JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DATA ANALYTICS – NEAR INFRARED SPECTRAL DATA

[1]Nazifi Lawal Bashir, [2]Sunday Tarekakpo Odobai

1. Department of Petroleum Resources
2. Niger Delta University

Abstract
Near-infrared, NIR, spectroscopy has found great applications in pharmaceuticals, food and petrochemicals through developing models that relate the spectral absorbance and sample property of interest. The classical approach to develop these models involves a univariate linear regression at a single selected wavelength. This research was aimed at developing data driven empirical models using several wavelengths for the prediction of active substance content in a pharmaceutical tablet from Near-Infrared spectral data. Prior to model computation, spectral data was pre-processed to remove unwanted spectral variations in the data and the traditional partial least squares, PLS, regression technique was used to develop benchmark models, with cross-validation number of components, used to evaluate the performance of the data-driven models. Artificial neural networks, ANN, is a data driven model computation method that can model complex datasets, this method was used to develop the empirical models. Pre-processing of the data showed significant effect in PLS and ANN models with improved model performance observed when data is first pre-processed before model computation. Models developed using ANN performed better than the models developed using PLS with higher correlation coefficient and lower root mean squared error of prediction. In this research a combination of PLS and ANN in computing multivariate models is proposed where the PLS predictor scores with reduced dimension are used as input for computing the ANN model. By comparison of the models developed, the combination of PLS and ANN resulted in the best model with coefficient of determination of 0.97 and root mean squared error of prediction of 0.22. This therefore illustrates the potential application of combination of PLS and ANN in modelling NIR spectral data.

Keywords
Machine Learning, Supervised Learning, Near Infrared Spectroscopy, Chemometrics, Partial Least Square, Artificial Neural Network.

1.0 Introduction

Near-infrared spectroscopy, NIRS, is an analytical technique widely applied to describe the characteristics samples in solid, semi-solid, fluid and vapour forms. The technique typically involves passing a beam of near-infrared radiation of the electromagnetic spectrum to probe samples. Most often the objective of this analysis is to predict the quantity or concentration of one or more content in a sample, referred to as quantitative analysis. In other cases, the objective is to identify or classify samples into groups, referred to as qualitative analysis. In the former case, a quantitative relationship, or model, is developed that can generalize the relationship between digitised sample NIR spectra – denoted as x matrix – and the quantity or concentration of content of interest in the sample – denoted as y matrix (Chen, et al., 2007). In other applications, the quantity of interest to be determined be a physical parameter such as strength and viscosity of a polymer, thickness of a tablet coating, hardness of a tablet, or in gasolines, octane number.

NIRS has a universal response characteristic and for that reason, the technique has found applications in diverse areas, different analytical purposes and almost any type of samples (Pasquini, 2018).

The development of quantitative relation, or model, is achieved by using regression methods with the first step mostly being to pre-process the spectral data. Spectral data is the result of NIRS and quite often, this data comprise of both useful information and systematic variations that is not relevant to the responses y. The objective of pre-processing is to remove these unwanted systematic variations and minimize spectral variability that is contained in the spectra which is not relevant to the purpose of the quantitative model. These additional variations in x that do not relate to y add unwanted information to the modelling process and thus reduce model accuracy.

Regression methods, also known as calibration methods are used to develop models from pre-processed spectra data for the purpose of extracting useful information using mathematical and statistical methods, this approach is generally referred to as chemometrics. Once regression models are developed, they can be used to predict the quantity or concentration of interest from an independent set to evaluate the extent of model generalization.

Another approach to developing quantitative models is through machine learning methods, using algorithms and statistical tools to efficiently model a system without need for explicit knowledge about the system.

In recent decades, there has been tremendous developments in the field of computational intelligence, leading to advances in artificial intelligence and machine learning. Machine learning methods can be used to compute empirical models, models derived from purely from specific sets of experimental data, have great capability to analyse data about a system and establish a generalized relationship between the state variables (input, internal and output variables) of the system with no need for any information on the physical behaviour of the system. These empirical models are also known as data-driven models.

Artificial neural networks, ANN, is a data-driven modelling approach originally developed to simulate the behaviour of biological neural networks in processing signals. ANN model data by using a machine learning algorithm to approximate a relation between a system's inputs and outputs, a process known as training were the data used, to a great extent covers, the variation in the system. The computed model is then tested using a separate set of data, data not used during training from the same system to evaluate the extent to which the model generalised the relation.

NIRS instruments typically provide huge amount of very high dimensional data as results that need fast and efficient processing to extract useful analytical information from the spectral data (Blanco & Peguero, 2010). Classically, NIR spectral data is modelled by the traditional univariate method, this typically involves developing a linear regression model to predict sample property of interest using one selected wavelength with obvious peak and assuming a linear relationship between the measured spectral response x and the known reference value y (Naes, et al., 2002). This may not describe the actual underlying relation between the spectral response and the reference values of the samples (Wiesner, et al., 2014).

NIR spectra are primarily characterized by several overlapping and broad peaks due to the presence of more than one absorber in the sample. Using a single wavelength absorbance to develop a calibration model for predicting the concentration one absorber is not always possible and because of this, it very difficult to interpret the data and extract information using the univariate approach. Measurements based in NIRS typically consist of information

relating to the sample and the instrument used as well with several physical and chemical effects (Naes, et al., 2002). These effects may be in the form of external disturbance like light scattering effects and instrumental variation of noise (Wang, et al., 2011). This therefore requires extensive application of multivariate regression methods to efficiently analyse data and extract useful information. Multivariate regression techniques, using several wavelengths, provide better optimal model by combining the information from all the wavelengths (Naes, et al., 2002).

Again, each individual wavelength in the spectrum contains information some certain, using only one selected wavelength for developing models is not ideal because some individual wavelengths in the complete spectrum may be affected by noise or other spectral variability. Therefore, by averaging the information in all the wavelengths a more robust model often results, providing a better result than the univariate approach (Brereton, 2000).

Data-driven models have the advantage of being non-parametric with ability to implicitly capture complex non-linear relationships between predictor and response variables. If sufficient data is available, an ANN model can approximate any "smooth" relationship between dependent and independent variables to any accuracy (Chen, 2018).

3.0 Methodology

The methodology adopted in developing calibration models for predicting the quantity of active substance in a pharmaceutical tablet is summarized in figure 1. The activities involved in model development can be broadly grouped in laboratory level activities and computation level. Complete experimental procedure for sample preparation, NIR spectroscopy measurements and reference method analysis can be found in the published journal by (Dyrby, et al., 2002).

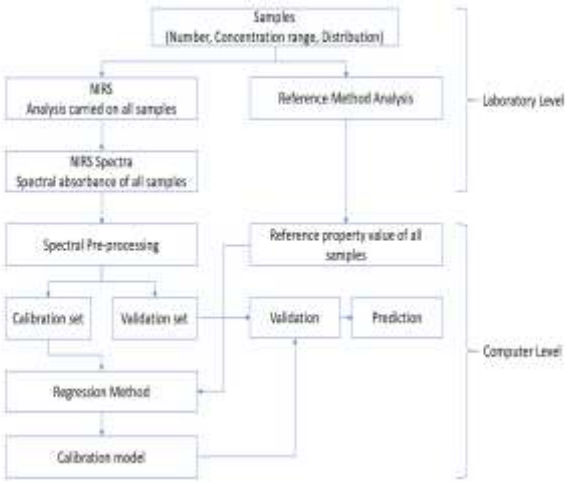

Figure 1: Steps involved in calibration model development

3.1 Data Collection

The data used in this work is NIR transmittance measurement of 30 tablets, Escitalopram ® from H. Lundbeck A/S pharmaceutical company. In total, the spectra contain 404 variables present for 310 samples at the wavelength range 7400 – 10507 cm^{-1} with a resolution of 16 cm^{-1} averaging 128 scans per sample. The data was generated during the work of (Dyrby, et al., 2002) and they used ABB Bomem FT-NIR model MB-160 spectrometer to record the spectra. The spectrometer was equipped with a presenting device purposely for pharmaceutical tablets and a detector made of 1.7-µm InGaAs.

Four different dosages of tablets, 5, 10, 15 and 20 mg, were used in the work and although each of them have the same concentration their shapes and sizes were different. Analysis was carried on 31 batches each of them having 10 tablets and (Dyrby, et al., 2002) reported the presence of excipients in the tablets mainly microcrystalline cellulose. Other excipients present in smaller quantity are magnesium stearate, talc and a material containing titanium dioxide.

Dyrby, et al. (2002) recorded the background resistance with the internal ceramic standard, Spectralon® 99%, and because measurement was done in transmittance mode, tablet spectra was converted to absorbance units using the background transmittance. With Spectralon® 99% standard, the range of intensities recorded by the detector was narrow because the intensity level of the reference spectrum and the intensity level of the sample spectrum were within the same range.

3.2 Reference Method analysis

Reference values of quantities of active substance in the tablet samples are needed for any calibration model building. For this case, (Dyrby, et al., 2002) used high performance liquid chromatography (HPLC) to evaluate the values. The reference method has a combined sampling and apparatus error estimated as ±3.5% and the reference values of the active substance content were given in mg. The weight percent, %w/w, was evaluated by dividing the reference active substance content by the weight of the tablet.

3.3 Model Development

3.3.1 Data Sampling

For proper calibration, spectra data set must be divided into subsets, calibration and validation sets, which determine the performance of any subsequent model. The specification of the calibration or training set is a critical step in model development (Alam, et al., 2017) because a lot of considerations need to be done to ensure that the calibration set contains as many samples as possible and include all variations such that it represents the complete samples dataset. Generally, the predictive ability of developed models is improved by an increased number of samples in the calibration set (Porep, et al., 2015).

Because it is difficult to determine the samples to be included in the calibration set (Roggo, et al., 2007), the calibration and validation set were partitioned using the Kennard-Stone algorithm (Kennerd & Stone, 1969) selecting 217 samples in the training set, representing about two thirds of the dataset, and 93 samples in the validation set. The algorithm selects the samples considered the most representative of the whole data set by choosing the sample with close value to the mean of the set followed by sequential selection of samples that with the highest distances of projection form the previously selected samples.

Drawbacks of data partitioning have already been discussed in the literature and because of that, 10-fold cross-validation was also used to validate the models with associated root mean squared error of cross-validation, RMSECV.

3.3.2 Pre-processing

Pre-processing of data is a very critical step in chemometric modelling that influences the final prediction ability of the model. Variables measured in spectral data have high variations and a relationship between them, the main aim of pre-processing is to adjust these variations and the relationship such that the data is now more suited to the multivariate analysis goals.

Multivariate calibration models were developed to predict the reference values of active substance content using recorded spectral response. With many different methods available to perform data pre-processing, there is no distinct procedure in selecting a certain method for a multivariate analysis, eventual choice depending on several different properties of the dataset (Engel, et al., 2013).

In this work, combinations of several pre-processing techniques were applied to the spectral data to get the model with the best predictive ability. Normalization and as spectral derivatives were all applied to the spectra singly or in combination to test prediction accuracy.

3.3.3 Multivariate Regression Analysis

Calibration may be considered as a form of regression with three basic aims (Brereton, 2018):

i. Form a simple functional model between two sets of measurement,
ii. Validate the model
iii. Make future predictions using the model

3.3.3.1 Partial least squares regression models

Partial least squares, PLS, regression was used to develop calibration models for the prediction of active substance content in a pharmaceutical tablet from NIR spectral data. For selecting the optimum number of factors or components in the model, the ten-fold-cross validation was used. This technique uses the root mean square error of cross validation, RMSECV, to specify the number of components for which the error is minimum thereby optimizing the model (Bleye, et al., 2012).

Raw and pre-processed data, using different pre-processing techniques, were used to compute the PLS models. The RMSECV and coefficient of determination was calculated for all models. The range of wavelength 7400 – 10507 cm^{-1} was used throughout all model computations.

The PLS regression algorithm assumes a linear relationship between measured spectral response $X_{n,p}$ and concentration of analyte $y_{n,1.}$ Where n is the number of samples and p is the number of variables or wavelengths (Mehmoud, et al., 2012). The equation may be written as:

$$y = \alpha + X\beta \qquad 24$$

The terms α and β represents the regression parameters and error term respectively. For a single response case, the variables are first cantered and scaled such that:

$$X_0 = X - 1\bar{x}' \quad 25$$

$$y_0 = y - 1\bar{y} \quad 26$$

For a number of components 'A' that are relevant for prediction by the model (Naes & Holland, 1993), where the number must be less than the number of variables, p, the models runs for a = 1, 2, 3…., A in the following steps:

1. Calculation of loading weights.
   $w_a = X'_{a-1}y_{a-1}$ 27
   The weights calculated describe the direction of maximum covariance with y_{a-1} that is covered by X_{a-1}. The loading weights are normalized such that their length equals 1.
2. Calculation of the score vector t_a:
   $t_a = X_{a-1}w_a$ 28
3. Regression of variables in X_{a-1} on the score vector to compute X-loadings, p_a:
   $p_a = X'_{a-1}\frac{t_a}{t'_a t_a}$ 29
   The Y-loadings, q_a, are computed in the same way:
   $q_a = y'_{a-1}\frac{t_a}{t'_a t_a}$ 30
4. Reducing X_{a-1} and y_{a-1} by removing the contribution of the score vector t_a:
   $X_a = X_{a-1} - t_a p'_a$ 31
   $y_a = y_{a-1} - t_a q_a$ 32
5. This is continued by returning to step 1 for all $a < A$.

If the matrix $W = [w_1, w_2, \dots, w_A]$ represents the loading weights, $T = [t_1, t_2, \dots, t_A]$ the scores, $P = [p_1, p_2, \dots, p_A]$ the X-loadings, and $q = [q_1, q_2, \dots, q_A]$ the Y-loadings; regression coefficients are estimated by:

$$\hat{\beta} = W(P'W)^{-1}q \quad 33$$

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} \quad 34$$

This can be generalized for multiple response regression by introducing a loading matrix $Q = [q_1, q_2, \dots, q_A]$ instead of the Y-loadings.
At the present time partial least squares (PLS) and its modifications are the most widely used, considered standard (Pasquini, 2003), multivariate regression method for analytical data analysis based on NIR spectroscopy (Pasquini, 2018).

3.3.3.2 Artificial neural networks models

Artificial neural networks, ANN, models were also built using the pre-processed spectral data. The artificial neural network fitting app in the software used allows specification of input, spectral data; and target (referred to as output onwards), reference values of active substance quantities, data. The app also allows varying the number of hidden neurons and of training iterations for model learning to compare performance of developed models. The app also allows varying the learning algorithm to be used for training, throughout this work the Levenberg-Marquardt algorithm was used. The algorithm stops training at the minimum mean squared error of validation. The software, which randomly divides the input data into training, validation and testing sets, provides analyses of mean squared errors and regression plots for the training, validation and testing sets. Sampling ratio of 70%, 15% and 15% was used for training, validation and testing sets respectively.
The predictor scores output of the best PLS models were also used as inputs for the ANN to observe the effect of reduced factors by the PLS on the performance of subsequent ANN models.

3.3.4 Model Performance

Root mean squared error of prediction, RMSEP, for which root mean squared error of cross-validation, RMSECV, was also used, was computed for PLS models to evaluate their performance in predicting active substance quantity for the validation set samples not used in the during the calibration. The coefficient of determination, R^2, was also computed to determine the extent to which the model covers the variation in the data.
ANN model's performance was evaluated by comparing the mean squared error, MSE, which is the average squared difference between predicted values and reference values; and the correlation coefficient, R.

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad 35$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{val}}(\hat{y}_i - y_i)^2}{n_{val}}} \quad 36$$

In the expression, $\hat{y}_i$ is the property value predicted by the calibration model, y_i is the property value from a reference method for the sample, n_{val} is the number of samples in the validation set.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

37

Where, $\bar{y}_i$ is the mean value of the reference property values.

3.3.5 Software and Computing

Data pre-processing, normalization and Savitzky-Golay derivatives; PLS models and ANN models were all computed using MATLAB R2018, 64-bit (The MathWorks, Inc.).

4.0 Results and Discussion

Data driven empirical models were developed for the prediction of active substance content in a pharmaceutical tablet. Partial least squares regression models, PLS, and artificial neural networks models, ANN, were developed to use NIR transmittance spectra as the independent variables and predict active substance content as the dependent variables. Reference values of the dependent variables were measured from the method of high-performance liquid chromatography (Dyrby, et al., 2002) and these were used to evaluate the accuracy of the developed models.
The available dataset contains recorded NIR transmittance within the range 7400 – 10500 cm^{-1}. Figure 2 shows the spectra of a 20mg tablet and that of the pure active substance as measured by (Dyrby, et al., 2002) with a characteristic visual band associated with the active substance attributed to the second overtone of the aromatic C – H stretch. This can be observed by the peak of the active substance spectra at 8830 cm^{-1}, partially overlapping with the one at 8200 cm^{-1}, as a result of the primary excipient – microcrystalline cellulose (Dyrby, et al., 2002). The reference values of the active substance content is characterized by a mean of 7.4282mg, a median of 7.9589mg and a standard deviation of 1.2954mg.

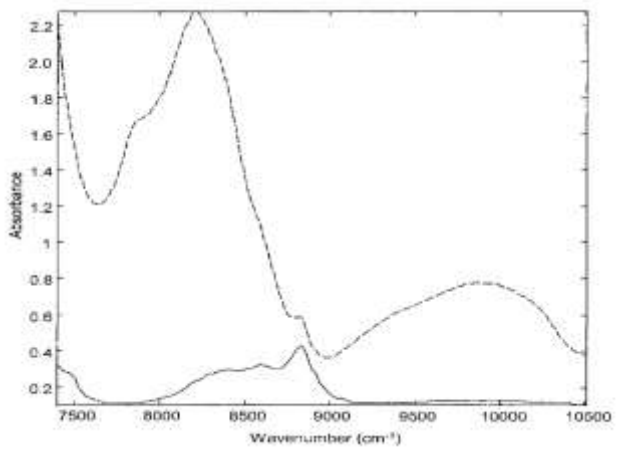

Figure 2: NIR raw spectra of a 20 mg tablet represented by the dashed line, and the pure active substance represented by the solid line (Dyrby, et al., 2002).

4.1 Data Pre-processing

4.1.1 Normalization

Raw NIR of the obtained data is shown in figure 3 and normalization pre-processing was applied to the raw data. Mean centering was applied to the raw spectral data set because it is one of the most straightforward and common method to transform the data (Kuhn & Johnson, 2013). Data scaling on the other hand was done by dividing through values of the predictor variables by the standard deviation. This was carried out to ensure numerical stability of subsequent calculations.

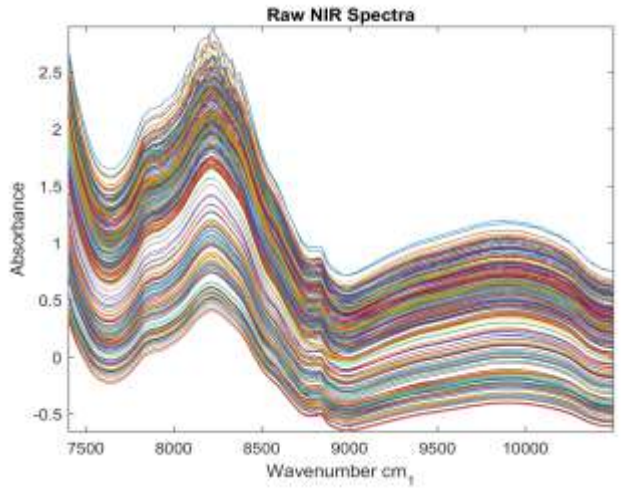

Figure 3: Raw NIR spectra of 310 samples at range 7400 – 10507 cm^{-1}.

The consequence of normalization is shown is shown in figure 4, This in effect subtract the mean from all predictor variables values thereby resulting in the predictor variables to have zero mean, also bringing the standard deviation to one for all predictor variables.

4.1.2 Spectral derivatives

Savitzky-Golay spectral derivative filter was applied to the raw data for smoothing, both first-derivative and second-derivative showed effective transformation of the spectra as shown in figure 5 and 6 respectively. Because the combination of more than one pre-processing technique is feasible, and, Savitzky-Golay first derivative and normalization showing great suitability (Blanco & Peguero, 2010), these combinations showed varying effects and the position of the overtone representing the active substance content. Figures 7 and 8 show the application of combinations of normalization and Savitzky-Golay first-derivative and second-derivative smoothing to the raw data.

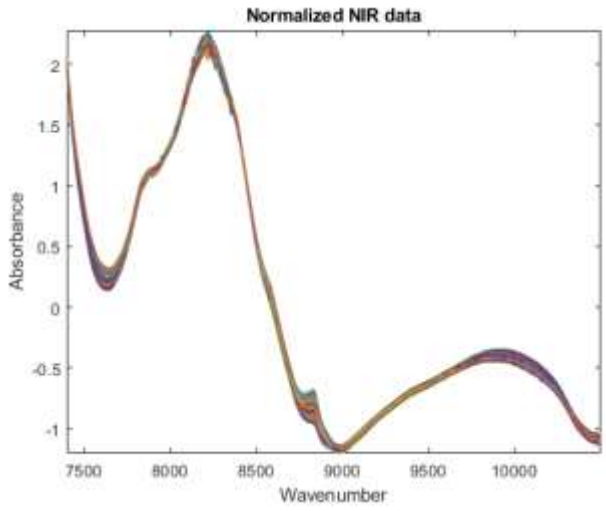

Figure 4: Application of Normalization on the raw data.

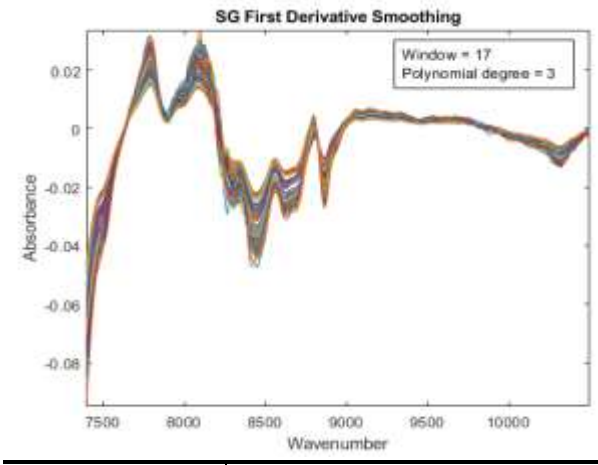

Figure 5: Savitzky-Golay first-derivative smoothing.

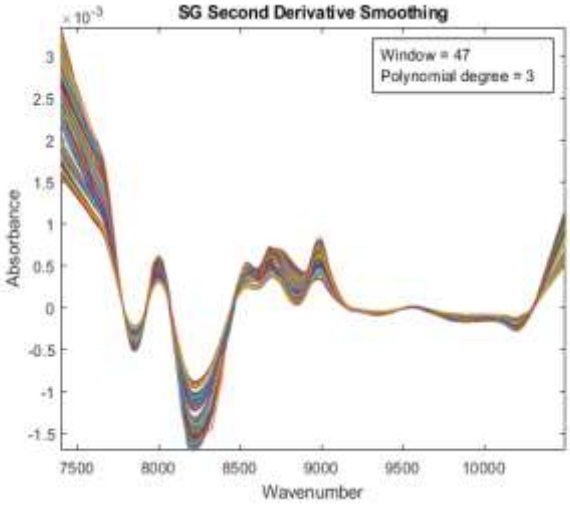

Figure 6: Savitzky-Golay second-derivative smoothing.

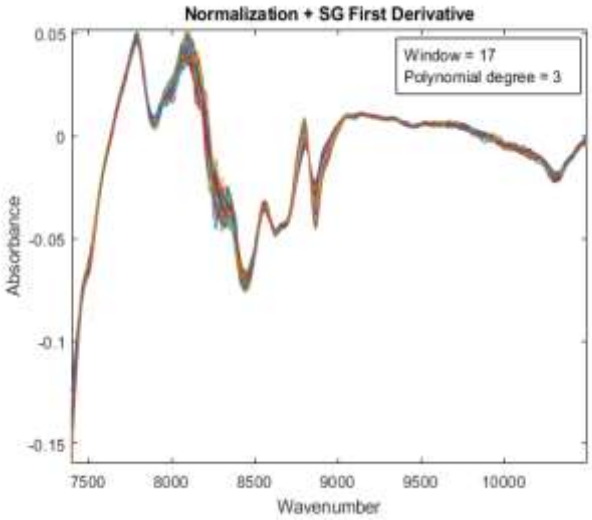

Figure 7: Normalization followed by Savitzky-Golay first-derivative smoothing.

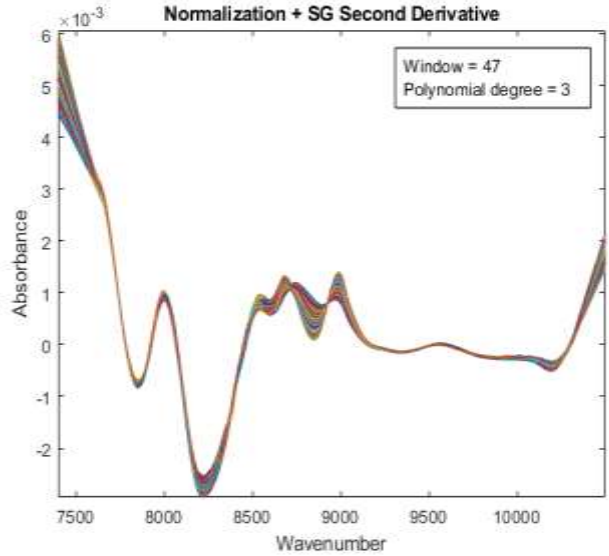

Figure 8: Normalization followed by Savitzky-Golay second-derivative smoothing.

4.2 Partial Least Square Models

4.2.1 PLS Regression models validated by cross validation

Partial least square, PLS, regression models were developed by using 10-fold cross validation over the complete dataset to strategically determine the appropriate number of components used in the model through minimised estimated prediction error. Figure 9 shows a typical number of components that ensures minimised prediction error. Figures 10 - 15 show PLS regression models developed using number of components determined by cross validation.

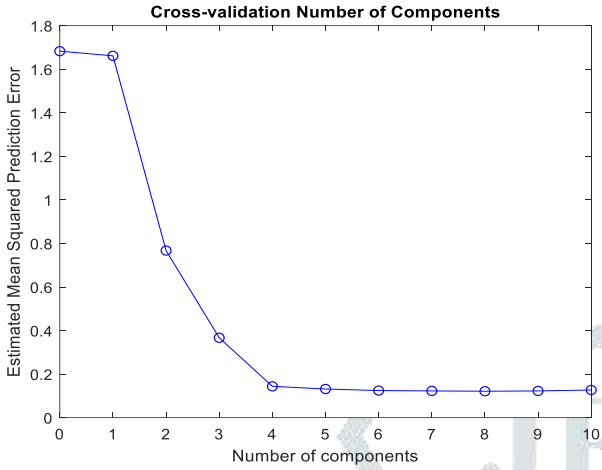

Figure 9: Selecting the number of components using cross validation over whole dataset.

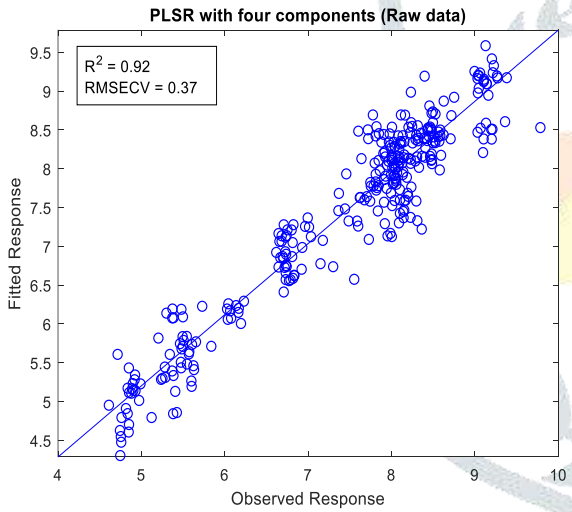

Figure 10: Four component PLS Regression model using raw NIR data.

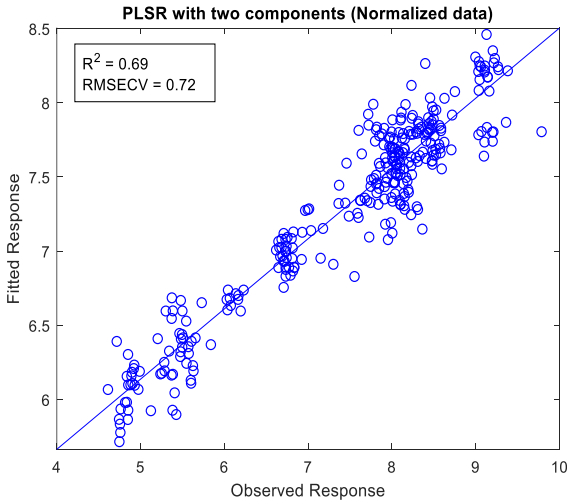

Figure 11: Two component PLS Regression model using normalized data.

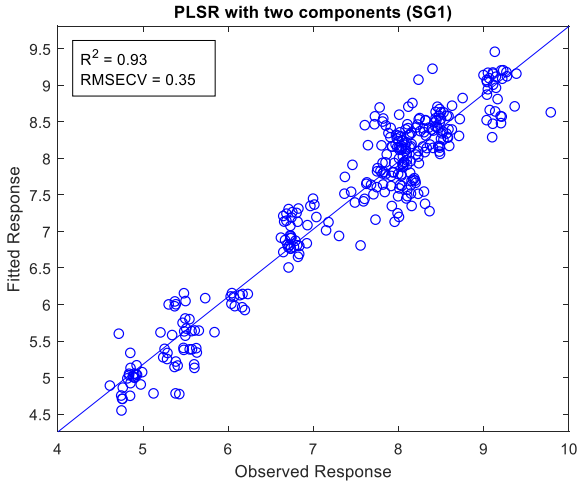

Figure 12: Two component PLS Regression model using data pre-processed by Savitzky-Golay first-derivative smoothing.

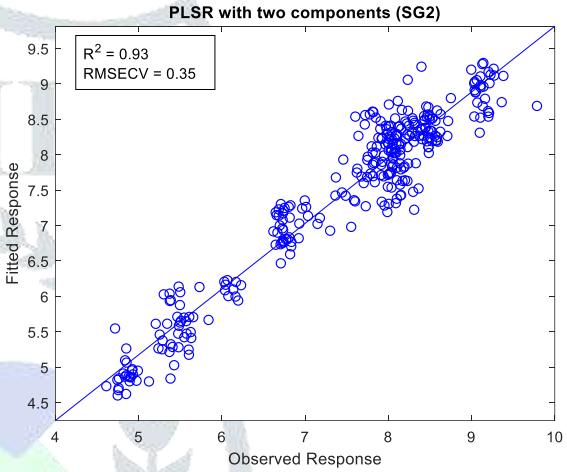

Figure 13: Two component PLS Regression model using data pre-processed by Savitzky-Golay second-derivative smoothing.

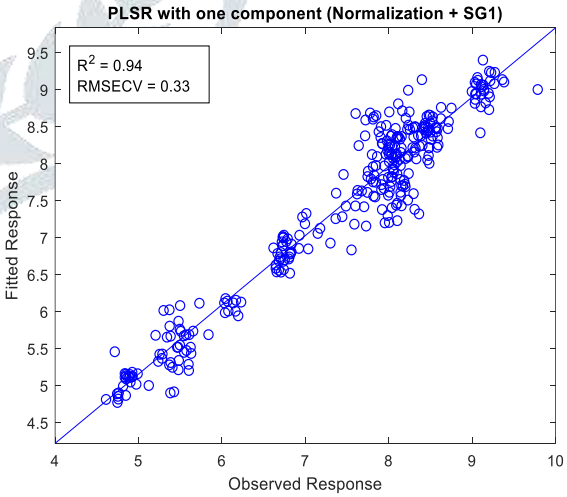

Figure 14: Two component PLS Regression model using data pre-processed by normalization followed by Savitzky-Golay first-derivative smoothing.

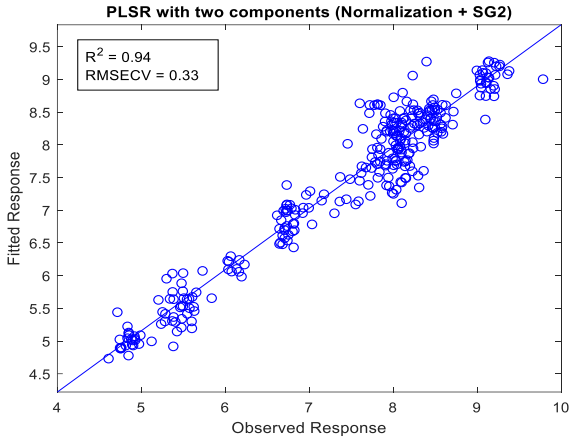

Figure 15: Two component PLS Regression model using data pre-processed by normalization followed by Savitzky-Golay second-derivative smoothing.

4.2.2 PLS Regression models validated by test set

With the dataset partitioned into training set and test set using the Kennard-Stone sampling algorithm, models were developed using training set, and the subsequent models were validated using test sets that were not used in the model construction. Figures 16 – 21 show the models developed and validated using a test set.

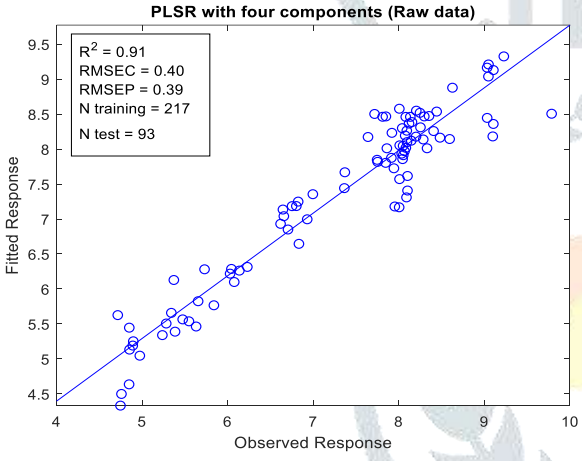

Figure 16: Four component PLS Regression model using raw data partitioned into training and test sets.

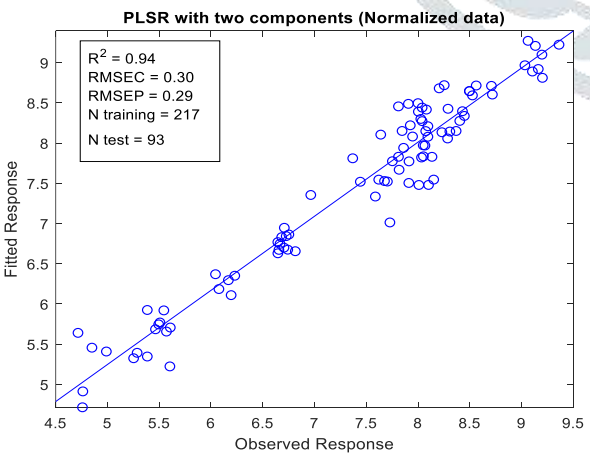

Figure 17: Two component PLS Regression model using normalized data partitioned into training and test.

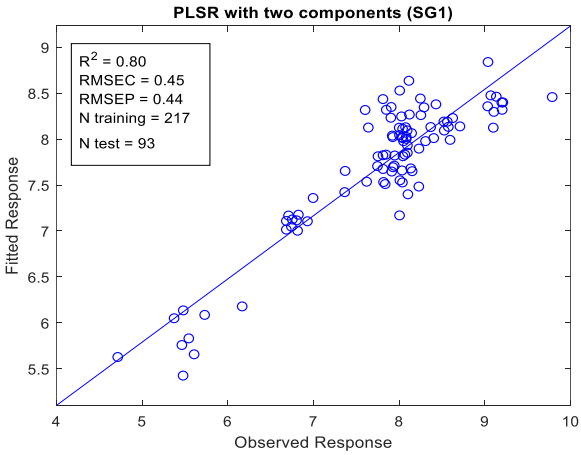

Figure 18: Two component PLS Regression model using pre-processed (Savitzky-Golay first-derivative smoothing) data partitioned into training and test.

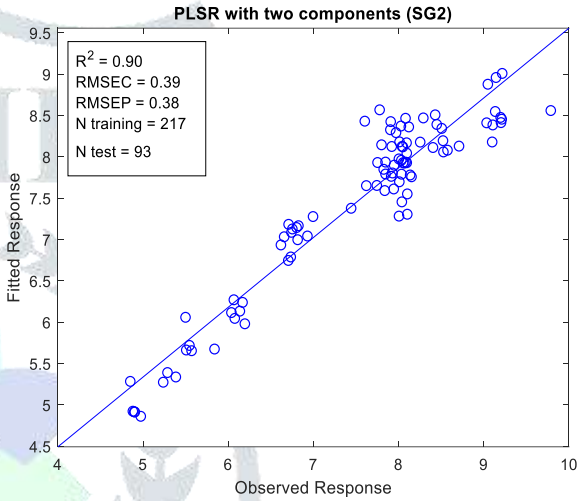

Figure 19: Two component PLS Regression model using pre-processed (Savitzky-Golay second-derivative smoothing) data partitioned into training and test.

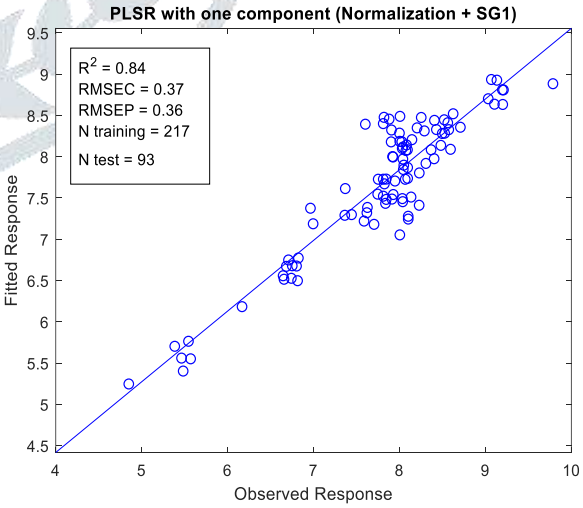

Figure 20: One component PLS Regression model using pre-processed (Normalization followed by Savitzky-Golay first-derivative smoothing) data partitioned into training and test.

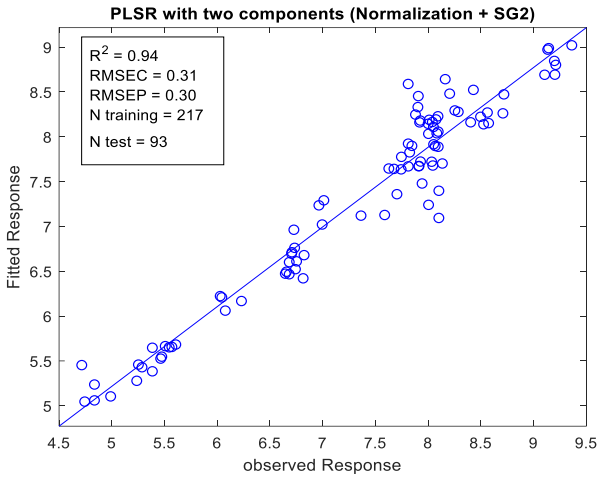

Figure 21: Two component PLS Regression model using pre-processed (Normalization followed by Savitzky-Golay second-derivative smoothing) data partitioned into training and test.

Table 1: Comparison of different pre-processing methods on the performance of PLS Regression models.

Models Validated by Cross-Validation				Partitioned Data		
Pre-processing	Comp's	RMSECV	R^2	RMSEC	RMSEP	R^2
None	4	0.37	0.92	0.40	0.39	0.91
Normalization	2	0.72	0.69	0.30	0.29	0.94
S-G first-derivative	2	0.35	0.93	0.45	0.44	0.80
S-G second-derivative	2	0.35	0.93	0.39	0.38	0.90
Normalization + S-G first-derivative	1	0.33	0.94	0.37	0.36	0.84
Normalization + S-G second-derivative	2	0.33	0.94	0.31	0.30	0.94

4.3 Artificial Neural Networks Models

Artificial neural networks, ANN, models developed are based on the Levenberg-Marquardt training algorithm, and because the fitting allows for specification of network architecture, number of hidden neurons in the network were altered to check the effect on model performance, the number of hidden neurons were changed from 5 – 20. Raw tablets data as well as pre-processed data were used as input to the app while reference values for the active substance content were used as the output, predictor scores form PLS regression were also used as input. Figures 22 – 30 show ANN models developed using ten hidden neurons and pre-processed data.

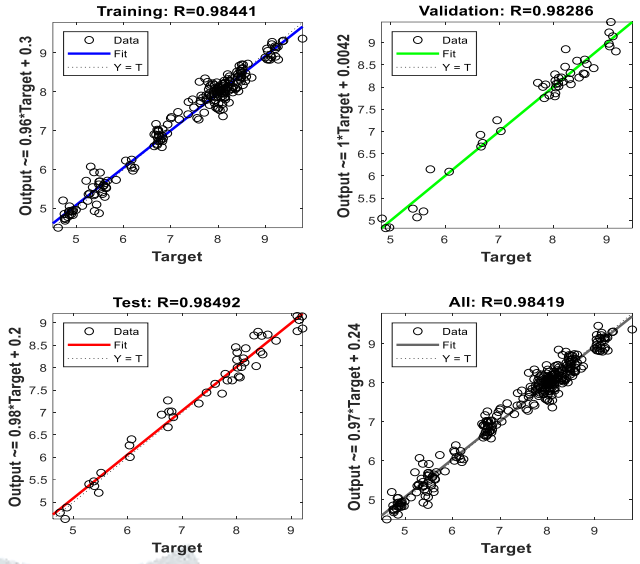

Figure 22: ANN model with five hidden neurons using raw predictor data as input

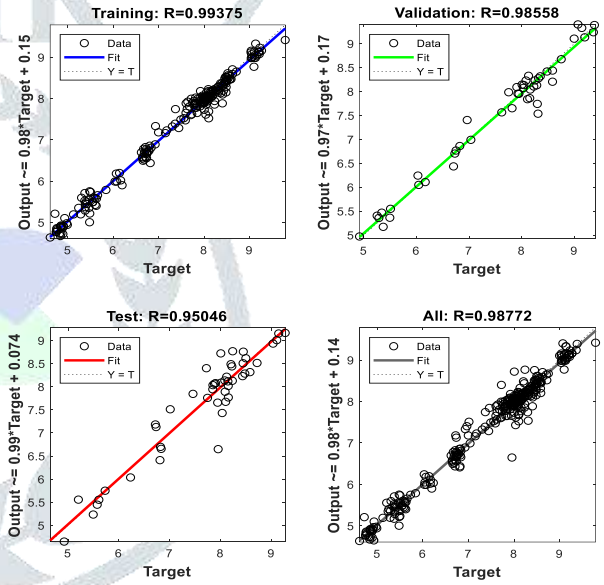

Figure 23: ANN model with ten hidden neurons using raw predictor data as input

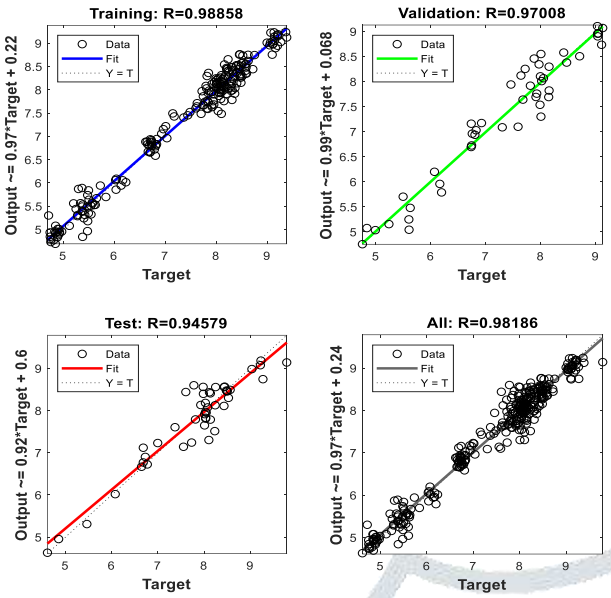

Figure 24: ANN model with twenty hidden neurons using raw predictor data as input

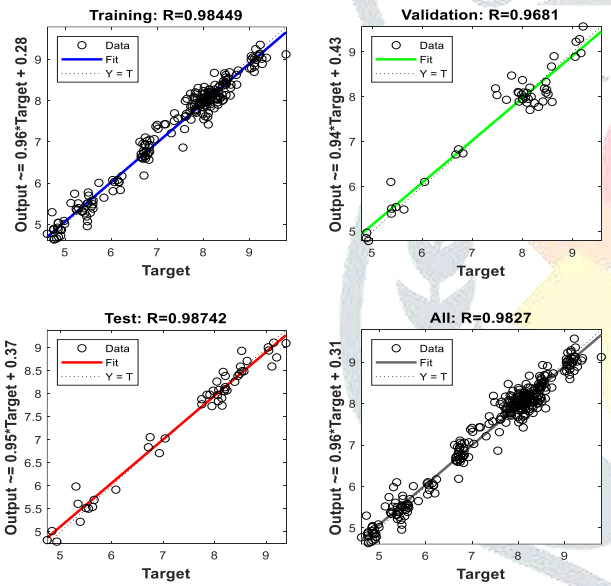

Figure 25: ANN model developed using four component PLS predictor scores of raw data as input

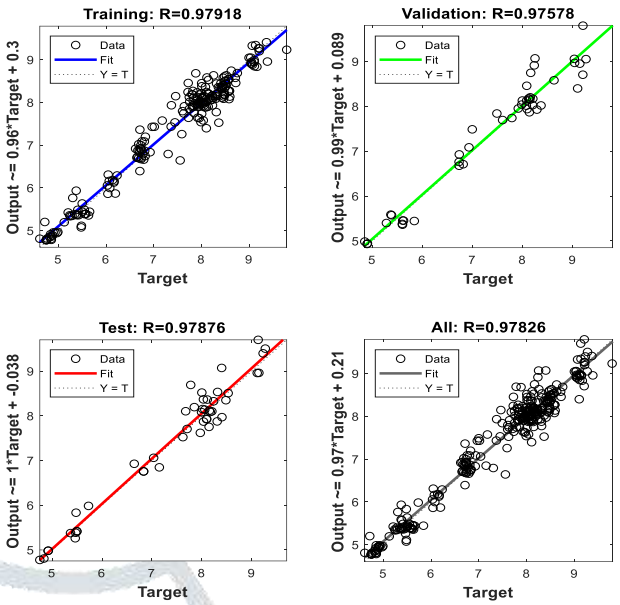

Figure 26: ANN model developed using two component PLS predictor scores of normalized data as input.

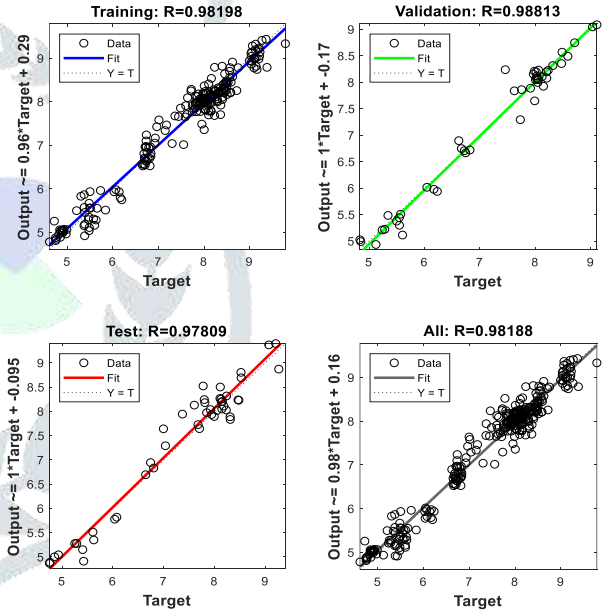

Figure 27: ANN model developed using two component PLS predictor scores of pre-processed (Savitzky-Golay first-derivative smoothing) data as input.

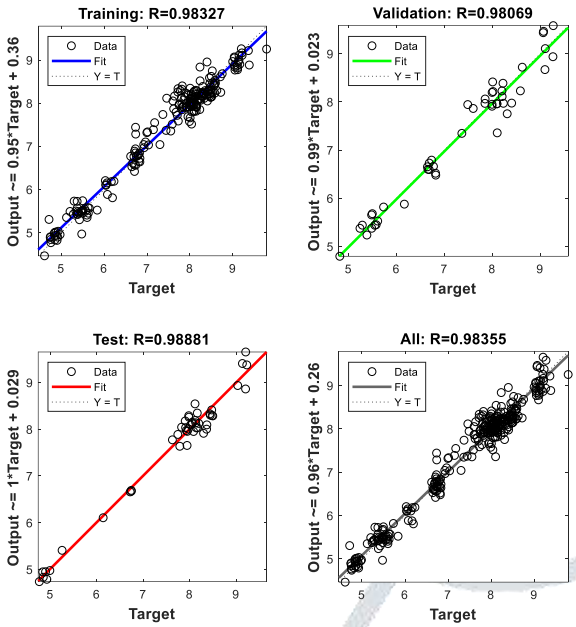

Figure 28: ANN model developed using two component PLS predictor scores of pre-processed (Savitzky-Golay second-derivative smoothing) data as input

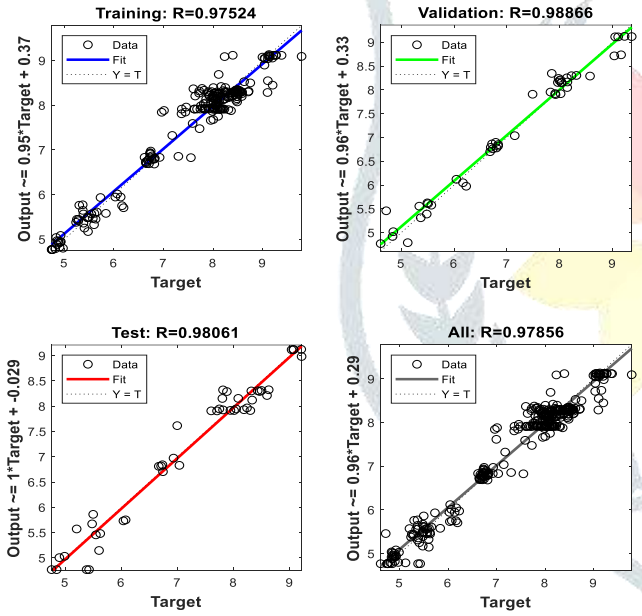

Figure 29: ANN model developed using one component PLS predictor scores of pre-processed (normalization followed by Savitzky-Golay first-derivative smoothing) data as input

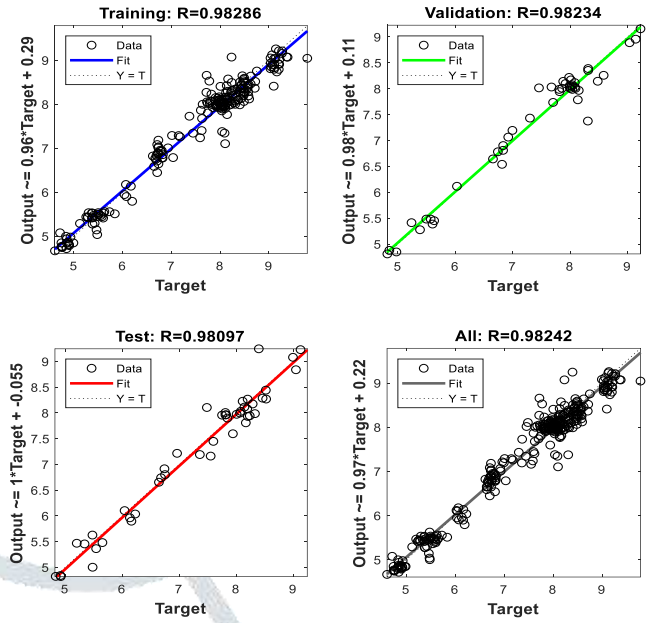

Figure 30: ANN model developed using two component PLS predictor scores of pre-processed (normalization followed by Savitzky-Golay second-derivative smoothing) data as input.

Table 2: ANN model performance using five hidden neurons and raw predictor data as input

	Samples	MSE	RMSE	R	R^2
Training	216	0.0530	0.23	0.9844	0.97
Validation	47	0.0515	0.23	0.9829	0.97
Testing	47	0.0508	0.23	0.9849	0.97

Table 3: ANN model performance using ten hidden neurons and raw predictor data as input

	Samples	MSE	RMSE	R	R^2
Training	216	0.0234	0.15	0.9938	0.99
Validation	47	0.0450	0.21	0.9856	0.97
Testing	47	0.0121	0.11	0.9505	0.90

Table 4: ANN model performance using twenty hidden neurons and raw predictor data as input

	Samples	MSE	RMSE	R	R^2
Training	216	0.0410	0.20	0.9886	0.98
Validation	47	0.0925	0.30	0.9701	0.94
Testing	47	0.1160	0.34	0.9458	0.89

Table 5: ANN model performance using four component PLS predictor scores of raw data as input

	Samples	MSE	RMSE	R	R^2
Training	216	0.0520	0.23	0.9845	0.97
Validation	47	0.0972	0.31	0.9681	0.94
Testing	47	0.0464	0.22	0.9874	0.97

Table 6: ANN model performance using two component PLS predictor scores of normalized data as input.

	Samples	MSE	RMSE	R	R^2
Training	216	0.0675	0.26	0.9792	0.96
Validation	47	0.0872	0.30	0.9758	0.95
Testing	47	0.0812	0.28	0.9788	0.96

Table 7: ANN model performance using two component PLS predictor scores of pre-processed (Savitzky-Golay first-derivative smoothing) data as input.

	Samples	MSE	RMSE	R	R^2
Training	216	0.0601	0.25	0.9820	0.96
Validation	47	0.0406	0.20	0.9881	0.98
Testing	47	0.0816	0.29	0.9781	0.96

Table 8: ANN model performance using two component PLS predictor scores of pre-processed (Savitzky-Golay second-derivative smoothing) data as input.

	Samples	MSE	RMSE	R	R^2
Training	216	0.0561	0.25	0.9833	0.97
Validation	47	0.0659	0.20	0.9807	0.96
Testing	47	0.0376	0.29	0.9888	0.98

Table 9: ANN model performance using one component PLS predictor scores of pre-processed (normalization followed by Savitzky-Golay first-derivative smoothing) data as input.

	Samples	MSE	RMSE	R	R^2
Training	216	0.0759	0.28	0.9752	0.95
Validation	47	0.0443	0.21	0.9889	0.98
Testing	47	0.0750	0.27	0.9806	0.96

Table 10: ANN model performance using two component PLS predictor scores of pre-processed (normalization followed by Savitzky-Golay second-derivative smoothing) data as input.

	Samples	MSE	RMSE	R	R^2
Training	216	0.0605	0.25	0.9829	0.97
Validation	47	0.0498	0.22	0.9823	0.96
Testing	47	0.0603	0.25	0.9810	0.96

4.4 Discussion

4.4.1 PLS Regression models

As discussed in the literature pre-processing of NIR data transforms the data such that important information extraction from the data is maximized by eliminating sources of redundancy and unwanted information. PLS models benefit from the predictors being on a common scale (Kuhn & Johnson, 2013) and raw data was normalized, as shown in figure 4, by mean centering and standard deviation scaling. Additive and multiplicative effects in the data were treated by the Savitzky-Golay first and second-derivative smoothing (Pasti, 1997), for specifying the parameters window size and degree of polynomial, this was done by checking the effect of various combinations of window size and polynomial degree and observing improvements in the spectra. Polynomial degree 3 was chosen for both first and second derivatives because the improvement in results deteriorate above degree 3, window size 17 and 47 were also chosen respectively in the same manner. Figures 4 to 8 show the effect of applications of pre-processing techniques to the original data. The effect of combinations of pre-processing techniques on PLS model's performance was investigated and the results are shown in table 1.

PLS regression models were developed based on the variables in the full spectrum, 7400 – 10500 cm^{-1}, because this interval contains variables with important spectral information (Dyrby, et al., 2002). Raw data, as well as pre-processed data was used to compute the models, in each case, 10-fold cross-validation used to select the number of components (figure 9). The selection of number of components by cross-validation becomes imperative because of the need to minimize error of prediction. Whereas an insufficient number of components means some relevant information have been ignored by the model and thus under fits the model; a higher than optimal number of components may usually contain irrelevant information in the model, causing over fitting in the model.

For the purpose of PLS model validation, two scenarios were considered. The first scenario involves model validation by 10-fold cross-validation with associated errors of cross-validation, and the second scenario involves validation by data sampling into training (217 samples) and validation sets (93 samples) based on the Kennard-Stone sampling algorithm, where PLS models were computed using the training set and validating the models by predicting the active substance content in the test set, having associated calibration and prediction errors. Models validated by cross-validation showed better performance in prediction and have lower associated errors of prediction (Table 1) than models validated by a test set. This attributed to the fact that cross-validation ensures that all samples are used during the training and validation at least once. The best PLS model was obtained when raw NIR data was pre-processed by normalization followed by a Savitzky-Golay first derivative smoothing having the lowest root mean square error of cross–validation, 0.33, and the highest coefficient of determination, 0.94, with only one component, the lowest possible number of components for the PLS model. Caution is usually exercised when dealing with one component PLS model (Williams, et al., 2017), but this may be justified by the presence of one dominant absorber in the tablet.

Observed response, the reference values for active substance content, versus fitted response, the values of active substance content calculated by the model, plots for PLS models computed from partitioned data are shown in figures 16 – 21. These models have associated root mean square error of calibration, RMSEC, and root mean square error of prediction, RMSEP. The models are computed for both raw data and pre-processed data, and the best model was obtained with data pre-processed by only normalization, a two component PLS model with RMSEC of 0.29, RMSEP of 0.29 and coefficient of determination of 0.94.

Comparing the two scenarios, 10-fold cross-validation holds the advantage of repeated training of the PLS models thereby acquiring more information on the data as a result, and thus has better model performance characteristics. On the other hand, PLS models computed from partitioned data are only limited to data information from the training set therefore the test set used to validate the model is have predictor variables completely unknown to the model.

4.4.2 Artificial neural networks models

Artificial neural networks, ANN, models were developed to predict the quantity of active substance (the output layer of the ANN model) in a pharmaceutical tablet using NIR spectral data (the input layer of the ANN model). The software used in this research allows for specification of ANN architecture be defining the number of hidden neurons, but selection of the optimal number hidden neurons remains largely based on trial and error method (Chen, et al., 2001). Generally, a low number of hidden neurons results into low prediction performance of the ANN model while a higher number of neurons may result in higher accuracy of prediction but then this is only significant for the root mean square error, RMSEP, for predicting the training set. A high number of hidden neurons may cause overfitting of the model (Kuhn & Johnson, 2013). The effect of varying the number of hidden neurons can be seen in tables 2 – 4 where raw NIR data was used as input, with the coefficient of determination for the testing samples reducing from 0.97 for 5 hidden neurons, to 0.90 for ten hidden neurons, and to 0.89 for twenty hidden neurons indicating possible tendency of model over fitting. When 10 hidden neurons were used and the raw NIR data as input to the input layer the computed model has the lowest RMSEP of 0.11.

PLS predictor scores from raw, and pre-processed data were also used as inputs to the input layer of the ANN with 10 hidden neurons, this is observed as shown in tables 5 – 10. Because the PLS predictor scores have reduced dimension than the original predictor data, ANN models computed using ten hidden layers show improvement on the results from the PLS models.

Changes in coefficient of determination and RMSE are not pronounced when different PLS predictor scores were used in the input layer of the ANN, for example; using predictor scores of one component PLS model computed with data pre-processed by normalization followed Savitzky-Golay first-derivative smoothing (table 9) as the input layer, and using predictor scores of two component PLS model computed with data pre-processed by normalization

followed by Savitzky-Golay second-derivative smoothing (table 10), resulted to approximately the same coefficient of determination, 0.96 and 0.96, and RMSEP, 0.27 and 0.25, for the testing set. The Best ANN model was obtained when four component PLS predictor scores from raw data was used as input with the lowest RMSEP of 0.22 and coefficient of determination of 0.97 for the test set.

As mentioned earlier during ANN model computation the number of hidden neurons affects model performance another consideration was the computation time and thus cost. Where using raw NIR data as input with 5 hidden layers of neuron takes from a few seconds to about 1 minute to compute the model, using the same input with 20 hidden layers of neuron takes from 3 minutes to just less than 5 minutes to compute the model. Therefore, a high number of hidden neurons may not only over fit the model but also increases the computation time and thus cost of computation. On the contrary computation of PLS models takes just a few seconds of computation time.

5.0 Conclusion

This research investigated the performance of artificial neural networks models over the multivariate partial least squares regression models for the prediction of active substance content in a pharmaceutical tablet from a NIR spectral data. ANN is a data driven empirical model that is not restricted by the linear assumption of the PLS between sample spectral response and the active substance content. The application of ANN on raw tablet data returned a better model performance results than the PLS models. Further improvement in ANN model performance was observed when predictor scores from a PLS model was used as the input in computing the ANN model. This shows the dimensionality reduction of the spectral data by PLS enhances the performance of an ANN model when the predictor scores are used as input.

Both PLS regression and ANN have been used to develop multivariate models for the prediction of active substance content, and the combination of the two methods by using the PLS predictor scores as the input for ANN, presents an interesting method with great potential to develop quantitative models based in NIR spectroscopy.

6.0 Recommendations

Although ANN performed better than PLS, concerns have been raised on interpretation of the models with difficulty to understand the results (Cozzolino, et al., 2011). Further research should be carried out with towards this issue. While ANN model computation involved random partitioning of the datasets into training, validation and test set, the developed models can be further validated using an independent data set to check the accuracy of prediction.

References

Afseth, N. K. & Kohler, A., 2012. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, Volume 117, pp. 92-99.

Alam, M. A., III, J. D. & Anderson, C., 2017. Designing a Calibration set in spectral space for efficient development of an NIR method for tablet analysis. *Journal of Pharmaceutical and BIomedical Analysis*, Volume 145, pp. 230-239.

Barclay, V. J., Bonner, R. F. & Hamilton, I. P., 1997. Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data Set Compression. *Analytical Chemistry*, 69(1), pp. 78-90.

Bi, Y. et al., 2016. A local pre-processing method for near infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica Chimica Acta*, Volume 909, pp. 30-40.

Blanco, M. & Peguero, A., 2010. Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Trends in Analytical Chemistry*, 29(10), pp. 1127-1136.

Blanco, M. & Villarroya, I., 2002. NIR spectroscopy: a rapid response analytical tool. *Trends in Analytical Chemistry*, 21(4), pp. 240-250.

Bleye, C. D. et al., 2012. Critical review of near-infrared spectroscopic methods validation in pharmaceutical applications. *Journal of Pharmaceautical and Biomedical Analysis*, Volume 69, pp. 125-132.

Brereton, R. G., 2000. Introduction to multivariate calibration in analytical chemistry. *Thr Royal Society of Chemistry*, Volume 125, pp. 2125-2154.

Brereton, R. G., 2018. *Chemometrics: Data Driven Extraction for Science*. 2nd ed. Chichester, West Sussex: John Wiley & Sons Ltd.

Chen, T., 2018. *Lecture 8 - Introduction to empirical models.* Guidford: University of Surrey.

Chen, T., Morris, J. & Martin, E., 2007. Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*, Volume 87, pp. 59-71.

Chen, Y. et al., 2001. Prediction of Drug Content and Hardness of Intact Tablets Using Artificial Neural Network and Near-Infrared Spectroscopy. *Drug Development and Industrial Pharmacy*, 27(7), pp. 623-632.

Ciurczak, E. W. & Igne, B., 2015. *Pharmaceutcal and Medical Applications of Near Infrared Spectroscopy.* 2nd ed. Boca Raton: CRC Press, Taylor & Francis Group.

Cozzolino, D., Cynkar, W., Shah, N. & Smith, P., 2011. Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International*, Volume 44, pp. 1888-1896.

Dyrby, M. et al., 2002. Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. *Applied Spectroscopy*, 56(5), pp. 579-585.

Engel, J. et al., 2013. Breaking with trends in pre-processing?. *Trends in Analytical Chemistry*, Volume 50, pp. 96-106.

Freitas, M. P. et al., 2005. Prediction of drug dissolution profiles from tablets using NIR diffuse reflectance spectroscopy: A rapid and nondestructive method. *Journal of Pharmaceutical and Biomedical Analysis*, Volume 39, pp. 17-21.

Gallagher, N. B., Blake, T. A. & Gassman, P. L., 2006. Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil. *Journal of Chemometrics*, Volume 19, pp. 271-281.

Hart, J. H. & Norris, K. H., 1996. Direct spectrophotometric determination of moisture content of grain and seeds. *Journal of Near Infared Spectroscopy*, Volume 4, p. 23.

Herberger, K., 2008. Chapter 7 - Chemoinformatics—multivariate mathematical–statistical methods for data evaluation. *Medical Applications of Mass Spectrometry*, pp. 141-169.

Herschel, F. W., 1800. Experiments on the Refrangibility of the Invisible Rays of the Sun. *Philosophical Transactions of the Royal Academy*, pp. 284-292 .

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2014. *An introduction to Statistical Learning with Applications in R.* 1st ed. New York: Springer.

Kennerd, R. W. & Stone, L. A., 1969. Computer Aided Design of Experiments. *Technometrics*, 11(1), pp. 137-148.

Kuhn, M. & Johnson, K., 2013. *Applied Predictive Modeling.* 1st ed. New York: Springer.

Massart, D. L. et al., 1998. *Handbook of Chemometrics and Qualimetrics: Part A.* Amsterdam: Elsevier.

McClure, W. F., 2003. 204 years of near infrared technology: 1800-2003. *Journal of Near Infrared Spectroscopy*, Volume 11, pp. 487-518.

Mehmoud, T., Liland, K. H., Snipen, L. & Saebo, S., 2012. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, Volume 118, pp. 62-69.

Naes, T. & Holland, I., 1993. Relevant components in regression. *Scandinavian Journal of Statistics*, Volume 20, pp. 2239-250.

Naes, T., Isaksson, T., Fearn, T. & Davies, T., 2002. *A User-Friendly Guide to Multivariate Calibration and Classification.* 1st ed. Chichester: NIR Publications.

Ni, W., Norgaard, L. & Morup, M., 2014. Non-linear calibration models for near infrared spectroscopy. *Analytica Chimica Acta*, Volume 813, pp. 1-14.

Norgaard, L., Lagerholm, M. & Westerhaus, M., 2013. *Artificial Neural Networks and Near Infrared Spectroscopy - A case study on protein content in whole wheat grain*, Hilleroed: FOSS.

Norris, K. H. & Williams, P. C., 1984. Optimization of Mathematical Treatments of Raw Near-Infrared Signal in the Measurement of Protein in Hard Red Spring Wheat. I. Influence of Particle Size. *Cereal Chemistry*, 61(2), pp. 158-165.

Pasquini, C., 2003. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *Journal of Brazilian Chemical Society*, 14(2), pp. 198-219.

Pasquini, C., 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. *Analytica Chimica Acta* , Volume 1026, pp. 8-36.

Pasti, L., 1997. *Derivative computation using the Savitzky-Golay algorithm*, Vlaardingen: ChemoAc.

Porep, J. U., Kammerer, D. R. & Carle, R., 2015. On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology*, Volume 46, pp. 211-230.

Reich, G., 2005. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Advanced Drug Delivery Reviews*, Volume 57, pp. 1109-1143.

Rinnan, A., Berg, F. v. d. & Engelsen, S., 2009. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, 28(10), pp. 1201-1222.

Rinnan, A. et al., 2009. Data Pre-processing. In: D. Sun, ed. *Infrared Spectroscopy for Food Quality Analysis and Control.* New York: Elsevier Inc, pp. 29-50.

Roggo, Y. et al., 2007. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, Volume 44, pp. 683-700.

Savitzky, A. & Golay, M. J. E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), pp. 1627-1638.

Stuth, J., Jama, A. & Tolleson, D., 2003. Direct and indirect means of predicting forage quality through near infrared reflectance spectroscopy. *Field Crops Research*, Volume 84, pp. 45-56.

Wang, J. C., 2001. A New Approach to Near Infrared Spectral Data Analysis Using Independent Component Analysis (ICA). *Journal of Chemical Information and Computer Sciences*, pp. 992-1001.

Wang, K., Chen, T. & Lau, R., 2011. Bagging for robust non-linear multivariate calibration spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, Volume 105, pp. 1-6.

Wiesner, K., Fuchs, K., Gigler, A. M. & Pastusiak, R., 2014. Trends in Near Infrared Spectroscopy and Multivariate Data Analysis from an Industrial Perspective. *Procedia Engineering*, Volume 87, pp. 867-870.

Williams, P., Dardenne, P. & Flinn, P., 2017. Tutorial: Items to be included in a report on a near infrared spectroscopy project. *Journal of Near Infrared Spectroscopy*, 25(2), pp. 85-90.

Winding, W., Shaver, J. & Bro, R., 2008. Loopy MSC: A Simple Way to Improve Multiplicative Scatter Correction. *Applied Spectroscopy*, 62(10), pp. 1153-1159.

Wold, S., Antti, H., Lindgren, F. & Ohman, J., 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, Volume 44, pp. 175-185.

Yuanyuan, C. & Zhibin, W., 2018. Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks. *Chemometrics and Intelligent Laboratory Systems*, Volume 181, pp. 1-10.