NIR spectrometer was used to collect 100 mixed NIR and MIR range spectra for 100 mixtures composed of human serum albumin and glucose in a phosphate buffer solution. The preparation of buffer solution was done by adding phosphate powder into 1 liter of distilled water to prepare the buffer solution with a concentration of 0.01M/dl and pH value was adjusted to 7.4. For each sample four NIR & four MIR spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [113]. The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose and human serum albumin concentrations of the prepared 100 samples are from 5 to 500 mg/dl and 5 g/dl respectively. The collected spectra spanned the spectral region from 2100 nm to 3210 nm with a spectral resolution of 1.7 nm. The collected raw spectra are divided into training and testing data subsets using Kennard-Stone (KS) algorithm such that the training set consisting of 75 spectra samples and the test set has 25 spectra samples

C:\Users\user\Desktop\Miner correction folder\chapter 3 figures\dataset 4 plot.tiff Figure 3.5 Represent the plot of the raw collected dataset 4 which consist of 3 replicates of 30 chemical samples of glucose, urea and triacetin. The plot shows the level of absorbed spectra as a function of wavelength and glucose level concentration.

Figure 3.6 Preparation process of dataset 1 which include dissolving potassium dihydrogen and phosphate in distilled water then adding urea, triacetin and glucose with pH =7.4.

Distilled water

Phosphate powder

**+**

Adjusting the buffer solution pH to 7.4

HSA

**+**

Glucose

Aqueous mixture 2

Potassium dihydrogen 3.4023 g

Distilled water

Sodium mono hydrogen phosphate 3.0495 g

**+**

Adjusting the buffer solution pH to 7.4

Triacetin

Urea

**+**

Glucose

Aqueous mixture 1

Figure 3.7 Preparation process of dataset 2, 3 and 4 which include dissolving Human serum albumin and phosphate in distilled water then adding glucose with pH =7.4.

**Chapter Four**

**Scatter correction techniques coupled with bandpass filtering**

**4.1 Introduction**

This Chapter proposes a novel pre-processing method based on combining bandpass filtering with scatter correction techniques Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) to enhance the prediction capability of the linear regression models Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR) in near infrared (NIR) and Mid infrared spectroscopy. The proposed model has been compared to the pre-processing technique RreliefF, the developed regression models has been validated to predict the glucose concentration from NIR and MIR spectra of an aqueous mixture of human serum albumin and glucose in a solution of distilled water and phosphate buffer. The results confirms that the proposed pre-processing method not only improves the prediction capability of both the linear calibration models PCR and PLSR models but also achieves better results than the RReliefF pre-processing method by reducing the Root Mean Square Error Prediction RMSEP.

**4.2 NIR and MIR spectroscopy**

Near Infrared (NIR) and Mid Infrared (MIR) spectroscopy is one of the promising non-invasive approaches for the monitoring of blood glucose levels [114-116]. However, the prediction of glucose concentration from NIR spectra of aqueous mixtures remains a challenge. The two most commonly used linear calibration models in the quantitative analysis of glucose from NIR spectra are the Principal Component Regression (PCR) and the Partial Least Squares Regression (PLSR) [117-122]. However, the collected NIR spectra are affected by baseline variations, scattering and instrumental noise. Hence, the spectra should be pre-processed before applying to the calibration model [123]. Bandpass filtering is a well-known pre-processing technique in chemometrics for removal of the baseline variations and the high frequency noise components in the spectra, however the design of an optimal bandpass filter that matches the raw spectra and the regression model remains a challenge [124]. Scattering techniques are known to preserve the peaks in the spectra without degrading the signal to noise ratio of the spectra. Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) are techniques that can compensate for the degradation in the signal to noise ratio resulting from the light scattering [125]. Feature weighting methods are used to eliminate the irrelevant features in the raw spectra and can be adopted in pre-processing as a class on non-linear techniques. The Regressional Relief-F (RReliefF) is identified as a successful feature weighting algorithm due to its simplicity and effectiveness. RReliefF pre-processing has proven to be effective pre-processing methods in NIR spectroscopy to eliminate the irrelevant features from both low and high-frequency components and therefore help to improve the prediction capability of linear regression models [124]. However, such techniques tend to be complex to implement and difficult to adapt to different noise environments.

In this chapter, a novel linear pre-processing technique is proposed that combines linear ‘feature preserving’ smoothing techniques such as MSC and SNV with digital Bandpass filtering to improve the glucose prediction performance of the linear regression models PCR and PLSR. The proposed methods of Preprocessing are evaluated against the non-linear RReliefF method. Corresponding models were developed, tested and evaluated using experimental data to predict the glucose concentration from a mixture of glucose and human serum albumin in a solution of distilled water and phosphate buffer. The results obtained show not only that the proposed model enhances the prediction ability of both PLSR and PCR, but also achieves better prediction performance than RReliefF.

**4.3 Light scattering and baseline effects**

Nonlinearities are undesired scatter effects in NIR spectroscopy caused by the particles size in chemical samples or the wavelength size of electromagnetic radiation, therefore; applying suitable pre-processing techniques can eliminate these effects on a large scale [125]. Differences in the effective path length and light scattering are the main reasons for undesired systematic variation which represents a significant part of the sample total variation and can be identified as a baseline shifting (multiplicative effects) in addition to other non-linear effects [125]. During the reflectance measurement of the NIR sample, two reflected radiation can be obtained. These radiations are specular and diffusive. Sampling geometry and instrument design can minimize specular reflection as it does not contain any chemical information [64] while the main source of information in the NIR spectra is the diffusively reflected light because it reflects in a wide range of directions but the collected information will contain microstructure (scattering) in addition to the absorption of the sample chemical composition [125]. The main form of light scattering are Lorenz-Mie and Rayleigh and both are generated due to scattering of electromagnetic radiation such as density fluctuation, droplets, bubbles, small particles, cells, micro-organ cells and fibers [125]. Rayleigh occurs when the wavelength of the electromagnetic radiation is much larger than particles diameter (λ ˃ 10d), hence, it is wavelength dependent (~1⁄λ^4) [64]. In the case of NIR and MIR spectroscopy, the wavelength is smaller than the particle size and that's when Lorentz-Mie scattering is predominant and unlike Rayleigh scattering it depends on the scattering particles shape instead of the wavelength. The diffused and reflected spectrum of the samples is usually affected by noise resulting from scattering. If the glucose concentration levels in the sample is very low, the effects of spectral scattering could take over the spectral information. To reduce the effects of scattering interference, scatter correction methods must be applied to the spectral data as a pre-treatment process. Scatter correction techniques are pre-treatment methods that can separate the analyte absorbed informative and the scattered signal contained in the spectral data. They can reduce the variation in the spectral in the same set of samples due to particle size inhomogeneous. Scatter correction pre-processing methods can be represented by many techniques such as Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), Extended Inverse MSC, This work concentrates on MSC and SNV as pre-processing techniques to correct the scattering effects in order to enhance the regression model prediction of glucose.

**4.3-A** **Multiplicative Scatter Correction (MSC)**

Multiplicative Scatter (signal) Correction is one of the pre-processing techniques used in NIR spectroscopy. Mertens et al introduced it for the first time in 1983 [48] and then Geladi et al have done further elaboration on it in 1985 [47], the main concept of MSC is to remove the unwanted scatter effects from the data matrix before modeling the data in two steps [125].

1. Estimating the correction coefficients (Multiplicative and additive contributions).

=+ . + (4-1)

1. Correcting the recorded spectrum.

= = (4-2)

Where is the measurement done by NIR spectrometer on one original spectra sample, is the reference spectrum required for the dataset pre-processing. is a part of which is not modeled (residual). is the corrected spectra and, are scalar parameters varies from sample to sample.

**4.3-B Standard Normal Variate (SNV)**

SNV is an important pre-processing technique for correcting scattered Data [125-126]. In general the SNV format is similar to MSC.

= (4-3)

Where represents the required sample spectrum average value to be corrected, is the sample spectrum standard deviation. The main concept of SNV to correct a signal is very similar to MSC with one exception that no common reference signal is required, just proceding each observation on its own and separated from the remains of the set. It could be considered an important practical advantage when a common reference is not required [125]. The obvious similarity between MSC and SNV can be summarized in the following approximation

≈. + (4-4)

Where is the standard deviation average of the whole spectra and is the all spectra grand mean over.

**4.4 Experimental data preparation**

In this work, three data sets were used to test the proposed model dataset 2, dataset 3 and dataset 4 explained in chapter 3. A Fourier transform NIR/MIR spectrometer has been used to collect these 3 datasets. Dataset 2 represent the NIR spectral region, dataset 3 represent a mixed spectral region from NIR and MIR starts with 2100 nm to 3120 nm and dataset 4 represent the MIR spectral region. Each dataset consist of 100 NIR and MIR range spectra for 100 mixtures composed of human serum albumin and glucose in a phosphate buffer solution with chemical levels similar to the actual levels in the human blood, the idea was to increase the number of used samples from 30 sample in dataset 1 to 100 sample in the current datasets to provide more detailed and realistic results. The preparation of buffer solution was done by adding phosphate powder into 1 liter of distilled water to prepare the buffer solution with a concentration of 0.01M/dl and pH value was adjusted to from 7 to 7.4 in order to follow the instructions of the manufacturing company of the phosphate powder and to reach the same level in human body. For each sample four NIR & four MIR spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [113]. The individual component concentrations were selected such that these cover their physiological range in blood. The ranges of the glucose and human serum albumin concentrations of the prepared 100 samples are from 5 to 500 mg/dl and 5 g/dl respectively. The collected NIR spectra spanned the spectral region from 2100 nm to 2500 nm for dataset 2 and from 2100 nm to 3120 for dataset 3 while dataset 4 varied from 1250 cm-1 to 4000 cm-1 with a spectral resolution of 1.7 nm. The collected raw spectra were divided into training and testing data subsets using Kennard stone algorithm [127] such that the training set consisting of 75 spectra samples and the test set has 25 spectra samples.

**4.5 Discussion and experimental results**

Initially, the PCR and the PLSR models were implemented without pre-processing [128]. In this work, the libPLS (an integrated library for Partial Least Squares Regression and Discriminant Analysis) toolbox [129] (http://www.libpls.net/) was used to build the PLSR model. The PLSR model is then pre-processed using the RReliefF processing and filtering methods. The bandpass filter prediction is controlled by the center frequency C, the bandwidth W, the Ripple R and the filter order N. Both the PCR and PLSR models were developed with the different pre-processing techniques and the results obtained are detailed in Tables 4.1, 4.2 and 4.3 for data sets 2, 3 and 4 respectively. Figure 4.1 and Figure 4.2 shows the root mean squared error obtained for PCR and PLSR with different scattering techniques for dataset 2 the results show that the joint combination of the bandpass filtering with the light scatter correction techniques in the NIR spectral region can reduce the prediction error significantly and after 6 latent and 8 principal components values, the proposed model cannot provide further enhancement which indicate that this is the optimal number of regression factor required to provide the best prediction for glucose using this model. Figures 4.3 and 4.4 for dataset 3 these figures show that the proposed model of the in the combined spectral region between NIR and MIR can reduce the prediction error significantly and after 7 latent values and 9 principal components the proposed model cannot provide further enhancement which indicate that this is the optimal number of regression factor required to provide the best prediction for glucose using this model while, figures 4.5 and 4.6 for dataset 4 where the results show that the proposed model of the joining between bandpass filtering and the light scatter correction techniques in the combined spectral region between MIR can reduce the prediction error significantly and after 6 latent values and 8 principal components the proposed model cannot provide further enhancement which indicate that this is the optimal number of regression factor required to provide the best prediction for glucose using this model. Table 4.1, shows that the glucose prediction is enhanced by using the PCR and PLSR with different pre-processing techniques and the experimental results show the ability of the PLSR to predict the same Root Mean Square Error of the PCR for less regression coefficients which means reducing the complexity of the regression model. The coupling of MSC or SNV with Chebyshev has prediction accuracy by 31 % for PCR and 31% for PLSR in comparison to both the models with R-ReliefF processing, Table 4.2 shows that the glucose prediction is enhanced by using the PCR and PLSR with different pre-processing techniques and the experimental results show the ability of the PLSR to predict of the PCR for less regression coefficient indicating reducing the complexity of the regression model. The coupling of MSC or SNV with Chebyshev has improved the prediction accuracy by 39% for PCR and 40% for PLSR when no pre-processing is applied, and finally for Table 3 and the proposed technique has also improved the prediction accuracy by 44 % for PCR and 49% for PLSR in comparison to both the models with R-ReliefF.

**4.6 Conclusion**

In this chapter, a novel pre-processing technique based on the combination of scatter correction with bandpass filtering has been proposed to enhance the prediction ability of linear regression models. The proposed method has been compared to the pre-processing technique RReliefF and the developed regression models has been validated to predict the concentration of glucose from NIR spectra of an aqueous mixture of human serum albumin and glucose in a phosphate buffer solution. The results confirm that the proposed pre-processing method not only improves the prediction capability of both the linear calibration models PCR and PLSR models but also achieves better results than the non-linear RReliefF pre-processing method. The reduction in the RMSEP was not as high as expected when applying the proposed model to dataset 2 in the NIR region which can be related to the deleted useful information during the bandpass filtering treatment to the baseline effects and the noise from high frequency components. It could be also due to that our data set in the NIR region has one glucose region which is the combination absorption band which varies from 2050 nm to 2300 nm and missing the first and the overtone absorption regions which falls in the wavelength range from 700 nm to 1500 nm hence we decided to avoid using dataset 2 with SG method and focus on dataset 3 and 4 to further investigate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression model | Pre-processing |  | RMSEP (mg/dL) | Filtering Factors |
| PLSR | None | 0.914 | 47.732 | None |
| PLSR | RReliefF | 0.967 | 41.587 | K=3 |
| PLSR | Band pass | 0.99 | 34.22 | C=0.0107f W=0.0032f R=2.2dB N=1 |
| PLSR | Band pass with SNV | 0.990 | 28.335 | C=0.0107f W=0.0032f R=2.2dB N=1 |
| PLSR | Band pass with MSC | 0.990 | 28.545 | C=0.0107f W=0.0033f  R=2.3dB N=1 |
| PCR | None | 0.972 | 47.622 | None |
| PCR | RReliefF | 0.978 | 41.776 | K=13 |
| PCR | Band pass | 0.98 | 32.52 | C=0.0125f W=0.0032f  R= 1.8dB N=2 |
| PCR | Band pass with SNV | 0.985 | 32.748 | C=0.0033f W=0.003f  R= 3 dB N=2 |
| PCR | Band pass with MSC | 0.986 | 30.674 | C=0.0041f W=0.0038f  R= 2.1dB N=1 |

Table 4.1 Results for dataset 2 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

Table 4.2 Results for dataset 3 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression model | Pre-processing |  | RMSEP (mg/dL) | Filtering Factors |
| PCR | None | 0.970 | 94 | None |
| PCR | RReliefF | 0.976 | 93 | K=13 |
| PCR | Chebyshev Bandpass | 0.977 | 78.52 | C=0.8525 W=0.0050  R=2.7 |
| PCR | Chebyshev Band pass with SNV | 0.983 | 73.43 | C=0.3173f W=0.0024f  R= 0.8 dB N=2 |
| PCR | Chebyshev Band pass with MSC | 0.984 | 71.33 | C=0.8567f W=0.0107f  R= 3dB N=2 |
| PLSR | None | 0.912 | 100 | None |
| PLSR | RReliefF | 0.965 | 100 | K=3 |
| PLSR | Chebyshev Bandpass | 0.98 | 78.6 | C=0.0079 W=0.0037 R=0.7 N=2 |
| PLSR | Chebyshev Band pass with SNV | 0.988 | 67.45 | C=0.3173f W=0.053f  R=3dB N=2 |
| PLSR | Chebyshev Band pass with MSC | 0.988 | 66.10 | C= 0.3820f W=0.0027f  R=0.8dB N=2 |

Table 4.3 Results for dataset 4 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Regression model | | Pre-processing | |  | RMSEP (mg/dL) | Filtering Factors |
| PLSR | None | | 0.914 | | 35.7 | None |
| PLSR | RReliefF | | 0.967 | | 31 | K=3 |
| PLSR | Chebyshev Band pass | | 0.99 | | 25.23 | C=0.0019 W=0.0012  R=1.2 |
| PLSR | Chebyshev Band pass with SNV | | 0.990 | | 17.5 | C=0.0040f W=0.0017f  R=2.8dB N=1 |
| PLSR | Chebyshev Band pass with MSC | | 0.990 | | 17.45 | C=0.0040f W=0.0011f  R=1.3dB N=2 |
| PCR | None | | 0.972 | | 38.7 | None |
| PCR | RReliefF | | 0.978 | | 34 | K=13 |
| PCR | Chebyshev Band pass | | 0.98 | | 26.17 | C=0.0055 W=0.0011  R=2.4 |
| PCR | Chebyshev Band pass with SNV | | 0.985 | | 17.4 | C=0.0040 W=0.0017  R=2.8 N=1 |
| PCR | Chebyshev Band pass with MSC | | 0.986 | | 17.4 | C=0.0041f W=0.0017f  R= 2.7dB N=1 |

C:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\NIR Mixed\PLSR\figures\PLSR NIR mix with different scattering techniques new.tiffC:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\NIR 2100-2500\PCR\Figures\PCR NIRF with different scattering techniques new final.tiffFigure 4.1 explain the results of the regression model as a function of the RMSEP for PCR dataset 2 using different preprocessing methods.

Figure 4.2 explain the results of the regression model as a function of the RMSEP for PLSR dataset 2 using different preprocessing methods.

C:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\NIR 2100-2500\PLSR\figures\PLSR NIR with bandpass and  different preprocessing new final.tiffC:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\MIRF\PLSR\figures\PLSR MIR with different scattering techniques new 1.tiffFigure 4.3 explain the results of the regression model as a function of the RMSEP for PCR dataset 3 using different preprocessing methods.

Figure 4.4 explain the results of the regression model as a function of the RMSEP for PLSR dataset 3 using different preprocessing methods.

C:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\NIR Mixed\PCR\Figures\PCR with different scatter correction techniques new.tiffC:\Users\user\Desktop\thesi results\Chapter 3 results for scatter correction paper\matlab codes\MIRF\PCR\Figures\MIR PCR with different scattering techniques new.tiffFigure 4.5 explain the results of the regression model as a function of the RMSEP for PCR dataset 4 and different preprocessing methods.

Figure 4.6 explain the results of the regression model as a function of the RMSEP for PLSR dataset 4 with different preprocessing methods.

**Chapter Five**

**Savitzky Golay optimization coupled with bandpass filtering**

5.1 Introduction

This Chapter introduces a novel pre-processing method based on optimizing the parameters of the Savitzky-Golay filter coupled with bandpass filtering to further enhance the performance of the linear regression models PLSR and PCR to predict glucose concentration from three different data sets. Data set1 extracted from 89 samples of a chemical mixture consist of glucose, urea and phosphate buffer and data set 3 consists of 100 samples of glucose, human serum albumin and phosphate buffer to cover the mixed range spectral region and dataset 4 consist of 100 sample of glucose and human serum albumin in a phosphate buffer and cover the Mid Infrared spectral region. The proposed method is compared to the bandpass filtering coupled with Savitzky-Golay using fixed parameters pre-processing technique for further evaluation. The developed prediction models have been validated to predict the concentration of the glucose from the mentioned datasets. The results confirm that the proposed preprocessing method enhance the prediction performance of the linear calibration models PCR and PLSR and achieves better results than the method of using bandpass filtering with Savitzky-Golay using fixed parameters.

**5.2 Savitzky Golay filtering**

In the process of FT-IR spectroscopy analysis, the sample volume, sample preparation, the measuring method, and the measuring parameters, such as the choice of the scanning times and the scanning resolution will more or less bring in inevitable noise to the spectral data [5]. In order to make full use of the informative data and to reduce noise, the data pre-treatment is regularly necessary for the spectra before establishing the calibration model. Savitzky-Golay (SG) [77], smoothing is a widely used pre-treatment method that can effectively eliminate the noise like baseline-drift, tilt and reverse [123]. It consists of three different smoothing modes. The smoothing parameters are the degree of the polynomial (PD), the derivative order of polynomials (DOP) and the size of the smoothing window (SW). Here these parameters are very meaningful. A narrow SW is prone to cause calculation error, causing a decreased model precision, while a wide SW could over smooth and polish the spectral data, leading to the decreased accuracy. A reasonable choice of SW is crucial for SG smoothing. The SW could be appropriately selected according to the regression model prediction result by combining with the choice of the regression model components. The SG smoothing originally uses a symmetric SW which requires an equal number of data points on both sides of the center point and as a result, the technique discards points on both ends of the spectra during the pre-processing. Fourier Transform Infrared (FTIR) spectroscopy is one of the promising non-invasive approaches for the monitoring of blood glucose levels [114-116].

The most basic method for derivation is finite differences: the first derivative is estimated as the difference between two subsequent spectral measurement points; the second order derivative is then estimated by calculating the difference between two successive points of the first-order derivative spectra:

- (5.1)

= - = (5.2)

Where represents the first derivative and the second derivative at point (wavelength) . This method is extremely simple, but, unfortunately, cause noise inflation; it should almost always be avoided in practice.

The SG derivation is a basic method developed to avoid the noise inflation in finite differences. The SG derivation includes two steps.

1. Smoothing of the spectra, where averaging over a given number of points is performed:

(5.3)

Where is the number of points in the smoothing window centered around the current measurement point .

1. For first-order derivation, take the difference between two smoothed values with a given gap size between them (larger than zero); for second-order derivation, take twice the smoothed value at point and the smoothed value at a gap distance on either side.

- (5.4)

+ (5.5)

As can be seen from Equations (5.4) and (5.5), the actual derivation mimics a finite difference (Equations (5.1) and (5.2)). By applying a smoothing prior to the calculation and by introducing a gap size, the problem of decreasing the signal-to-noise ratio is reduced.

However, the prediction of glucose concentration from FTIR spectra of aqueous mixtures remains a challenge. The two most commonly used linear calibration models in the quantitative analysis of glucose from FTIR spectra are the Principal Component Regression (PCR) and the Partial Least Squares Regression (PLSR) [117-122]. However, the collected spectra are affected by baseline variations, scattering and instrumental noise. Hence, the spectra should be pre-processed before applying to the calibration model [70]. Bandpass filtering is a well-known pre-processing technique in chemometrics for removal of the baseline variations and the high-frequency noise components in the spectra, however, the design of an optimum bandpass filter that matches the raw spectra and the regression model remains a challenge [124]. Savitzky-Golay (SG) filter is one of the most commonly used pre-processing techniques which can reduce the degradation in the signal to noise ratio resulting from the light scattering [125]. However, the high-frequency noise components present in the spectra cannot be eliminated using this technique. Furthermore, matching the filter parameters to the raw data and the regression techniques used to preserve the important features remains a challenge [70]. A bandpass filtering technique based on combining the bandpass method with SG smoothing filter stage with fixed smoothing parameters has been introduced in [109] using dataset 3. In this chapter, the technique proposed in [109] is further improved by the optimization of the SG smoothing parameters to yield a better prediction performance than the model introduced in [109]. The proposed technique is also expanded in this chapter to cover the dataset 4 in the MIR region and dataset1 for NIR explained earlier in chapter 3. The method is evaluated and compared with the SG with fixed parameters method. The developed models are tested and evaluated to predict the concentration of glucose from a mixture of human serum albumin and glucose in a phosphate buffer solution. The results obtained show that the SG parameters optimization improves the prediction performance of both PCR and PLSR when using the model introduced in [109] and also achieves better prediction performance than the SG with fixed parameters pre-processing technique.

**5.3 Discussion and experimental results**

Starting the regression process by passing the raw data to the PCR and the PLSR models without pre-processing and the no of (PC) or (LV) is shown in Figure 5.1 for dataset 1 and figure 5.2 for dataset 3 and in figure 5.3 for dataset 4. The x-axis shows the number of PCR components in red and PLSR latent values in blue while the y-axis represents the estimated mean square error of prediction. Figure 5.1 shows 7 LV are required for PLSR and 10 PC for PCR in the case of dataset 1, while Figure 5.2 shows that 7 LV are required for PLSR and 9 PC for PCR in the case of dataset 3. Figure 5.3 shows that 6 LV are required for PLSR and 8 PC for PCR in the case of dataset 4 [128] these figures were generated using the equations in chapter 3 for PCR and PLSR which represent the RMSEP for the raw data before applying any preprocessing or filtering steps to explain why we choose this number of components as the optimum number as it contain the maximum variance representing the raw data matrix. In this work, the libPLS (an integrated library for Partial Least Squares Regression and Discriminant Analysis) toolbox [129], was used to perform the PLSR. The PLSR model is then pre-processed using Savitzky-Golay and Bandpass filtering each one aside. The bandpass filter prediction is controlled by four factors, center frequency C, bandwidth W, the Ripple R and the filter order N. The Lower and the upper cut-off frequencies can be calculated from the following equations.

(5.6)

(5.7)

Tables 5.1, 5.2 and 5.3 below explain the glucose prediction enhancement for Datasets 1, 3 and 4 explained in chapter 3 in this thesis. By using the PCR and PLSR with different size of SG window with and without modifying the polynomial degree and the derivative order and recording the impact of each technique on the Root Mean Square Error of Prediction (RMSEP) and the determination coefficient R2, the experimental results show the ability of the proposed models to enhance PLSR and PCR Glucose prediction for the two datasets and the NIR and MIR spectral regions. The digital bandpass filter removes the low-frequency noise and the high-frequency components and with the help of smoothing techniques, it can perform better as shown in figures 5.4 to 5.9. Figures 5.10 to 5.15 describes the effect of changing the size of the smoothing window on the RMSEP with fixed and variable derivative and polynomial degree these figures shows the different values of the RMSEP for different smoothing windows, different center frequency and different bandwidth so the RMSEP value with a smoothing window of 7 points has a different center frequency and bandwidth from the RMSEP value with a smoothing window of 9 points and so on, hence come the fluctuation in the results for some figures as they are not a function of the smoothing window size only but a function of the center frequency and the bandwidth as well which is being optimized every time we run the simulation to find these optimum parameters. The coupling of the proposed models with Bandpass filtering has reduced the RMSEP from 63 to 14.5 mg/dL for PCR and from 52.4 to 15 mg/dL for PLSR using Dataset 1. The proposed models have also reduced the RMSEP from 47 to 23 mg/dL for PCR and from 47.7 to 24.5 mg/dL for PLSR using Dataset 3 and for dataset 4 the RMSEP has been reduced from 38.7 to 15.9 mg/dL for PLSR and from 35.7 to 20 mg/dL for PCR.

**5.4 Conclusion**

In this chapter, a method based on combining bandpass filtering with the optimization of the Savitzky-Golay parameters, smoothing window size, polynomial order and the derivative degree has been proposed to improve the prediction performance of the standard linear models PLSR and PCR. The proposed method has been compared to the same pre-processing technique with fixed parameters and the developed regression models has been validated to predict the concentration of glucose from NIR spectra of an aqueous mixture of human serum albumin and glucose in a phosphate buffer solution. The results confirm that the proposed pre-processing method not only improves the prediction capability of both the linear calibration models PCR and PLSR models but also achieves better results than the bandpass with SG filtering using fixed parameters pre-processing method. This model reveals the importance of using multiple treatments to the spectral data at the same time, the tradeoff between the size of smoothing window, derivative order and the degree of the polynomial provides butter results in glucose prediction levels than the scatter correction approach but the model in some cases requires higher number of regression factors to reach stability and decay from desirable results after fewer factors which can be due to the bandpass filtering process as well and thus we C:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\oldata\PLSr vs PCR CV plot.tiffdecided to adopt a different approach to bandpass filtering.

Figure 5.1 explain the results of the regression model as a function of the RMSEP for PLSR vs PCR for dataset 1

C:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\MIR\PLSR vs PCR CV plot.tiff

C:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\NIR\PLSR vs PCR CV plot.tiffFigure 5.2 explain the results of the regression model as a function of the RMSEP for PLSR vs PCR for dataset 3.

Figure 5.3 explain the results of the regression model as a function of the RMSEP for PLSR vs PCR for dataset 4 without any filtering.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression model | Pre-processing |  | RMSEP (mg/dL) | Filtering Factors |
| PLSR | None | 0.89 | 52.4 |  |
| PLSR | SG variable parameters | 0.94 | 24.9 | Window=47 Order=3 Derive= 1 |
| PLSR | SG fixed parameters | 0.95 | 37 | Window=19 Order=2 Derive= 0 |
| PLSR | Chebyshev filter | 0.96 | 18.1 | C=0.0293 W=0.009 R=0.5 |
| PLSR | SG optimum parameters with Chebyshev | 0.96 | 15 | C=0.0281 W=0.0038 R=2.3  Window=35 Order=3 Derive= 2 |
| PLSR | Savitzky-golay with Chebyshev | 0.97 | 19 | C=0.0353 W=0.0061 R=2  Window=15 Order=0 Derive= 2 |
| PCR | None | 0.88 | 63.6 |  |
| PCR | SG fixed | 0.94 | 29.4 | Window=15 Order=2 Derive= 0 |
| PCR | Chebyshev filter | 0.96 | 35.3 | C=0.0427 W=0.0031 R=2.4 |
| PCR | SG variable | 0.97 | 25.5 | Window=80 Order=0 Derive= 2 |
| PCR | SG optimum parameters with Chebyshev | 0.97 | 14.5 | C=0.0402 W=0.0040 R=2.3 Window=89 Order=2 Derive= 2 |
| PCR | SG with Chebyshev | 0.97 | 15.7 | C=0.0072 W=0.0029 R=1.7 Window=83 Order=1 Derive= 2 |
| PCR | SG with Chebyshev | 0.97 | 16 | C=0.0130 W=0.0025 R=1.9 Window=105 Order=1 Derive= 0 |

Table 5.1 Results for dataset 1 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression model | Pre-processing |  | RMSEP (mg/dL) | Filtering Factors |
| PLSR | None | 0.91 | 47.7 | none |
| PLSR | Savitzky-golay | 0.97 | 37.5 | Window size=21 |
| PLSR | Chebyshev filter | 0.98 | 31.9 | C=0.0093f W=0.0051f R= 2.3dB  N=1 |
| PLSR | Savitzky-golay with Chebyshev | 0.99 | 28.9 | C=0.0095f W=0.0014f R=3dB  N= 1 window size =35 |
| PLSR | Savitzky-golay with Chebyshev | 0.99 | 28.8 | C=0.0128f W=0.003f R=0.6dB  N= 1 window size =17 |
| PLSR | Savitzky-golay with Chebyshev | 0.99 | 28.5 | C=0.0192f W=0.0096f R=1.3dB  N= 1 window size =15 |
| PLSR | Savitzky-golay with Chebyshev | 0.99 | 24.5 | C=0.0169f W=0.0206f R=2.7 dB  N=2 window size =25 |
| PCR | None | 0.97 | 47.7 | none |
| PCR | Savitzky-Golay | 0.97 | 40.8 | Window size=13 |
| PCR | Chebyshev filter | 0.99 | 30.2 | C=0.005f W=0.0035f R=0.5dB N=2 |
| PCR | Savitzky-golay with Chebyshev | 0.99 | 27.8 | C=0.007f W=0.005f R=1.4dB  N=2 window size=15 |
| PCR | Savitzky-golay with Chebyshev | 0.99 | 25.0 | C=0.0023f W=0.0011f R=2.8dB  N=1 window size=39 |
| PCR | Savitzky-golay with Chebyshev | 0.99 | 24.0 | C=0.01f W=0.0014f R=0.8dB   N=1 window size=21 |
| PCR | Savitzky-golay with Chebyshev | 0.99 | 23 | C=0.0093f W=0.0076f R=1dB  N=1 window size=5 |

Table 5.2 Results for dataset 2 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression model | Pre-processing |  | RMSEP (mg/dL) | Filtering Factors |
| PLSR | None | 0.91 | 35.7 | none |
| PLSR | Savitzky-golay | 0.97 | 27.6 | Window size=35 |
| PLSR | Chebyshev filter | 0.98 | 31.9 | C=0.0093f W=0.0027f R=2.3dB N=1 |
| PLSR | Savitzky-Golay with Chebyshev | 0.99 | 30.4 | C=0.0026f W=0.0017f R=1.8dB N=1 window size=33 |
| PLSR | Savitzky-Golay with Chebyshev | 0.99 | 26.6 | C=0.0015f W=0.0019f R=3dB N=2 window size=9 |
| PLSR | Savitzky-Golay with Chebyshev | 0.99 | 21.8 | C=0.0039f W=0.001f R=1.7dB N=1 window size=7 |
| PLSR | Savitzky-Golay with Chebyshev | 0.99 | 20.9 | C=0.0014f W=0.0017f R=3dB N=2 window size=19 |
| PCR | None | 0.97 | 38.7 | none |
| PCR | Savitzky-Golay | 0.97 | 28.3 | Window size=35 |
| PCR | Chebyshev filter | 0.99 | 30.2 | C=0.005f W=0.0035f N=2 R=0.5dB |
| PCR | Savitzky-Golay with Chebyshev | 0.99 | 40.6 | C=0.0029f W=0.0019f R=2.9dB N=1 window size=15 |
| PCR | Savitzky-Golay with Chebyshev | 0.99 | 35.2 | C=0.0029f W=0.0019f R=2.9dB N=1 window size=30 |
| PCR | Savitzky-Golay with Chebyshev | 0.99 | 34.2 | C=0.0029f W=0.0019f R=2.9dB N=1 window size=5 |
| PCR | Savitzky-Golay with Chebyshev | 0.99 | 15.9 | C=0.0029f W=0.0019f R=2.9dB N=1 window size=25 |

Table 5.3 Results for dataset 4 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

C:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\NIR\PCR\PCR NIR chebpass with SG optimum parameters.tiffC:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\oldata\PLSR\PLSR chebypass  with SG optimum parameters new.tiffFigure 5.4 explain the results of the regression model as a function of the RMSEP for PCR dataset 1 with different window sizes

Figure 5.5 explain the results of the regression model as a function of the RMSEP for PLSR dataset 1 with different window sizes.

C:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\MIR\PLSR\figures\PLSR MIR Chebpass with different SG windows new final.tiffC:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\MIR\window only\PCR\Figures\PCR MIR Chebpass with SG different windows new.tiffC:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\NIR\PLSR\figures\PLSR NIR Chebpass with optimum parameters new.tiffFigure 5.6 explain the results of the regression model as a function of the RMSEP for PCR dataset 3 with different window sizes.

Figure 5.7 explain the results of the regression model as a function of the RMSEP for PLSR dataset 3 with different window sizes.

C:\Users\user\Desktop\PLSR CBP with SG variable parameters new.tiffC:\Users\user\Desktop\thesi results\final figures and codes for the thesis\chapter 3 SG window\new data\MIR\window only\PCR\Figures\PCR MIR Chebpass with SG different windows new.tiffFigure 5.8 explain the results of the regression model as a function of the RMSEP for PCR dataset 4 with different window sizes.

Figure 5.9 explain the results of the regression model as a function of the RMSEP for PLSR dataset 4 with different window sizes.

C:\Users\user\Desktop\thesi results\Chapter 4 results for SG paper\oldata\PLSR\SG OLDATA all parameters optimization only.tiffFigure 5.10 explain the performance of the regression model with an without optimum parameters for PCR dataset 1 and for different window sizes.

C:\Users\user\Desktop\thesi results\Chapter 4 results for SG paper\oldata\PCR\SG OLDATA all parameters optimizationvs window size.tiff

Figure 5.11 explain the performance of the regression model with and without optimum parameters for PLSR dataset 1 and with different window sizes.

C:\Users\user\Desktop\thesi results\optimum parameters\SG\NIR\PCR\NIR PCR CBP window size vs optimum parameters plot.tiffC:\Users\user\Desktop\thesi results\Chapter 4 results for SG paper\NIR\PLSR\SGCBP NIR mix RMSEP vs window both fixed and optimized.tiffFigure 5.12 explain the performance of the regression model with and without optimum parameters for PCR dataset 3 for different window sizes.

Figure 5.13 explain the performance of the regression model with and without optimum parameters for PLSR dataset 3 and with different window sizes.

C:\Users\user\Desktop\thesi results\optimum parameters\SG\MIR\PLSR\SGCBP PLSR MIR RMSEP optimum parameters vs window size.tiffC:\Users\user\Desktop\thesi results\Chapter 4 results for SG paper\MIR\PCR\PCR SG window size vs optimum parameters new.tiffFigure 5.14 explain the performance of the regression model with and without optimum parameters for PCR dataset 4 and with different window sizes.

Figure 5.15 explain the performance of the regression model with and without optimum parameters for PLSR dataset 4 and with different window sizes.

**Chapter six**

**Bandstop (band reject) filtering in the quantitative analysis of glucose**

**6.1 Introduction**

This chapter proposes the use of bandstop filtering as a pre-treatment method in the quantitative analysis of glucose from both dataset 2 in the NIR region and dataset 4 in the MIR region explained in chapter 3 in this thesis. The proposed method is investigated and evaluated against the traditional bandpass filtering and implemented with the linear calibration models Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) to predict the glucose from an aqueous mixture consisting of glucose and Human Serum Albumin (HSA) dissolved in a phosphate buffer solution. The results obtained show that BSF pre-treatment achieves better prediction performance than bandpass filtering in both the NIR and MIR spectral regions by eliminating a specific spectral narrow band that does not contain any information related to analyte of interest which is in this case the glucose, this can be interpreted as a pure noise reducing the signal to noise ration unlike the bandpass which removes higher and lower frequencies which might contain related information about the glucose and hence reducing the signal to noise ratio such as shown in chapters 4 and 5 previously. For detailed analysis, the BPF and BSF were implemented under both the Butterworth and Chebyshev filter configurations in both bands; in the NIR region the Butterworth bandstop filtering combined with the PLSR model provides the best glucose prediction by reducing the Root Mean Square Error of Prediction (RMSEP) from 100 mg/dL without filtering to 34 mg/dL with a coefficient of determination R2 of 0.982. In the MIR region the Chebyshev bandstop filtering combined with either PLSR or PCR improves the glucose prediction by reducing the RMSEP by 54% compared to 45% when using BPF and with R2 of 0.996.

**6.2 Background**

NIR and MIR spectroscopy has been proposed as a promising technologies for non-invasive glucose measurement [130-134]. Multivariate calibration algorithms such as the Partial Least Squares Regression (PLSR) and the Principal Component Regression (PCR) have been introduced to correlate the variation of the IR spectra with the information of the analyte of interest within multiple wavelengths [117-122]. Despite their fundamental similarities, Near Infrared (NIR) and Mid Infrared (MIR) have traditionally used different instrumentation and hence are commonly introduced as distinct measurement techniques for the same applications[135]. There are several parameters affecting collected NIR and MIR spectra such as instrumental noise, baseline variations and scattering. This is why, when processing raw spectra, a pre-processing step is often performed before the calibration model [70]. Mid-Infrared (MIR) spectroscopy is usually used for the qualitative analysis of organic substances. One of the MIR measurement advantages is its sensitivity to both organic constituents and mineral components. Glucose has two absorption bands, the combination and overtone, in the Near-infrared region. Previously implemented methods using MIR radiation introduced a strong glucose absorption in this region which is a crucial factor for the improvement of the measurements sensitivity compared to the NIR region measurements as it has been confirmed in this paper [45]. Also, molecule fundamental vibrations are detected in the MIR region and also there is a promising correlation between MIR and NIR spectra [136].

Bandpass filtering has been introduced as a pre-processing technique in chemometrics for the removal of the variations in the baseline which occur in low frequencies, and the high-frequency noise in the spectra components, however the design of an optimal bandpass filtering with a corresponding regression model remains a challenge [23] [137]. Furthermore, during the digital bandpass filtering operation useful information which might exist in the very low and high frequencies, will be removed which will affect the prediction accuracy. It is stipulated in this work that bandstop filtering could overcome these limitations as information that exists in the high and low-frequency regions of the spectra could be preserved; this is because bandstop filtering offers better selectivity and hence the ability to eliminate only a specific band of frequencies by design, based on the range of the filter’s rejection band. Bandstop filters can be classified into wide band and narrow band and have been extensively used in a wide range of applications including communications, biomedical, audio systems and Raman spectroscopy [138-139]. However, to the best of our knowledge, no work has been attempted to investigate the adoption of the bandstop filtering for the pre-processing in the context of NIR/MIR spectroscopy. In this chapter, the use of BSF is investigated as a pre-treatment method in NIR/MIR spectroscopy and evaluated against bandpass filtering. A novel model is proposed for the quantitative analysis of glucose from both NIR and MIR spectra based on the bandstop filtering pre-processing coupled with PLSR and PCR regression models. The efficacy of the proposed model is validated to determine the concentration of glucose from an aqueous mixture consisting of glucose and Human Serum Albumin (HSA) in a phosphate buffer solution. It is shown that bandstop filtering pre-treatment achieves better prediction performance than the traditional BPF in both NIR and MIR spectral regions. Two filters Butterworth and Chebyshev have been implemented in both regions. The results indicate that using dataset 2 the Butterworth bandstop filtering combined with the PLSR model provides the best prediction by reducing the Root Mean Square Error of Prediction (RMSEP) from 100 mg/dL to 34 mg/dL with a coefficient of determination (R2) of 0.982; the proposed model also reduces the required number of Latent values from 7 to 5, thus reducing the complexity of the regression model as well. In dataset 4 the Chebyshev bandstop filtering achieved the best performance by reducing the RMSEP by 54% and the required number of principal components from 12 to 8 with a coefficient of determination equal to 0.996.

**6.3 Theory**

The quantitative analysis of NIR/MIR spectra requires sophisticated signal processing and data analysis. The signal processing techniques deal mostly with ‘de-noising’ the raw data and are often referred to as pre-processing. The data analysis deals mostly with the interpretation and extraction of the desired information from the signals; these data analysis techniques are often referred to as multivariate calibration.

In order to provide meaningful information, the signals can be represented as absorbance spectra, interferogram, or frequency domain spectrum.

**6.4 Multivariate Calibration**

This corresponds to regression algorithms that formulate the extent of the relationship between a response (in this case the spectra) and the predictors (the Glucose concentration) via a calibration model. The model can then be adopted to estimate new responses. PCR and PLSR are two widely used techniques in Chemometrics [102-103].

Generally, the performance of the calibration model can be evaluated by metrics such as the and .

**6.5 Preprocessing**

The prediction performance of the regression algorithms and hence the calibration models depend heavily on the quality of the raw spectra. Practically collected spectra in non-controlled environments suffer from baseline variations, scattering, high frequency noise, noise due to high background spectra such as water, and noise due to temperature variation. The aim of pre-processing is to enhance both the quality and resolution of the raw spectra, by using signal processing techniques for smoothing, filtering, scatter correction and spectral enhancement.

**6.6 Digital Filtering**

Digital filtering corresponds to multiple mathematical steps in the frequency domain that manipulate the spectral characteristics of the data in a controlled way to provide desirable effects. Frequency spectrum encoded signals such as NIR/MIR spectra, where the required information is preserved in their Fourier transform frequency components, are examples of data that can be enhanced by digital filtering. However, the design of an optimal filter in practice is challenging; the Grid search optimization algorithm was proposed by Arnold [108], has been adopted in many reported publications to define the optimum filter parameters.

**6.6-A Digital Bandpass Filtering**

Digital BPF has been introduced as one of the most popular pre-processing techniques in the quantitative analysis of glucose from NIR spectra [23] [108-110]. Bandpass filtering is used to eliminate the variation of the baseline and noise present in the high frequencies of the raw spectra. Center frequency and bandwidth are the main components that define the digital bandpass filter and these two factors must be optimized to find the desired band which represents the maximum information about the analyte of interest which in our case is the glucose concentration. The raw spectra are first pre-processed using the digital bandpass filtering and then the PCR and the PLSR models are coupled with bandpass filter using a specific number of Principal Components (PC) and Latent Variables (LV) [23] [109-110]

**6.6-B Digital Bandstop Filtering**

Bandstop filters are very useful in removing unwanted signals. These filters are designed with methods very similar to the BPF mentioned above. The only difference is that the frequency of the poles is the same as the frequency of the zeros [111-112]. Bandstop filtering has not been used before to predict the concentration of glucose from NIR or MIR spectra. Unlike the case of bandpass filtering, the bandstop filtering targets a specific undesired band and removes it with minimum effects on the concentration of glucose. The main parameters that require optimization are cut-off bandwidth and the centre frequency that define the desirable frequency band that has optimum information about the glucose concentration. The PCR and the PLSR models are then built using the filtered spectra with a specific number of PCs and LVs [23] [109-110].

**6.6-C Butterworth and Chebyshev filters**

The Butterworth filter has a monotonic characteristic of its magnitude function of the n-poles in both the stopband and passband of the filter. Alternatively, the Chebyshev filter has a ripple in the passband. The reason for flexible ripples in the passband or stopband is to gain a sharper transition between the passband and stopband compared to the Butterworth filter [111]. The filters characteristics are controlled by the filter order and ripple. The filter ripple can be in both the passband or in the stopband, this work will concentrate on both bands to identify the difference between these bands in order to improve the prediction capability of the regression model [23] [109-110]

**6.7 Data Preparation**

**6.7-A Chemical samples preparation**

To obtain spectral data, a chemical mixture has been designed in the laboratories of the chemistry department in Sheffield University, the samples for this experiment has been prepared in an uncontrolled environment. The mixture used in this study consists of glucose, Human Serum Albumin (HSA) and phosphate buffer solution. This has been done by preparing two mixtures primarily with glucose, HSA and phosphate buffer and secondary with HSA and Phosphate buffer only. The concentration of the chemical components in the primary mixture has the glucose concentration at 500 mg/ dL, HSA at 5 g/dL and the concentration of the phosphate buffer at 0.01 m/ dL; the secondary mixture has the same concentrations for HSA and the phosphate except for the glucose as it does not contain any glucose. Both mixtures have the same pH level which is 7.4. The first sample with a volume of 5 mg is collected from the primary mixture which contains a glucose concentration of 500 mg/ dL then the same volume from the secondary mixture is added to the primary mixture flask to prepare sample two. Now the glucose concentration is 495 mg/ dL in the primary mixture with the other chemical components at the same level. Repeating the same process until all the 100 samples are collected with 5 mg/ dL glucose concentration decrement in each sample from the previous one, but the HSA and phosphate levels remain constant and at the end there will be 100 samples with glucose concentration varies from 500 mg/ dL to 5 mg/dl but HSA and phosphate level stabilized at 5 g/ dL and 0.01 m/ dL respectively.

**6.7-B Spectra collection**

NIR and MIR spectra have been collected in the materials department at Sheffield University. Using a Frontier MIR/NIR spectrometer with Two temperature-stabilized DTGS (deuterated triglyceride sulfate) detectors: one optimized for MIR with an enhanced SNR of 15,000:1 peak-to-peak for a 5 second scan, and one suited for NIR with <10 µA RMS noise over 250 nm range for a 1 minute scan. Each sample of the 100 samples prepared previously has been processed using the spectrometer with a 1.7 nm separation and a wavelength range varies from 2100 nm to 8000 nm. 4 measurements were collected and averaged for each sample to enhance spectra precision [113]. The collected spectra have been divided into two ranges the NIR which varies from 2100 nm to 2500 nm and MIR which varies from 4000 cm-1 to 1250 cm-1 (2500 nm to 8000 nm). Figure 6.1 shows the plot of the collected spectra. The prediction process requires the collected raw spectra to be divided into training and testing data subsets using Kennard and stone algorithm [127], so that the training set consisting of 75 spectra samples and the testing set has 25 spectra samples. The idea behind using Kennard and Stone algorithm is to distribute the training set samples in the most diverse way based on a maxima criterion. This is done by identifying the two most distant samples in the data set and remove them to the training set then repeating the same to the remaining samples until we get the entire desired number of the training samples [140].

**6.8 Proposed Model**

Applying the regression model should be the first step to perform prediction. The PCR has been implemented without any filtering to the raw data to recognize the variance captured by each PC for PCR. After implementing the regression models using dataset 3 in the NIR region, 98% of the captured variance has been represented by 12 PCs for PCR for dataset 2 and 4 in the NIR and MIR regions as shown in Figure 6.2. The optimum number of LVs which will provide the best variance representation for the PLSR for datasets 2 and 4 in the NIR and MIR region is shown in Figure 6.3. The PLSR model requires 7 LVs to represent 95% of the variance for both spectral regions. The libPLS toolbox (an integrated library for discriminant analysis and partial least square regression) [129] [141], has been implemented to develop the PLSR model in this study (<http://www.libpls.net>) based on the Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm [103]. The second step will be applying the bandpass and bandstop filtering methods to the raw data. The Chebyshev filter consists of three parameters that control its performance: the center frequency (C), the Bandwidth (B) and the Ripple (R). The Butterworth filter does not require any ripple. The output signal from the pre-treatment is then processed and applied to the regression model to calculate the RMSEP which will indicate how close the predicted glucose concentrations to the actually prepared samples as shown in the methodology diagram in Figure 6.4.

C:\Users\user\Desktop\All data plot with bold font.tiff

Figure 6.1 explain the raw data plot for Dataset 2 and 4 for each sample as a function of the absorbed light of the spectra.

C:\Users\user\Desktop\Chemometrics paper\figures\figure 2 PLSR Variance NIR vs MIR\PLSR Variance NIR vs MIR.tiffC:\Users\user\Desktop\Chemometrics paper\figures\figure 3 PCR Variance NIR vs MIR\Variance  MIR vs NIR PCR alone.tiffFigure 6.2 explains the represented variance for NIR and MIR when applying PCR only without any filtering.

Figure 6.3 explains the represented variance for NIR and MIR when applying PLSR only without any filtering.

**Raw Data**

Bandpass Filtering

Bandstop Filtering

Butterworth

Chebyshev

Regression Model PLSR and PCR

Figure 6.4 explains the system flow chart steps that the raw dataset passes through to obtain the required results.

**6.9 Experimental Results and Discussion**

In order to evaluate the ability of the proposed model to predict the glucose concentration from both dataset2 and 4, the regression models were implemented with and without the pre-processing methods and the results have been compared to the bandpass method under the same conditions. This has been done by combining the regression models with both bandpass and bandstop pre-processing methods separately and compared the results using the standard measure, the RMSEP. The optimum parameters for the filters are calculated using the grid search method by sweeping the entire spectrum range till the lowest RMSEP is obtained. The RMSEP for both the linear regression models without any pre-processing is found to be 100 mg/dL in the NIR region and the RMSEP was reduced to 42 mg/dL for PCR and 34 mg/dL for PLSR when coupled with bandstop filtering methods respectively. Wide and narrow response filters have been used in this experiment (Chebyshev and Butterworth). The results illustrate that there is a recognized reduction in the RMSEP value using both pre-treatment methods, however the bandstop filtering method outperforms the BPS method by about 30% in the case of NIR Chebyshev with PLSR and it reaches 32% in the case of Butterworth with PLSR. This indicates that the ability of bandstop filter to reject the unwanted noise is better than bandpass in the NIR region. Furthermore, Bandstop filtering still provides better prediction than the bandpass with both Chebyshev and Butterworth; there is a 24.7% improvement in prediction performance of the bandstop filter than the bandpass method for the PLSR model with a Chebyshev filter in the MIR region and 16% improvement was found for the PCR model in the MIR region when combined with the Butterworth filter. The NIR and MIR regions slight performance variation could be attributed to that MIR measures the fundamental vibrations of a molecule [136], whilst NIR spectroscopy measures the combinations and overtones of bonds and fundamental vibrations. Bandstop filtering still provides better prediction than the bandpass with both Chebyshev and Butterworth.

Figure 6.5 and 6.6 explain the performance of both filters in both NIR and MIR regions using the RMSEP value and demonstrate the stability of the prediction performance of the bandstop in comparison to the bandpass. For example, in figure 6.5 the bandpass requires more latent values and never reaches the RMSEP level of bandpass while the bandstop requires fewer components and achieves better RMSEP than bandpass in several latent values. In Figure 6.5 to 6.8 the bandpass performs as expected for several components then decays from the expected results unlike the bandstop filtering which remains with constant level of error prediction, this is an indication that the bandpass filtering removes some of the required information during its filtering process which is necessary for predicting the glucose from the chemical mixture. Bandpass filtering is known to treat high frequency noise variation in the base line but it does not treat the overlapping noise between the glucose and other mixture components; although advances in spectrometers produce an enhanced signal with a very low instrument noise level, the overlapping noise remains a challenge since it is distributed all over the spectrum. BSF is more capable to tackle overlapping noise with a narrower rejection bandwidth as shown in Figures 6.9 to 6.16. In these figures the RMSEP results for the entire spectrum are plotted, it is clear that every band rejected using bandstop produces a desired RMSEP unlike bandpass only 1 or 2 bands provide a desired RMSEP and all the rest have RMSEP values higher than the RMSEP without filtering; this indicates that the variance explained in the data filtered by bandstop is higher than bandpass and since the regression models PCR and PLSR are variance dependent, bandstop can be considered a better solution to enhance the prediction ability of these linear regression models than bandstop.

**6.10 Conclusion**

In this chapter, the use of bandstop filtering has been proposed as pre-treatment method in the quantitative analysis of glucose prediction and investigated using a mixture consists of glucose and human serum albumin in a phosphate buffer solution. The proposed model has been applied to two spectral ranges NIR (dataset 2) and MIR (dataset 4) spectral ranges. The bandstop filtering method has been compared to the well-known pre-processing method based on bandpass filtering. The experimental results indicate that the proposed model provides better glucose prediction than the bandpass filtering in both NIR and MIR ranges and for wide and narrow response filters such as Chebyshev and Butterworth. This could be due to missing glucose information removed by bandpass filtering during the process of eliminating the high-frequency components and baseline variations. The proposed model enhances the prediction ability of the linear regression models the PLSR and the PCR significantly compared to the bandpass filtering and also require less number of regression factors (LVs) and (PCs), based on these results we decided to repeat the tests with scatter correction and SG filtering using bandstop filtering instead of bandpass to further investigate and compare the performance.

C:\Users\user\Desktop\Chemometrics paper\figures\figure 5 PLSR NIR\PLSR NIR.tiff

C:\Users\user\Desktop\Chemometrics paper\figures\figure 6 PCR NIR\PCR NIR.tiffFigure 6.5 explains the glucose prediction enhancement as a function of the RMSEP for PLSR dataset 2 with different filters

Figure 6.6 explains the glucose prediction enhancement as a function of the RMSEP for PCR dataset 2 with different filters.

C:\Users\user\Desktop\Chemometrics paper\figures\figure 5 PLSR NIR\MIRPLSR.tiff

C:\Users\user\Desktop\Chemometrics paper\figures\figure 6 PCR NIR\MIRPCR.tiffFigure 6.7 explains the glucose prediction enhancement as a function of the RMSEP for PLSR dataset 4 with different filters.

Figure 6.8 explains the glucose prediction enhancement as a function of the RMSEP for PCR dataset 4 with different filters.

C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\BBP vs BBS PLSR MIR optimum parameters.tiff

C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\CBS vs BP PLSR MIR optimum Parameters.tiffFigure 6.9 explains the glucose prediction enhancement as a function of the RMSEP for Butterworth bandpass vs bandstop PLSR dataset 2

Figure 6.10 explains the glucose prediction enhancement as a function of the RMSEP for Chebyshev Bandpass vs Bandstop PLSR dataset 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Pre-processing | RMSEP (mg/dL) | R2 | Optimum Parameters |
| PLSR | None | 100 | 0.879 | None |
| PLSR | Butterworth bandpass | 66.317 | 0.947 | C=0.0553f W=0.0254f |
| PLSR | Chebyshev Bandpass | 67.37 | 0.945 | C=0.3147f W=0.0052f R=2.5dB |
| PLSR | Butterworth bandstop | 34.422 | 0.986 | C=0.8491f W=0.0245f |
| PLSR | Chebyshev bandstop | 37.349 | 0.983 | C=0.7225f W=0.0844f R=0.5dB |
| PCR | None | 102.9 | 0.872 | None |
| PCR | Butterworth bandpass | 75.070 | 0.932 | C=0.8525f W=0.0054f |
| PCR | Chebyshev Bandpass | 71.520 | 0.938 | C=0.8525f W=0.0050f R=2.7dB |
| PCR | Butterworth bandstop | 44.867 | 0.975 | C=0.3089f W=0.0188f |
| PCR | Chebyshev bandstop | 42.728 | 0.978 | C=0.3090f W=0.0263f R=1.8 dB |

Table 6.1 Results for dataset 2 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

Table 6.2 Results for dataset 4 explaining the output of different preprocessing methods as a function of the Root mean square error of prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Preprocessing | RMSEP (mg/dL) | R2 | Optimum Parameters |
| PLSR | None | 35.961 | 0.984 | None |
| PLSR | Butterworth bandpass | 26.358 | 0.995 | C=0.0041 W=0.0020 |
| PLSR | Chebyshev Bandpass | 25.228 | 0.992 | C=0.0019 W=0.0012 R=1.2 |
| PLSR | Butterworth bandstop | 18.644 | 0.995 | C=0.1514 W=0.0036 |
| PLSR | Chebyshev bandstop | 16.334 | 0.996 | C=0.0337 W=0.0024  R=2.9dB |
| PCR | None | 38.754 | 0.982 | None |
| PCR | Butterworth bandpass | 26.683 | 0.991 | C=0.0013 W=0.0015 |
| PCR | Chebyshev Bandpass | 21.176 | 0.994 | C=0.00041f W=0.0018f  R=2.1db |
| PCR | Butterworth bandstop | 20.495 | 0.995 | C=0.1499 W=0.0016 |
| PCR | Chebyshev bandstop | 17.559 | 0.996 | C=0.4350 W=0.0014 |

C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\CBP vs CBS PCR NIR optimum parameters.tiffFigure 6.11 explains the glucose prediction enhancement as a function of the RMSEP for Butterworth bandpass vs bandstop PLSR dataset 2.

C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\CBS vs BP PLSR MIR optimum Parameters.tiff

Figure 6.12 explains the glucose prediction enhancement as a function of the RMSEP for Chebyshev bandpass vs bandstop PCR dataset 2.

C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\BBP vs BBS PLSR NIR optimum parameters.tiffC:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\BBP vs BBS PCR NIR optimum parameters.tiffFigure 6.13 explains the glucose prediction enhancement as a function of the RMSEP for Butterworth bandpass vs bandstop PCR dataset 2.

Figure 6.14 explains the glucose prediction enhancement as a function of the RMSEP for Chebyshev bandpass vs bandstop PLSR dataset 4.

**C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\CBP vs CBS PLSR NIR optimum parameters.tiff**C:\Users\user\Desktop\Miner correction folder\Chapter 6 figures\BBP vs BBS PCR MIR optimum parameters.tiff Figure 6.15 explains the glucose prediction enhancement as a function of the RMSEP for Chebyshev bandpass vs bandstop PCR dataset 4.

Figure 6.16 explains the glucose prediction enhancement as a function of the RMSEP using Butterworth bandpass vs bandstop PCR dataset 4.

Chapter Seven

**Bandstop filtering with preprocessing techniques**

**7.1 Introduction**

This Chapter sums the work from all other chapters and proposes a two novel pre-processing methods, the first method is based on combining bandstop filtering from chapter 6 with SG smoothing filter from chapter 5 and the second method based on combining bandstop filtering with scatter correction technique SNV from chapter 4 to enhance the prediction capability of the linear regression models Partial Least Squares Regression (PLSR) in near infrared (NIR) and Mid infrared spectroscopy. The proposed model has been compared to the effective pre-processing technique bandpass filtering combined with the same preprocessing methods SG and SNV proposed in chapters 4 and 5 respectively, the developed regression models has been validated to predict the glucose concentration from datasets 2 and 4 explained in chapter 3 featuring NIR and MIR spectra of an aqueous mixture of human serum albumin and glucose in a solution of distilled water and phosphate buffer. The results confirms that the proposed pre-processing methods not only improves the prediction capability of the linear calibration PLSR model but also achieves better results than the methods in chapter 4, 5 and 6 by reducing the Root Mean Square Error Prediction RMSEP and increasing the determination coefficient R2.

**7.2 Proposed model**

In this chapter, two novel linear pre-processing techniques were proposed that combines linear ‘feature preserving’ smoothing techniques such as SNV and SG smoothing filter with digital Bandstop filtering to improve the prediction performance of the linear regression model PLSR. The proposed methods of Preprocessing are evaluated against the bandpass filtering method combined with the same Preprocessing techniques SG and SNV. Corresponding models were developed, tested and evaluated using experimental data to predict the glucose concentration from a mixture of glucose and human serum albumin in a solution of distilled water and phosphate buffer. The results obtained show not only that the proposed model enhances the prediction ability of both PLSR, but also achieves better prediction performance than bandpass method.

**7.3 Pre-Processing**

The prediction performance of the regression algorithms and hence the calibration model depend heavily on the quality of the raw spectra. Practically collected spectra in non-controlled environments suffer from baseline variations, scattering, high frequency noise, noise due to high background spectra such as water, and noise due to temperature variation. The aim of pre-processing is to enhance both the quality and resolution of the raw spectra, by using signal processing techniques for smoothing, filtering, scatter correction and spectral enhancement.

**7.4 Digital Filtering**

Digital filtering corresponds to multiple mathematical steps in the frequency domain that manipulate the spectral characteristics of the data in a controlled way to provide desirable effects. Frequency spectrum encoded signals such as NIR/MIR spectra, where the required information is preserved in their Fourier transform frequency components, are examples of data that can be enhanced by digital filtering. However, the design of an optimal filter in practice is challenging; the Grid search optimization algorithm was proposed by Arnold [108] has been adopted in many reported publications to define the optimum filter parameters.

**7.4-A Digital Bandpass Filtering**

Digital bandpass filtering has been introduced in chapters 4 and 5 as one of the most popular pre-processing techniques in the quantitative analysis of glucose from NIR spectra [23] [108-110]. Bandpass filtering is used in joint combination with scatter correction technique SNV in chapter 4 and it also has been used combined with SG filtering in chapter 5 to eliminate the variation of the baseline and noise present in the high frequencies of the raw spectra. Center frequency and bandwidth are the main components that define the digital bandpass filter and these two factors must be optimized to find the desired band which represents the maximum information about the analyte of interest which in our case is the glucose concentration. The raw spectra are first pre-processed using the SNV and SG each a side then applying the digital bandpass filtering and finally the PLSR model is coupled with bandpass filtering using a specific number of Latent Variables (LV) to determine the RMSEP as a measuring factor for the glucose prediction accuracy [23] [109-110].

**7.4-B Digital Bandstop Filtering**

Bandstop filtering are very useful in removing unwanted signals. These filters are designed with methods very similar to the bandpass filtering mentioned above. The only difference is that the frequency of the poles is the same as the frequency of the zeros [111-112]. Bandstop filtering has been proposed to predict the concentration of glucose from NIR or MIR spectra alternatively dataset 2 and dataset 4 in chapter 6. Unlike the case of Bandpass filtering, the bandstop filtering targets a specific undesired band and removes it with minimum effects on the concentration of glucose. The main parameters that require optimization are cut-off bandwidth and the centre frequency that define the desirable frequency band that has optimum information about the glucose concentration. The PLSR model is then built using the filtered spectra with a specific number of LVs [23] [109-110].

**7.4-C Butterworth filtering**

The Butterworth filter has a monotonic characteristic of its magnitude function of the n-poles in both the stopband and passband of the filter [111]. The filter characteristics are controlled by the filter order, bandwidth and center frequency this work will concentrate on both bands to identify the difference between these bands in order to improve the prediction capability of the regression model [23] [109-110].

**7.5 Light scattering and baseline effects**

Nonlinearities are undesired scatter effects in NIR spectroscopy caused by the particles size in chemical samples or the wavelength size of electromagnetic radiation, therefore; applying suitable pre-processing techniques can eliminate these effects on a large scale [125]. Differences in the effective path length and light scattering are the main reasons for undesired systematic variation which represents a significant part of the sample total variation and can be identified as a baseline shifting (multiplicative effects) in addition to other non-linear effects [125]. During the reflectance measurement of the NIR sample, two reflected radiation can be obtained. These radiations are specular and diffusive. Sampling geometry and instrument design can minimize specular reflection as it does not contain any chemical information [64] while the main source of information in the NIR spectra is the diffusively reflected light because it reflects in a wide range of directions but the collected information will contain microstructure (scattering) in addition to the absorption of the sample chemical composition [125]. The main form of light scattering are Lorenz-Mie and Rayleigh and both are generated due to scattering of electromagnetic radiation such as density fluctuation, droplets, bubbles, small particles, cells, micro-organ cells and fibers [125]. Rayleigh occurs when the wavelength of the electromagnetic radiation is much larger than particles diameter (λ ˃ 10d), hence, it is wavelength dependent (~1⁄λ^4) [64]. In the case of NIR spectroscopy, the wavelength is smaller than the particle size and that's when Lorentz-Mie scattering is predominant and unlike Rayleigh scattering it depends on the scattering particles shape instead of the wavelength. The diffused and reflected spectrum of the samples is usually affected by noise resulting from scattering. If the glucose concentration levels in the sample is very low, the effects of spectral scattering could take over the spectral information. To reduce the effects pf scattering interference, scatter correction methods must be applied to the spectral data as a pre-treatment process. Scatter correction techniques are pre-treatment methods that can separate the analyte absorbed informative and the scattered signal contained in the spectral data. They can reduce the variation in the spectral in the same set of samples due to particle size inhomogeneous. Scatter correction pre-processing methods can be represented by many techniques such as Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV). This chapter concentrates on SNV as pre-processing techniques in combination with bandstop filtering to correct the scattering effects in order to enhance the regression model prediction of glucose.

**7.6 Experimental data preparation**

In this chapter, a Fourier transform NIR/MIR spectrometer has been used to collect 2 datasets dataset 2 and dataset 4 explained in chapter 3, each dataset consist of 100 sample for 100 mixtures consist of Human Serum Albumin and glucose in a phosphate buffer solution. The preparation of the buffer solution was achieved by adding Phosphate powder into 1 liter of distilled water to prepare the buffer solution with a concentration of 0.01 M/dl and the pH value was adjusted to 7.4. For each sample, four spectra were measured and then averaged without removing the sample from the spectrometer to improve the precision of the measurement [113]. The individual component concentrations were selected in a way to cover their physiological ranges in blood. The ranges of the glucose and Human serum Albumin concentrations of the prepared 100 samples are from 5 to 500 mg/dl and 5 g/dl respectively. The collected NIR spectra spanned the spectral region from 2100 nm to 2500 nm and the mixed dataset varies from 2100 nm to 3120 nm with a spectral resolution of 1.7 nm and the MIR spectral range varied from 125 cm-1 to 4000 cm-1. The collected raw spectra were divided into training and testing data subsets using Kennard stone algorithm [127], such that the training set consisting of 75 spectra samples and the test set has 25 spectra samples.

**7.7 Discussion and experimental results**

Starting the regression process by passing the raw data to the PLSR model without pre-processing and the no of (LV) is shown in chapter 6 Figure (6.1) dataset 2 and figure (6.2) for dataset 4. The x-axis shows the number PLSR latent values in Blue while the y-axis represents the estimated mean square error of prediction. Figure (6.1) shows 7 LV are required for PLSR in the case of dataset 2, while Figure (6.2) shows that 8 LV are required for PLSR in the case of dataset 4 [128]. In this chapter, the libPLS (an integrated library for Partial Least Squares Regression and Discriminant Analysis) toolbox [129] was used to perform the PLSR. The PLSR model is then pre-processed using SNV and Savitzky-Golay with Bandstop filtering each one aside. The bandstop filter prediction is controlled by four factors, center frequency C, bandwidth W, the Ripple R and the filter order N. The Lower and the upper cut-off frequencies can be calculated from the equations 6-1 and 6-2 from chapter 6. Tables 7.1 to 7.4 below explain the glucose prediction enhancement for dataset 2 in the NIR region and dataset 4 in the MIR region for the proposed model with Butterworth and Chebyshev filters. The digital bandstop filter removes the overlapping frequency components and with the help of SNV and SG filtering techniques, it can perform better as shown in figures 7.1 to 7.4. The coupling of the proposed models with Bandstop filtering has reduced the RMSEP from 94 to 26 for PLSR using dataset 2 in the NIR region. Applying Butterworth bandstop filtering to dataset 2 will reduce the RMSEP from 100 to 26 mg/dl using SG filtering with 57 point smoothing window, 1st derivative and 1st polynomial degree, whereas the use of SNV reduces the RMSEP to 31.4 mg/dL. Alternatively applying the Chebyshev bandstop filtering to dataset 2 reduces the RMSEP from 100 to 25 mg/dL using SG with 19 points smoothing window zero derivative and 2nd polynomial degree and to 29.6 mg/dL using SNV. By applying Butterworth bandstop filtering to dataset 4 reduce the error from 38 to 10 mg/dL using SNV and to 13 using SG whereas applying the Chebyshev reduce the error to 11 using SNV and to 12 using SG.

**7.8 Conclusion**

In this chapter, all the previously propose models in chapters 4,5 and 6 has been summand to produce two novel models based on the joint combination between bandstop filtering, SG filtering and SNV each one a side to further enhance the glucose prediction accuracy of the quantitative regression model PLSR in the NIR and MIR spectral region using dataset 2 and dataset 4, both bandstop with SG and bandstop with SNV performed better than the bandpass with SG and bandpass with SNV and that confirms the importance of using bandstop filtering instead of bandpass to predict glucose from NIR and MIR spectra which can be attributed to missing information of glucose during the bandpass filtering process. The experimental results proved that there is a 73% to 75% reduction in the RMSEP using the bandstop filtering combined with the preprocessing methods the SG and the SNV compared to 50% to 55% reduction in the same factor using the bandpass filtering with the SG, SNV and MSC preprocessing method. This indicate that the proposed models using bandstop filtering is 25% more accurate glucose prediction than bandpass filtering with the preprocessing methods in the NIR and MIR spectral regions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Pre-processing | RMSEP (mg/dL) | R2 | Filtering Factors |
| PLSR | None | 38.17 | 0.88 | None |
| PLSR | Butterworth bandpass | 29.6 | 0.936 | C=0.0020 W=0.0024 |
| PLSR | Butterworth bandstop | 19.6806 | 0.954 | C=0.0839 W=0.0036 |
| PLSR | Butterworth bandpass with SNV | 26.969 | 0.982 | C=0.0902 W=0.0639 |
| PLSR | Butterworth bandstop with SNV | 10.83 | 0.983 | C=0.6211 W=0.0900 |
| PLSR | Butterworth bandpass with SG | 18.4589 | 0.987 | C=0.0632 W=0.0044 Window=51 Order=1  Derive= 0 |
| PLSR | Butterworth bandstop with SG | 13.9032 | 0.984 | C=0.4448 W=0.014  Window=9 Order=1  Derive= 2 |

Table 7.1 Results for PLSR Butterworth dataset 4 which explains the results of the glucose prediction for different preprocessing methods as a function of the root mean square error pf prediction.

Table 7.2 Results for PLSR Butterworth dataset 2 which explains the results of the glucose prediction for different preprocessing methods as a function of the root mean square error pf prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Pre-processing | RMSEP (mg/dL) | R2 | Filtering Factors |
| PLSR | None | 100 | 0.879 | None |
| PLSR | Butterworth bandpass | 78.317 | 0.947 | C=0.0553f W=0.0254f |
| PLSR | Butterworth bandstop | 37.37 | 0.945 | C=0.3147f W=0.0052f |
| PLSR | Butterworth bandpass with SNV | 60.43 | 0.986 | C=0.8491f W=0.0245f |
| PLSR | Butterworth bandstop with SNV | 31.39 | 0.983 | C=0.7225f W=0.0844f |
| PLSR | Butterworth bandpass with SG | 54.5 | 0.986 | C=0.8491f W=0.0245f  Window=22 Order=2 Derive= 0 |
| PLSR | Butterworth bandstop with SG | 26.34 | 0.983 | C=0.7225f W=0.0844f  Window=57 Order=1 Derive= 1 |

Table 7.3 Results for PLSR Chebyshev dataset 2 which explains the results of the glucose prediction for different preprocessing methods as a function of the root mean square error pf prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Pre-processing | RMSEP (mg/dL) | R2 | Optimum Parameters |
| PLSR | None | 38.17 | 0.88 | None |
| PLSR | Chebyshev bandpass | 30.945 | 0.936 | C=0.0112 W=0.0042  R=1.2 |
| PLSR | Chebyshev bandstop | 20.193 | 0.954 | C=0.0098 W=0.0012  R=2 |
| PLSR | Chebyshev bandpass with SNV | 24.761 | 0.982 | C=0.0029 W=0.0067  R=2.7 |
| PLSR | Chebyshev bandstop with SNV | 11.219 | 0.983 | C=0.0153 W=0.0070  R=3 |
| PLSR | Chebyshev bandpass with SG | 21.361 | 0.987 | C=0.0096 W=0.0027 R=1.3 Window=33 Order=2 Derive= 1 |
| PLSR | Chebyshev bandstop with SG | 12.0651 | 0.984 | C=0.0134 W=0.0049 R=1.8 Window=29 Order=1 Derive= 0 |

Table 7.4 Results for PLSR Chebyshev dataset 4 which explains the results of the glucose prediction for different preprocessing methods as a function of the root mean square error pf prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Pre-processing | RMSEP (mg/dL) | R2 | Filtering Factors |
| PLSR | None | 100 | 0.88 | None |
| PLSR | Chebyshev bandpass | 73.317 | 0.936 | C=0.0033 W=0.0063  R=2.1 |
| PLSR | Chebyshev bandstop | 40.143 | 0.954 | C=0.0173 W=0.0024  R=1.4 |
| PLSR | Chebyshev bandpass with SNV | 68.43 | 0.982 | C=0.0092 W=0.0396  R=2.8 |
| PLSR | Chebyshev bandstop with SNV | 29.6 | 0.983 | C=0.0216 W=0.0900  R=2.4 |
| PLSR | Chebyshev bandpass with SG | 56.34 | 0.987 | C=0.1471 W=0.0053  R=1.9 Window=34 Order=1 Derive= 2 |
| PLSR | Chebyshev bandstop with SG | 25.12 | 0.984 | C=0.0592 W=0.0035  R=2.6 Window=16 Order=2 Derive= 0 |

C:\Users\user\Desktop\thesi results\Chapter 6 Bstop with preprossing\NIRF\PLSR\Butterworth\PLSR Butter bandstop with preprocessing.tiffC:\Users\user\Desktop\thesi results\Chapter 6 Bstop with preprossing\MIRF\PLSR\figures\PLSR MIR with different preprocessing new.tiffFigure 7.1 explains the glucose prediction enhancement as a function of the RMSEP for Dataset 2 PLSR using Butterworth with different preprocessing methods.

Figure 7.2 explains the glucose prediction enhancement as a function of the RMSEP for Dataset 4 PLSR using Butterworth with different preprocessing methods.

**C:\Users\user\Desktop\thesi results\Chapter 6 Bstop with preprossing\MIRF\PLSR\Chebychev\figures\PLSR NIRF with different processing new.tiff**C:\Users\user\Desktop\thesi results\Chapter 6 Bstop with preprossing\NIRF\PLSR\Chebychev\figures\PLSR NIRF with cheby bandstop and different preprocessing techniques final.tiffFigure 7.3 explains the glucose prediction enhancement as a function of the RMSEP for Dataset 2 PLSR using Chebyshev with different preprocessing methods.

Figure 7.4 explains the glucose prediction enhancement as a function of the RMSEP for Dataset 4 PLSR using Chebyshev with different preprocessing methods.

**Chapter Eight**

**Conclusion and future work**

**8.1 Conclusions**

In this chapter a brief conclusion about the proposed models and their advantages and drawbacks. Also a discussion about the future work which can enhance some of the mentioned drawbacks.

In chapter 4, a novel pre-processing technique based on the combination of scatter correction with bandpass filtering has been proposed to enhance the prediction ability of linear regression models. The proposed method has been compared to the pre-processing technique RReliefF and the developed regression models has been validated to predict the concentration of glucose from NIR spectra of an aqueous mixture of human serum albumin and glucose in a phosphate buffer solution. The results confirm that the proposed pre-processing method not only improves the prediction capability of both the linear calibration models PCR and PLSR models but also achieves better results than the non-linear RReliefF pre-processing method. The reduction in the RMSEP was not as high as expected when applying the proposed model to dataset 2 in the NIR region which can be related to the deleted useful information during the bandpass filtering treatment to the baseline effects and the noise from high frequency components. It could be also due to that our data set in the NIR region has one glucose region which is the combination absorption band which varies from 2050 nm to 2300 nm and missing the first and the overtone absorption regions which falls in the wavelength range from 700 nm to 1500 nm hence we decided to avoid using dataset 2 with SG method and focus on dataset 3 and 4 to further investigate.

In chapter 5, a method based on combining bandpass filtering with the optimization of the Savitzky-Golay parameters, smoothing window size, polynomial order and the derivative degree has been proposed to improve the prediction performance of the standard linear models PLSR and PCR. The proposed method has been compared to the same pre-processing technique with fixed parameters and the developed regression models has been validated to predict the concentration of glucose from NIR spectra of an aqueous mixture of human serum albumin and glucose in a phosphate buffer solution. The results confirm that the proposed pre-processing method not only improves the prediction capability of both the linear calibration models PCR and PLSR models but also achieves better results than the bandpass with SG filtering using fixed parameters pre-processing method. This model reveals the importance of using multiple treatments to the spectral data at the same time, the tradeoff between the size of smoothing window, derivative order and the degree of the polynomial provides butter results in glucose prediction levels than the scatter correction approach but the model in some cases requires higher number of regression factors to reach stability and decay from desirable results after fewer factors which can be due to the bandpass filtering process as well and thus we decided to adopt a different approach to bandpass filtering to .

In chapter 6, the use of bandstop filtering has been proposed as pre-treatment method in the quantitative analysis of glucose prediction and investigated using a mixture consists of glucose and human serum albumin in a phosphate buffer solution. The proposed model has been applied to two spectral ranges NIR (dataset 2) and MIR (dataset 4) spectral ranges. The bandstop filtering method has been compared to the well-known pre-processing method based on bandpass filtering. The experimental results indicate that the proposed model provides better glucose prediction than the bandpass filtering in both NIR and MIR ranges and for wide and narrow response filters such as Chebyshev and Butterworth. This could be due to missing glucose information removed by bandpass filtering during the process of eliminating the high-frequency components and baseline variations. The proposed model enhances the prediction ability of the linear regression models the PLSR and the PCR significantly compared to the bandpass filtering and also require less number of regression factors (LVs) and (PCs), based on these results we decided to repeat the tests with scatter correction and SG filtering using bandstop filtering instead of bandpass to further investigate and compare the performance.

In chapter 7, all the propose models has been summand to produce two novel models based on the joint combination between bandstop filtering, SG filtering and SNV each one a side to further enhance the glucose prediction accuracy of the quantitative regression model PLSR in the NIR and MIR spectral region using dataset 2 and dataset 4, both bandstop with SG and bandstop with SNV performed better than the bandpass with SG and bandpass with SNV and that confirms the importance of using bandstop filtering instead of bandpass to predict glucose from NIR and MIR spectra which can be attributed to missing information of glucose during the bandpass filtering process. The experimental results proved that there is a 73% to 75% reduction in the RMSEP using the bandstop filtering combined with the preprocessing methods the SG and the SNV compared to 50% to 55% reduction in the same factor using the bandpass filtering with the SG, SNV and MSC preprocessing method. This indicate that the proposed models using bandstop filtering is 25% more accurate glucose prediction than bandpass filtering with the preprocessing methods in the NIR and MIR spectral regions.

**8.2 Future work**

Future work for this thesis can be divided into three sections the data preparation, the regression models and the preprocessing methods.

Data preparation section will consider acquiring new data with more realistic characteristics such as blood plasma or real blood samples from diabetic volunteers and using spectrometers with NIR and MIR spectral ranges from 700 nm to 2500 nm for NIR and from 4000 cm-1 to 400 cm-1 for MIR then combine the spectral ranges together to get all the available information regard the glucose. It is also recommended to use higher spectral resolution results for example 1 nm spacing between measurements and more glucose samples with accurate glucose levels.

Regression model section will concentrate on using different regression models such as Recursive weighted Partial Least Squares (RPLS) and Support vector regression (SVR) models and compare them to the currently used models to determine their efficacy and evaluate their performance. It is also preferred to investigate nonlinear regression methods and discuss their performance and efficacy compared to the currently used linear methods.

The preprocessing section will focus on using other preprocessing techniques such as Extended Multiplicative Scatter correction EMSC, Inverse Multiplicative Scatter Correction IMSC to treat the scattering effects and compare these methods to the currently used ones, also it is recommended to investigate Norris William filtering method and compare it to the SG filtering method currently used to evaluate its performance and efficacy. It is recommended to use multiple filtering steps to treat the different drawbacks in the collected spectra as we discovered in the SG optimum parameters case where we keep modifying the filtering parameters till we get the best combination that suits the collected dataset

An additional step is recommended before the use of preprocessing methods is to use classification methods to classify the glucose levels based on their importance with the help of some compression techniques such SMOREG to shrink the size of the data matrices before any regression model used to speed up the prediction process and save time compared to the current time consuming methodology.

**8.3 Implementation cost**

**8.3 A NIR**

Two portable spectrometers can be used for the implementation of this project both span the NIR spectral range from 700 nm to 2500 nm and described as below.

**8.3A.1 DLP2010NIR**

The DLP2010NIR digital micro mirror device (DMD) acts as a spatial light modulator (SLM) to steer near-infrared (NIR) light and create patterns with speed, precision, and efficiency. Featuring high resolution in a compact form factor, the DLP2010NIR DMD is often combined with a grating single element detector to replace expensive In GaAs linear array-based detector designs, leading to high performance, cost-effective portable NIR Spectroscopy solutions. The DLP2010NIR DMD enables wavelength control and programmable spectrum and is well suited for low power mobile applications such as 3D biometrics, facial recognition, skin analysis, material identification and chemical sensing. This device cost about £117.03 from Mouser Electronics.

**8.3A.2 DLP4500NIR**

The DLP4500NIR digital micro mirror device (DMD) acts as a spatial light modulator (SLM) to steer near-infrared (NIR) light and create patterns with speed, precision, and efficiency. Featuring high resolution in a compact form factor, the DLP4500NIR DMD is often combined with a single element detector to replace expensive In GaAs array-based detector designs, leading to high performance, cost-effective portable solutions. This spectrometer costs £457 from Mouser Electronics.

**8.3B MIR**

# **8.3B.1 4100 Exo Scan Series**

The 4100 Exo Scan™ is a portable, hand-held, battery operated FTIR analyzer designed to measure a wide variety of solid and liquid samples. The instrument contains interchangeable ATR and external reflectance sample interfaces, a PDA based control panel that wirelessly links to a laptop for upload and download of data, methods etc, multi-level software and user interfaces, and a docking station allows the Exoscan to be used in a lab similar to a benchtop FTIR.

**References**

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

[2] A. Circulation, A. Low, I. Low, S. High, B. High, and a Langlois, ‘Chapter 2 : Background’, pp. 1–36, 1989.

[3] M. A. Arnold and G. W. Small, ‘Determination of physiological levels of glucose in an aqueous matrix with digitally filtered fourier transform near-infrared spectra’, *Anal. Chem.*, vol. 62, no. 14, pp. 1457–1464, 1990.

[4] L. a Marquardt, M. a Arnold, and G. W. Small, ‘Near-infrared spectroscopic measurement of glucose in a protein matrix.’, *Anal. Chem.*, vol. 65, no. 4, pp. 3271–3278, 1993.

[5] G. W. Small, M. a. Arnold, and L. a Marquardt, ‘Strategies for coupling digital filtering with partial least-squares regression: application to the determination of glucose in plasma by Fourier transform near-infrared spectroscopy.’, *Anal. Chem.*, vol. 65, no. 22, pp. 3279–89, 1993.

[6] K. H. Hazen, M. A. Arnold, and G. W. Small, ‘Temperature-insensitive near-infrared spectroscopic measurement of glucose in aqueous solutions’, *Appl. Spectrosc.*, vol. 48, no. 4, pp. 477–483, 1994.

[7] S. Pan, H. Chung, M. a Arnold, and G. W. Small, ‘Near-infrared spectroscopic measurement of physiological glucose levels in variable matrices of protein and triglycerides.’, *Anal. Chem.*, vol. 68, no. 7, pp. 1124–1135, 1996.

[8] H. Chung, M. A. Arnold, M. Rhiel, and D. W. Murhammer, ‘Simultaneous measurement of glucose and glutamine in aqueous solutions by near infrared spectroscopy’, *Appl. Biochem. Biotechnol.*, vol. 50, no. 2, pp. 109–125, 1995.

[9] H. Chung, M. A. Arnold, M. Rhiel, and D. W. Murhammer, ‘Simultaneous Measurements of Glucose, Glutamine, Ammonia, Lactate, and Glutamate in Aqueous Solutions by Near-Infrared Spectroscopy’, *Appl. Spectrosc.*, vol. 50, no. 2, pp. 270–276, Feb. 1996.

[10] a S. Bangalore, R. E. Shaffer, G. W. Small, and M. a Arnold, ‘Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy.’, *Anal. Chem.*, vol. 68, no. 23, pp. 4200–4212, 1996.

[11] R. E. Shaffer, G. W. Small, and M. a Arnold, ‘Genetic algorithm-based protocol for coupling digital filtering and partial least-squares regression: application to the near-infrared analysis of glucose in biological matrices.’, *Anal. Chem.*, vol. 68, no. 15, pp. 2663–2675, 1996.

[12] Q. Ding, G. W. Small, and M. a Arnold, ‘Genetic algorithm-based wavelength selection for the near-infrared determination of glucose in biological matrixes: initialization strategies and effects of spectral resolution.’, *Anal. Chem.*, vol. 70, no. 21, pp. 4472–4479, 1998.

[13] C. Fischbacher, K.-U. Jagemann, K. Danzer, U. A. Müller, L. Papenkordt, and J. Schüler, ‘Enhancing calibration models for non-invasive near-infrared spectroscopical blood glucose determination’, *Fresenius J. Anal. Chem.*, vol. 359, pp. 78–82, 1997.

[14] K. H. Hazen, M. A. Arnold, and G. W. Small, ‘Measurement of glucose and other analytes in undiluted human serum with near-infrared transmission spectroscopy’, *Anal. Chim. Acta*, vol. 371, no. 2–3, pp. 255–267, 1998.

[15] H. Arimoto, M. Tarumi, and Y. Yamada, ‘Temperature-insensitive measurement of glucose concentration based on near infrared spectroscopy and partial least squares analysis’, *Opt. Rev.*, vol. 10, no. 2, pp. 74–76, 2003.

[16] S. Wold, H. Antti, F. Lindgren, and J. Öhman, ‘Orthogonal signal correction of near-infrared spectra’, *Chemom. Intell. Lab. Syst.*, vol. 44, no. 1–2, pp. 175–185, 1998.

[17] A. K. Amerov, J. Chen, G. W. Small, and M. A. Arnold, ‘Scattering and absorption effects in the determination of glucose in whole blood by near-infrared spectroscopy’, *Anal. Chem.*, vol. 77, no. 14, pp. 4587–4594, 2005.

[18] S. Kasemsumran, Y. P. Du, K. Murayama, M. Huehne, and Y. Ozaki, ‘Near-infrared spectroscopic determination of human serum albumin, γ-globulin, and glucose in a control serum solution with searching combination moving window partial least squares’, *Anal. Chim. Acta*, vol. 512, no. 2, pp. 223–230, 2004.

[19] S. Kasemsumran, Y. P. Du, K. Maruo, and Y. Ozaki, ‘Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares’, *Chemom. Intell. Lab. Syst.*, vol. 82, no. 1-2 SPEC. ISS, pp. 97–103, 2006.

[20] M. Ren and M. A. Arnold, ‘Comparison of multivariate calibration models for glucose, urea, and lactate from near-infrared and Raman spectra’, *Anal. Bioanal. Chem.*, vol. 387, no. 3, pp. 879–888, 2007.

[21] J. Chen, M. a. Arnold, and G. W. Small, ‘Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions’, *Anal. Chem.*, vol. 76, no. 18, pp. 5405–5413, 2004.

[22] W. Saeys, K. Beullens, J. Lammertyn, H. Ramon, and T. Naes, ‘Increasing robustness against changes in the interferent structure by incorporating prior information in the augmented classical least-squares framework’, *Anal. Chem.*, vol. 80, no. 13, pp. 4951–4959, 2008.

[23] A. 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’, *2010 IEEE Work. Signal Process. Syst.*, pp. 243–248, 2010.

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

[25] 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.

[26] 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.

[27] B. Malik, M. Benaissa, and S. Member, ‘Determination of Glucose Concentration from Near-Infrared Spectra Using Locally Weighted Partial Least Square Regression Test’, pp. 6169–6171, 2012.

[28] B. A. Malik, ‘Infrared Spectra using Least Square Support Vector Machine’, no. Icic, pp. 475–478, 2015.

[29] K. C. Patchava, M. Benaissa, B. Malik, and H. Behairy, ‘Local linear embedded regression in the quantitative analysis of glucose in near infrared spectra’, *Anal. Methods*, vol. 7, no. 4, pp. 1484–1492, 2015.

[30] K. C. Patchava, M. Benaissa, and H. Behairy, ‘Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra’, *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2015-Novem, no. 1, pp. 2379–2382, 2015.

[31] S. Yu, D. Li, H. Chong, C. Sun, H. Yu, and K. Xu, ‘In vitro glucose measurement using tunable mid-infrared laser spectroscopy combined with fiber-optic sensor.’, *Biomed. Opt. Express*, vol. 5, no. 1, pp. 275–86, 2013.

[32] S. Liakat, K. a Bors, T.-Y. Huang, A. P. M. Michel, E. Zanghi, and C. F. Gmachl, ‘In vitro measurements of physiological glucose concentrations in biological fluids using mid-infrared light.’, *Biomed. Opt. Express*, vol. 4, no. 7, pp. 1083–90, 2013.

[33] D. H. Kim, I. K. Ilev, and J. U. Kang, ‘Using mid-infrared glucose absorption peak changes for high-precision glucose detection’, *Conf. Proc. - Lasers Electro-Optics Soc. Annu. Meet.*, no. Di, pp. 226–227, 2007.

[34] P. Wilks, ‘NIR vs Mid-IR – How to choose’, *Spectroscopy*, vol. 4, pp. 32–34, 2007.

[35] D. C. Klonoff, J. Braig, B. Sterling, C. Kramer, D. Goldberger, and R. Trebino, ‘Mid-infrared spectroscopy for noninvasive blood glucose monitoring’, *IEEE Lasers Electro-optics Soc Newsletters*, vol. 12, pp. 13–14, 1998.

[36] H. A. MacKenzie *et al.*, ‘Advances in photoacoustic noninvasive glucose testing’, *Clin. Chem.*, vol. 45, no. 9, pp. 1587–1595, 1999.

[37] S. Low-Ying, R. A. Shaw, M. Leroux, and H. H. Mantsch, ‘Quantitation of glucose and urea in whole blood by mid-infrared spectroscopy of dry films’, *Vib. Spectrosc.*, vol. 28, no. 1, pp. 111–116, 2002.

[38] Y.-J. Kim, S. Hahn, and G. Yoon, ‘Determination of glucose in whole blood samples by mid-infrared spectroscopy’, *Appl. Opt.*, vol. 42, no. 4, pp. 745–749, 2003.

[39] O. S. Khalil, ‘Non-invasive glucose measurement technologies: an update from 1999 to the dawn of the new millennium’, *Diabetes Technol. Ther.*, vol. 6, no. 5, pp. 660–697, 2004.

[40] 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.

[41] C. Vrančić *et al.*, ‘Continuous glucose monitoring by means of mid-infrared transmission laser spectroscopy in vitro’, *Analyst*, vol. 136, no. 6, pp. 1192–1198, 2011.

[42] M. A. Pleitez, T. Lieblein, A. Bauer, O. Hertzberg, H. von Lilienfeld-Toal, and W. Mäntele, ‘In vivo noninvasive monitoring of glucose concentration in human epidermis by mid-infrared pulsed photoacoustic spectroscopy’, *Anal. Chem.*, vol. 85, no. 2, pp. 1013–1020, 2012.

[43] L. Nybacka, ‘FTIR spectroscopy of glucose’. 2016.

[44] J. Kottmann, J. M. Rey, and M. W. Sigrist, ‘Mid-Infrared photoacoustic detection of glucose in human skin: towards non-invasive diagnostics’, *Sensors*, vol. 16, no. 10, p. 1663, 2016.

[45] R. Kasahara, S. Kino, S. Soyama, and Y. Matsuura, ‘Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers’, *Biomed. Opt. Express*, vol. 9, no. 1, p. 289, 2018.

[46] J. Y. Sim, C.-G. Ahn, E.-J. Jeong, and B. K. Kim, ‘In vivo Microscopic Photoacoustic Spectroscopy for Non-Invasive Glucose Monitoring Invulnerable to Skin Secretion Products’, *Sci. Rep.*, vol. 8, no. 1, p. 1059, 2018.

[47] P. Geladi, D. MacDougall, and H. Martens, ‘Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat’, *Appl. Spectrosc.*, vol. 39, no. 3, pp. 491–500, May 1985.

[48] H. Martens, S. A. Jensen, and P. Geladi, ‘Multivariate linearity transformation for near-infrared reflectance spectrometry’, in *Proceedings of the Nordic symposium on applied statistics*, 1983, pp. 205–234.

[49] B. G. Osborne, ‘Comparative study of methods of linearisation and scatter correction in near infrared reflectance spectroscopy’, *Analyst*, vol. 113, no. 2, pp. 263–267, 1988.

[50] T. Naes, T. Isaksson, and B. Kowalski, ‘Locally weighted regression and scatter correction for near-infrared reflectance data’, *Anal. Chem.*, vol. 62, no. 7, pp. 664–673, 1990.

[51] J. Sorvaniemi, A. Kinnunen, A. Tsados, and Y. Mälkki, ‘Using partial least squares regression and multiplicative scatter correction for ft-nir data evaluation of wheat flours’, *LWT - Food Science and Technology*, vol. 26, no. 3. pp. 251–258, 1993.

[52] M. S. Dhanoa, S. J. Lister, R. Sanderson, and R. J. Barnes, ‘The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra’, *J. Near Infrared Spectrosc.*, vol. 2, no. 1, pp. 43–47, 1994.

[53] I. S. Helland, T. Næs, and T. Isaksson, ‘Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data’, *Chemom. Intell. Lab. Syst.*, vol. 29, no. 2, pp. 233–241, 1995.

[54] S. Wold, H. Antti, F. Lindgren, and J. Öhman, ‘Orthogonal signal correction of near-infrared spectra’, *Chemom. Intell. Lab. Syst.*, vol. 44, no. 1–2, pp. 175–185, 1998.

[55] G. H. Geesink *et al.*, ‘Prediction of pork quality attributes from near infrared reflectance spectra’, *Meat Sci.*, vol. 65, no. 1, pp. 661–668, 2003.

[56] N. B. Gallagher, T. A. Blake, and P. L. Gassman, ‘Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil’, *J. Chemom.*, vol. 19, no. 5–7, pp. 271–281, 2005.

[57] Z.-P. Chen, J. Morris, and E. Martin, ‘Extracting Chemical Information from Spectral Data with Multiplicative Light Scattering Effects by Optical Path-Length Estimation and Correction’, *Anal. Chem.*, vol. 78, no. 22, pp. 7674–7681, 2006.

[58] S. Thennadil, M. Hy, and A. Kohler, ‘Physics-Based Multiplicative Scatter Correction Approaches for Improving the Performance of Calibration Models’, *Appl. Spectrosc.*, vol. 60, pp. 315–321, 2006.

[59] H. Martens, J. P. Nielsen, and S. B. Engelsen, ‘Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures’, *Anal. Chem.*, vol. 75, no. 3, pp. 394–404, 2003.

[60] M. R. Maleki, A. M. Mouazen, H. Ramon, and J. De Baerdemaeker, ‘Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy’, *Biosyst. Eng.*, vol. 96, no. 3, pp. 427–433, 2007.

[61] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel, ‘On the geometry of SNV and MSC’, *Chemom. Intell. Lab. Syst.*, vol. 96, no. 1, pp. 22–26, 2009.

[62] A. Kohler, M. Zimonja, V. Segtnan, and H. Martens, ‘Standard normal variate, multiplicative signal correction and extended multiplicative signal correction preprocessing in biospectroscopy’, 2009.

[63] F. Liu, Y. Jiang, and Y. He, ‘Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer’, *Anal. Chim. Acta*, vol. 635, no. 1, pp. 45–52, 2009.

[64] 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.

[65] F. Liu, Y. Jiang, and Y. He, ‘@article{moghimi2010vis, title={Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit}, author={Moghimi, Ali and Aghkhani, Mohammad H and Sazgarnia, Ameneh and Sarmad, Majid}, journal={Biosystems ’, *Anal. Chim. Acta*, vol. 635, no. 1, pp. 45–52, 2009.

[66] M. H. M. Killner, J. J. R. Rohwedder, and C. Pasquini, ‘A PLS regression model using NIR spectroscopy for on-line monitoring of the biodiesel production reaction’, *Fuel*, vol. 90, no. 11, pp. 3268–3273, 2011.

[67] B. Jamshidi, S. Minaei, E. Mohajerani, and H. Ghassemian, ‘Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of Valencia oranges’, *Comput. Electron. Agric.*, vol. 85, pp. 64–69, 2012.

[68] J. Engel *et al.*, ‘Breaking with trends in pre-processing?’, *TrAC Trends Anal. Chem.*, vol. 50, pp. 96–106, 2013.

[69] J. Engel, L. Blanchet, L. M. C. Buydens, and G. Downey, ‘Confirmation of brand identity of a Trappist beer by mid-infrared spectroscopy coupled with multivariate data analysis’, *Talanta*, vol. 99, pp. 426–432, 2012.

[70] H. Chen, Q. Song, G. Tang, Q. Feng, and L. Lin, ‘The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models’, *ISRN Spectrosc.*, vol. 2013, 2013.

[71] W. Windig, J. Shaver, and R. Bro, ‘Loopy MSC: a simple way to improve multiplicative scatter correction’, *Appl. Spectrosc.*, vol. 62, no. 10, pp. 1153–1159, 2008.

[72] S. Lohumi, S. Lee, H. Lee, and B.-K. Cho, ‘A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration’, *Trends Food Sci. Technol.*, vol. 46, no. 1, pp. 85–98, 2015.

[73] J. U. Porep, D. R. Kammerer, and R. Carle, ‘On-line application of near infrared (NIR) spectroscopy in food production’, *Trends Food Sci. Technol.*, vol. 46, no. 2, pp. 211–230, 2015.

[74] Y. Bi *et al.*, ‘A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation’, *Anal. Chim. Acta*, vol. 909, pp. 30–40, 2016.

[75] L. Gao *et al.*, ‘Improve the prediction accuracy of apple tree canopy nitrogen content through multiple scattering correction using spectroscopy’, *Agric. Sci.*, vol. 7, no. 10, pp. 651–659, 2016.

[76] C. Pasquini, ‘Near infrared spectroscopy: a mature analytical technique with new perspectives--a review’, *Anal. Chim. Acta*, 2018.

[77] 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.

[78] J. Steinier, Y. Termonia, and J. Deltour, ‘Smoothing and differentiation of data by simplified least square procedure’, *Anal. Chem.*, vol. 44, no. 11, pp. 1906–1909, 1972.

[79] H. H. Madden, ‘Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data’, *Anal. Chem.*, vol. 50, no. 9, pp. 1383–1386, 1978.

[80] K. Kitamura and K. Hozumi, ‘Effect of the degree of polynomials in the Savitzky—Golay method for calculation of second-derivative spectra’, *Anal. Chim. Acta*, vol. 172, pp. 111–118, 1985.

[81] P. A. Gorry, ‘General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method’, *Anal. Chem.*, vol. 62, no. 6, pp. 570–573, 1990.

[82] W. H. Press and S. A. Teukolsky, ‘Savitzky-Golay smoothing filters’, *Comput. Phys.*, vol. 4, no. 6, pp. 669–672, 1990.

[83] C. Ruffin and R. L. King, ‘The analysis of hyperspectral data using Savitzky-Golay filtering-theoretical basis. 1’, in *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS’99 (Cat. No. 99CH36293)*, 1999, vol. 2, pp. 756–758.

[84] J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh, ‘A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter’, *Remote Sens. Environ.*, vol. 91, no. 3, pp. 332–344, 2004.

[85] J. Luo, K. Ying, P. He, and J. Bai, ‘Properties of Savitzky-Golay digital differentiators’, *Digit. Signal Process.*, vol. 15, no. 2, pp. 122–136, 2005.

[86] S. Hargittai, ‘Savitzky-Golay least-squares polynomial filters in ECG signal processing’, in *Computers in Cardiology, 2005*, 2005, pp. 763–766.

[87] J. Luo, K. Ying, and J. Bai, ‘Savitzky-Golay smoothing and differentiation filter for even number data’, *Signal Processing*, vol. 85, no. 7, pp. 1429–1434, 2005.

[88] G. Vivó-Truyols and P. J. Schoenmakers, ‘Automatic Selection of Optimal Savitzky−Golay Smoothing’, *Anal. Chem.*, vol. 78, no. 13, pp. 4598–4608, 2006.

[89] B. Pan, H. Xie, Z. Guo, and T. Hua, ‘Full-field strain measurement using a two-dimensional Savitzky-Golay digital differentiator in digital image correlation’, *Opt. Eng.*, vol. 46, no. 3, p. 33601, 2007.

[90] M. Člupek, P. Matějka, and K. Volka, ‘Noise reduction in Raman spectra: Finite impulse response filtration versus Savitzky--Golay smoothing’, *J. Raman Spectrosc. An Int. J. Orig. Work all Asp. Raman Spectrosc. Incl. High. Order Process. also Brillouin Rayleigh Scatt.*, vol. 38, no. 9, pp. 1174–1179, 2007.

[91] S. R. Delwiche and J. B. Reeves, ‘A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: example with savitzky-golay filters and partial least squares regression’, *Appl. Spectrosc.*, vol. 64, no. 1, pp. 73–82, 2010.

[92] J. Xie, T. Pan, J. Chen, H. Chen, and X. Ren, ‘Joint optimization of Savitzky-Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose’, *Chinese J. Anal. Chem.*, vol. 38, no. 3, pp. 342–346, 2010.

[93] M. A. Awal, S. S. Mostafa, and M. Ahmad, ‘Performance analysis of Savitzky-Golay smoothing filter using ECG signal’, *Int. J. Comput. Inf. Technol.*, vol. 1, no. 02, 2011.

[94] D. Chen, Y. Chen, and D. Xue, ‘Digital Fractional Order Savitzky-Golay Differentiator’, *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 58, no. 11, pp. 758–762, Nov. 2011.

[95] Q. Quan and K. Y. Cai, ‘Time-domain analysis of the Savitzky-Golay filters’, *Digit. Signal Process. A Rev. J.*, vol. 22, no. 2, pp. 238–245, 2012.

[96] M. Chakraborty and S. Das, ‘Determination of signal to noise ratio of electrocardiograms filtered by band pass and Savitzky-Golay filters’, *Procedia Technol.*, vol. 4, pp. 830–833, 2012.

[97] S. R. Krishnan and C. S. Seelamantula, ‘On the Selection of Optimum Savitzky-Golay Filters’, *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 380–391, Jan. 2013.

[98] B. Zimmermann and A. Kohler, ‘Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy’, *Appl. Spectrosc.*, vol. 67, no. 8, pp. 892–902, 2013.

[99] Ç. Candan and H. Inan, ‘A unified framework for derivation and implementation of Savitzky--Golay filters’, *Signal Processing*, vol. 104, pp. 203–211, 2014.

[100] D. Acharya, A. Rani, S. Agarwal, and V. Singh, ‘Application of adaptive Savitzky--Golay filter for EEG signal processing’, *Perspect. Sci.*, vol. 8, pp. 677–679, 2016.

[101] G. Romo-Cárdenas *et al.*, ‘Nyquist-Shannon theorem application for Savitzky-Golay smoothing window size parameter determination in bio-optical signals’, *Results Phys.*, vol. 11, pp. 17–22, 2018.

[102] J. Shlens, ‘A Tutorial on Principal Component Analysis’, 2014.

[103] O. Yeniay and A. Goktas, ‘A comparison of partial least squares regression with other prediction methods’, *Hacettepe J. Math. Stat.*, vol. 31, no. 99, pp. 99–101, 2002.

[104] N. Faber, ‘A closer look at the bias–variance tradeoff in multivariate calibration’, *J. Chemom.*, vol. 192, no. February, pp. 185–192, 1999.

[105] L. Xu *et al.*, ‘On estimating model complexity and prediction errors in multivariate calibration: Generalized resampling by random sample weighting (RSW)’, *J. Chemom.*, vol. 25, no. 2, pp. 51–58, 2011.

[106] W. Wu and R. Manne, ‘Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications’, *Chemom. Intell. Lab. Syst.*, vol. 51, no. 2, pp. 145–161, 2000.

[107] K. Baumann, ‘Cross-validation as the objective function for variable-selection techniques’, *TrAC - Trends Anal. Chem.*, vol. 22, no. 6, pp. 395–406, 2003.

[108] M. A. Arnold and G. W. Small, ‘Determination of physiological levels of glucose in an aqueous matrix with digitally filtered Fourier transform near-infrared spectra’, *Anal. Chem.*, vol. 62, no. 14, pp. 1457–1464, 1990.

[109] 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’, in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, 2016, pp. 6210–6213.

[110] O. A. Alrezj, K. C. Patchava, M. Benaissa, and S. Alshebeili, ‘Coupling Scatter Correction with bandpass filtering for preprocessing in the quantitative analysis of glucose from near infrared spectra’, in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, 2017, pp. 1800–1803.

[111] R. C. Dorf, ‘The electrical engineering handbook’, *Piscataway, NJ IEEE Press. c1993, Ed. by Dorf, Richard C.*, 1993.

[112] G. L. Matthaei, ‘Interdigital band-pass filters’, *IRE Trans. Microw. theory Tech.*, vol. 10, no. 6, pp. 479–491, 1962.

[113] R. De Maesschalck *et al.*, ‘The development of calibration models for spectroscopic data using principal component regression’, *Internet J. Chem.*, vol. 2, no. 19, p. 1, 1999.

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

[115] M. A. Arnold, ‘Non-invasive glucose monitoring’, no. 7, pp. 46–49, 1996.

[116] S. F. Malin, T. L. Ruchti, T. B. Blank, S. N. Thennadil, and S. L. Monfre, ‘Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy’, *Clin. Chem.*, vol. 45, no. 9, pp. 1651–1658, 1999.

[117] U. A. Müller, B. Mertes, C. Fischbacher, K. U. Jageman, and K. Danzer, ‘Non-invasive blood glucose monitoring by means of near infrared spectroscopy: methods for improving the reliability of the calibration models.’, *Int. J. Artif. Organs*, vol. 20, no. 5, pp. 285–290, 1997.

[118] D. M. Haaland and E. V. Thomas, ‘Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information’, *Anal. Chem.*, vol. 60, no. 11, pp. 1193–1202, 1988.

[119] D. M. Haaland, M. R. Robinson, G. W. Koepp, E. V Thomas, and R. P. Eaton, ‘Reagentless Near-Infrared Determination of Glucose in Whole Blood Using Multivariate Calibration’, *Appl. Spectrosc.*, vol. 46, no. 10, pp. 1575–1578, Oct. 1992.

[120] L. Zhang, G. W. Small, and M. a. Arnold, ‘Calibration standardization algorithm for partial least-squares regression: Application to the determination of physiological levels of glucose by near-infrared spectroscopy’, *Anal. Chem.*, vol. 74, no. 16, pp. 4097–4108, 2002.

[121] P. H. Garthwaite, ‘An Interpretation od Partial Least Squares’, *J. Am. Stat. ssociation*, vol. 89, no. 425, pp. 122–127, 1994.

[122] N. Krämer, ‘The Degrees of Freedom of Partial Least Squares Regression’, *J. Am. Stat. Assoc.*, vol. 106, no. 494, pp. 697–705, 2011.

[123] H. Chen, Q. Song, G. Tang, Q. Feng, and L. Lin, ‘The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models’, *ISRN Spectrosc.*, vol. 2013, pp. 1–9, 2013.

[124] S. Yazdani, J. Shanbehzadeh, and M. T. Manzuri Shalmani, ‘RPCA: A Novel Preprocessing Method for PCA’, *Adv. Artif. Intell.*, vol. 2012, pp. 1–7, 2012.

[125] 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.

[126] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, ‘Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra’, *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, May 1989.

[127] R. W. Kennard and L. A. Stone, ‘Computer aided design of experiments’, *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[128] R. Kohavi, ‘A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection’, *Int. Jt. Conf. Artif. Intell.*, vol. 14, no. 12, pp. 1137–1143, 1995.

[129] H. Li, Q. Xu, and Y. Liang, ‘libPLS an integrated library for partial least squares regression and discriminant analysis’, *PeerJ Prepr.*, vol. 2, p. e190v1, 2014.

[130] C. Chou, C.-Y. Han, W.-C. Kuo, Y.-C. Huang, C.-M. Feng, and J.-C. Shyu, ‘Noninvasive glucose monitoring in vivo with an optical heterodyne polarimeter’, *Appl. Opt.*, vol. 37, no. 16, pp. 3553–3557, Jun. 1998.

[131] V. V Tuchin, *Handbook of optical sensing of glucose in biological fluids and tissues*. CRC press, 2008.

[132] X. Guo, A. Mandelis, and B. Zinman, ‘Noninvasive glucose detection in human skin using wavelength modulated differential laser photothermal radiometry’, *Biomed. Opt. Express*, vol. 3, no. 11, pp. 3012–3021, 2012.

[133] H. M. Heise, A. Bittner, and R. Marbach, ‘Clinical chemistry and near infrared spectroscopy: technology for non-invasive glucose monitoring’, *J. Near Infrared Spectrosc.*, vol. 6, no. 1, pp. 349–359, 1998.

[134] W. Zhang, R. Liu, W. Zhang, H. Jia, and K. Xu, ‘Discussion on the validity of NIR spectral data in non-invasive blood glucose sensing’, *Biomed. Opt. Express*, vol. 4, no. 6, pp. 789–802, 2013.

[135] W. M. Doyle, ‘Near-IR and mid-IR process analysis-a critical comparison’, *Adv. Instrum. Control*, vol. 50, no. Part 1, pp. 433–441, 1995.

[136] Y. W. Dong *et al.*, ‘Determination of Soil Parameters in Apple-Growing Regions by Near- and Mid-Infrared Spectroscopy’, *Pedosphere*, vol. 21, no. 5, pp. 591–602, 2011.

[137] G. W. Small, M. A. Arnold, and L. A. Marquardt, ‘Strategies for coupling digital filtering with partial least-squares regression: Application to the determination of glucose in plasma by Fourier-transform near-infrared spectroscopy’, *Anal. Chem.*, vol. 65, no. 22, pp. 3279–3289, 1993.

[138] L. R. Rabiner and B. Gold, ‘Theory and application of digital signal processing’, *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, 1975.

[139] A. D. Jeyarani and T. J. Singh, ‘Analysis of noise reduction techniques on QRS ECG waveform-by applying different filters’, in *Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010*, 2010, pp. 149–152.

[140] F. Westad and F. Marini, ‘Validation of chemometric models - A tutorial’, *Anal. Chim. Acta*, vol. 893, pp. 14–24, 2015.

[141] I. T. Jolliffe, ‘Principal Component Analysis. Second Edition’, *Springer Ser. Stat.*, vol. 98, p. 487, 2002.

[142] O. A. Alrezj, K. Patchava, M. Benaissa, and S. A. Alshebeili, ‘Pre-processing to enhance the quantitative analysis of glucose from NIR and MIR spectra’, in *IFMBE Proceedings*, 2017, vol. 65.