NIR-Spectroscopic Classification of Blood Glucose Level using Machine Learning Approach

Mohammad Habibullah
Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, Canada
moh220@mail.usask.ca

Mohammad Abdul Moin Oninda
Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, Canada
mohammad.abd.moin@usask.ca

Ali Newaz Bahar
Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, Canada
ali.bahar@usask.ca

Anh Dinh
Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon,
Canadaanv252@mail.usask.ca

Khan A. Wahid
Department of Electrical and Computer Engineering
University of SaskatchewanSaskatoon, Canada
khan.wahid@usask.ca

***Abstract*—Diabetes Mellitus (DM) or diabetes is one of the metabolic diseases exhibiting high blood glucose level over a prolonged period. The management of diabetes is associated with the proper monitoring of blood glucose level. Researchers have been working on developing robust techniques to monitor the level of glucose in the blood. This paper aims at predicting blood glucose levels based on NIR spectroscopic response data utilizing machine learning techniques. Blood glucose samples were prepared in a controlled environment and the NIR spectrums of the samples were obtained using NeoSpectraMicro development kit. Two machine learning approaches have been employed to analyze the experimental dataset. Firstly, the Random Forest Algorithm (RF) followed by Support Vector Machine (SVM) has been utilized that provides an accuracy of 67.5%. Then, a combination of Principle Component Analysis (PCA) and SVM is used. PCA followed by SVM shows a promising result of 77.5% accuracy compared to the previous technique. The numerical findings reveal that the NIR spectroscopy with appropriate data modeling algorithm can be a potential candidate for non-invasive blood glucose monitoring system.**

Keywords— Diabetes Mellitus (DM), NIR Spectroscopy, NeoSpectraMicro, SVM, PCA, Blood Glucose, Random Forest (RF)

I. INTRODUCTION

Diabetes can be referred to as one of the primary healthcare challenges of the modern epoch. It is expected that the total number of people mandated by diabetes will rise from 171 million in 2000 to over 366 million by 2030 securing the 6th dominant reason of death in the USA [2]. The inability of the human body to make sufficient insulin makes diabetes an incurable disease [3] causing high blood glucose level (i.e. hyperglycemia) or low blood glucose level (i.e. hypoglycemia) [4,5]. The pancreas produces insulin that initiates glucose uptake by the cells thereby reducing glucose concentration in blood [6,7]. Diabetes can be categorized into three groups, Type I diabetes mellitus, Type II diabetes mellitus, and gestational diabetes. An exciting frontier in diabetes research is the monitoring of the blood glucose level. The conventional testing and monitoring system is painful and introduces discomfort in patients daily life. Researchers have been working on developing techniques and instruments to closely monitor blood glucose level so that diabetes can be easily monitored and controlled thereby reducing the pain, discomfort, and cost that the patients with diabetes experiences.

Continuous development in blood glucose measuring equipment is being done seeking the qualities of being non-invasive, non-contact, painless, convenience for continuous real-time monitoring, and cost-effective having fast measurement capability and reduction in complication from usage. Continuous glucose monitoring (CGM) delivers elaborate information which can help to prevent unwanted hypo and hyperglycemic episodes [8]. Besides blood, other suitable body fluids for glucose measurement are intestinal fluid, ocular fluid, sweat, saliva, breath or urine [3].

Over the past few decades, a number of sensor-based blood glucose monitoring techniques have been developed and studied [3,5,8]. In [9] researchers have studied the enzymatic and non-enzymatic electrochemical glucose sensing approaches. The use of infrared spectroscopy in the monitoring of blood glucose were reported in [9-11] with an average level of accuracy. However, detecting blood glucose level using spectroscopy is gaining popularity.

This paper focuses on the classification of blood glucose level based on the NIR spectroscopy by employing the NeoSpectraMicro development kit. Data are analyzed using a suitable machine learning approach to be able to correctly predict the glucose level in blood. Section II of this paper provides a brief scientific background of the correlation of NIR light with blood glucose. Detail about the experimental setup is to be discussed in Section III. Section IV presents the results and simulation to validate the proposed concept. Finally, this paper is concluded in Section V.

II. SCIENTIFIC BACKGROUND

Near-infrared spectroscopy (NIRS) stands for the near-infrared region of the electromagnetic spectrum band from 780

978-1-7281-0319-8/19/$31.00 ©2019 IEEE

nm to 2500 nm [12]. The characteristic of NIRS to penetrate deeper into the sample makes it very useful, compared with mid-infrared spectra. When NIR light is incident on the sample, it interacts with the chemical components of the sample causing the light to be partially absorbed and scattered. The equation of light attenuations [13] can be described as follow:

$$I = I_o e^{-\mu_{eff} d} \tag{1}$$

Here,

I - intensity of reflected light

I_o - intensity of incident light

μ_{eff} - effective attenuation coefficient

d - length of the optical path

On the other hand, μ_{eff} can be expressed as a function,

$$\mu_{eff} = f(\mu_a, \mu_s) \tag{2}$$

where

μ_a- absorption coefficient

μ_s- scattering coefficient.

μ_a of a sample is highly influenced by the changes in glucose concentration corresponding to the changes in its intrinsic absorption. Changes in glucose concentration also affect the intensity of light scattered, i.e., μ_s. So, it is evident that variations of light intensity in NIR region can be effectively utilized to estimate glucose concentration by transmission through a glucose-containing the sample and reflected by it.

Fig. 1. NeoSpectraMicroTM Development Kit

This experiment utilizes Near-Infrared Spectroscopy (NIRS) from 1300nm–2600nm wavelength for characterization of glucose concentration in blood using the NeoSpectraMicro™ Development Kit [14] as the sensing device. The NeoSpectra Micro, shown in Figure 1, is a chip-sized NIR spectral sensor that delivers the spectral feedback of absorption of light by the materials for characterization and identification.

As the aim of the project is to approximate the blood glucose level from NIR-spectrum data, so finding the best model that suits the data set with the highest accuracy is the major part of the project. Several multivariate analyses-Random Forest (RF) algorithm, Principal Component Analysis (PCA) and Support Vector Machine (SVM) were performed and compared to build the model. All the data processing, modeling, and results were performed in PyCharm (version 2018.2) integrated development environment using Python 3.7 programming language.

III. Experimental Setup

A. Solution Preparation

For collecting data, ten artificial blood samples have been prepared based on standard human blood glucose levels (Table I). The amount of solute (glucose) can be obtained using equation (3).

$$Solute\ (g) = Concentration\left(\frac{mol}{L}\right) \times RFM \times Volume(L) \tag{3}$$

where, Relative Formula Mass (RFM) of glucose is 180.156g.

TABLE I. Standard Human Blood Glucose

Level	mg/dL	mmol/L	Risk
Dangerously High	315+	17.4	Very High
High	280	15.6	High
High	250	13.7	High
High	215	11.0	High
Borderline	180	10.0	Medium
Borderline	150	8.2	Medium
Borderline	120	7	Medium
Normal	108	6	No risk
Normal	72	4	No risk
Low	70	3.9	Medium
Dangerously Low	50	2.8	High

In this project, the volume was considered as 250 ml, and the mass of solute has been calculated. Table II shows the mass of solute in gram for different concentration of blood glucose.

TABLE II. Mass of Solute of Different Blood Sample

Sample	Concentration (mol/L)	RFM (g)	Volume (L)	Mass of solute (g)
C1	0.0040	180.156	0.250	0.180156
C2	0.0060	180.156	0.250	0.270230
C3	0.0070	180.156	0.250	0.315270
C4	0.0082	180.156	0.250	0.369300
C5	0.0100	180.156	0.250	0.450390
C6	0.0110	180.156	0.250	0.495400
C7	0.0137	180.156	0.250	0.617030
C8	0.0156	180.156	0.250	0.702600
C9	0.0174	180.156	0.250	0.783600
C10	0.0200	180.156	0.250	0.900780

B. Data Collection

NeoSpectra Micro development kit was used to obtain spectrum in Near-Infrared region (1300-2500 nm), and the data were visualized using the SpectroMost GUI shown in Fig. 2. A total of 100 samples data were taken from 10 different blood solutions (C1-C10) in accordance to Table II. The kit was

calibrated using a white and black background before the data collection step.

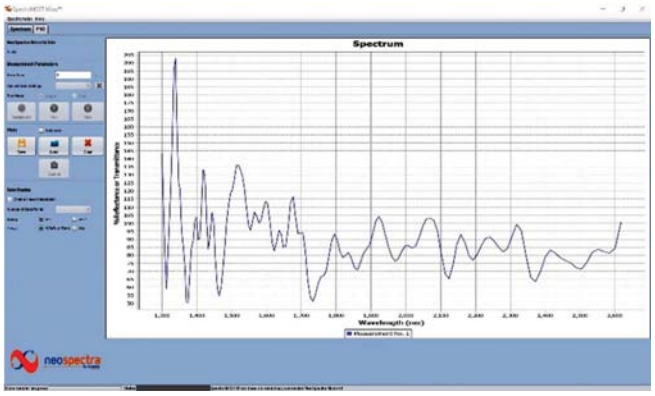

Fig. 2. Software GUI for NeoSpectra Modules (SpectroMost)

Different levels of blood glucose solutions exhibit distinctive light reflective characteristic while observed from 1300 nm-2300 nm. Ten samples from each class were averaged and plotted against the wavelengths for analysis. Fig. 3 shows the plot of variations of reflectance of ten classes.

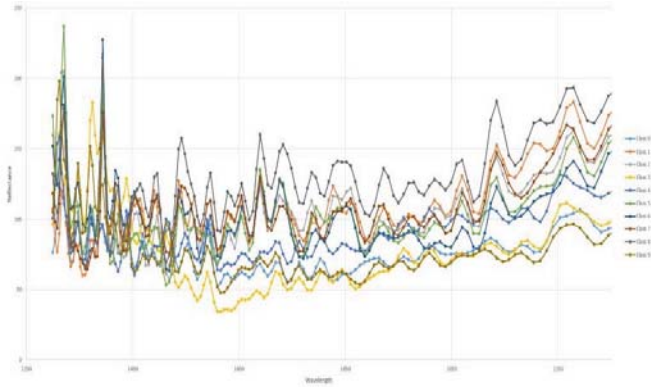

Fig. 3. A plot of reflectance vs wavelength of blood glucose solution

IV. Classification Techniques

The dataset contains reflectance of 100 samples at 156 different wavelengths, considered as features. The different predefined glucose levels (Table II) were labeled as classes. At first, the feature vector was scaled through standardization having properties of a standard normal distribution with a mean of zero and a standard deviation of one. The dataset was then modeled for classification in two techniques.

In the first techniques, RF was used for extracting important features. Here, the 1% threshold has been used that reduced the features from 156 to 28. After reducing the features, dataset were randomly shuffled and then fed into SVM classifier.

In the second method, PCA was applied at first for reducing the dimensionality. While choosing the number of principal components, 95% variance has used that means, it chose the minimum number of principal components such that 95% of the variance was retained. In this case, 95% of the variance resulted to 35 principal components. Now, SVM classifier was applied to those 35 components. For cross-validation, the holdout method was used in both the techniques splitting the dataset to 60% for training and 40% for testing.

V. Result Analysis

For evaluating the performances of the classification techniques, both the trained models were tested on the test set consisting of 40 samples. We followed standard evaluation metrics- accuracy, precision, recall, f1-score for classification performance analysis. All of them were calculated from the confusion matrix as follows:

$$Accuracy = \frac{Total\ number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}}$$

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}}$$

$$F1 - score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

Where i and j are the indices of the confusion matrix. Actual classes are depicted by the rows whereas the columns refer to the predicted classes.

The first techniques, RF followed by SVM, showed an average accuracy of 67.5% and precision of 70%. Detailed class wise performance is shown in Table III and plotted in Fig. 4.

TABLE III. Classification Report of Technique 1

Class	Precision	Recall	F1-score
1	1.00	0.6	0.75
2	0.60	1.00	0.75
3	0.25	0.25	0.25
4	0.50	1.00	0.67
5	1.00	1.00	1.00
6	0.40	0.40	0.40
7	0.50	0.25	0.33
8	0.67	0.67	0.67
9	1.00	1.00	1.00
10	1.00	1.00	1.00
Weighted Average	0.70	0.68	0.67

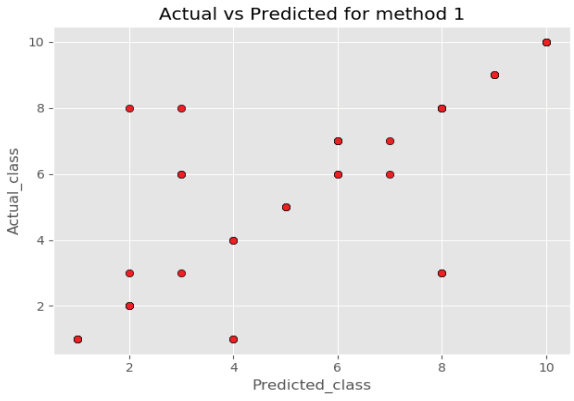

Fig. 4. A plot of actual class vs predicted class for the first technique

The second approach showed a better performance than first. This technique, PCA followed by SVM, predicted the classes with an average accuracy of 77.5% and precision of 82%. Table IV shows the classification report of the second technique.

TABLE IV. CLASSIFICATION REPORT OF TECHNIQUE 2

Class	Precision	Recall	F1-score
1	1.00	1.00	1.00
2	0.60	1.00	0.75
3	0.50	1.00	0.67
4	1.00	1.00	1.00
5	1.00	1.00	1.00
6	0.50	0.40	0.44
7	0.67	0.50	0.57
8	1.00	0.33	0.50
9	1.00	1.00	1.00
10	1.00	1.00	1.00
Weighted	0.82	0.78	0.76

The performance improvement from method one to method two can be visualized by comparing Figure 4 and Figure 5. Most of the outliers in the first technique were disappeared in the second techniques. For better comparison, the confusion matrices of both the method are shown in Table V.

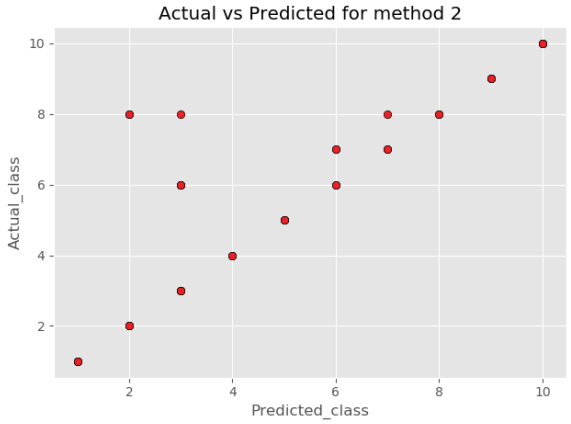

Fig. 5. A plot of actual class vs predicted class for the second technique

TABLE V. CONFUSION MATRICS

Technique 1	Technique 2
[**3** 0 0 2 0 0 0 0 0 0]	[**5** 0 0 0 0 0 0 0 0 0]
[0 **3** 0 0 0 0 0 0 0 0]	[0 **3** 0 0 0 0 0 0 0 0]
[0 1 **1** 0 0 0 0 2 0 0]	[0 0 **4** 0 0 0 0 0 0 0]
[0 0 0 **2** 0 0 0 0 0 0]	[0 0 0 **2** 0 0 0 0 0 0]
[0 0 0 0 **3** 0 0 0 0 0]	[0 0 0 0 **3** 0 0 0 0 0]
[0 0 2 0 0 **2** 1 0 0 0]	[0 0 3 0 0 **2** 0 0 0 0]
[0 0 0 0 0 3 **1** 0 0 0]	[0 0 0 0 0 2 **2** 0 0 0]
[0 1 1 0 0 0 0 **4** 0 0]	[0 2 1 0 0 0 1 **2** 0 0]
[0 0 0 0 0 0 0 0 **3** 0]	[0 0 0 0 0 0 0 0 **3** 0]
[0 0 0 0 0 0 0 0 0 **5**]	[0 0 0 0 0 0 0 0 0 **5**]

VI. CONCLUSION

In this article, a new approach is presented for monitoring blood glucose levels based on the Near-Infrared Spectroscopy. To evaluate and validate our system, Random Forest algorithm, Principal Component Analysis and Support Vector Machine algorithms have been applied. It was observed that Principal Component Analysis combined with the Support Vector Machine has the highest capability to predict glucose level in the sample data. All the experiments were performed on artificial blood sample with different concentrated glucose solutions. This method has not been applied to real blood yet. Moreover, our experimentally collected data does not fully mimic actual varieties of blood glucose levels. Nevertheless, this study reports that Near-Infrared Spectroscopy can be modeled with appropriate machine learning techniques for noninvasive blood glucose monitoring.

REFERENCES

[1] Steffes, M. W., & Sacks, D. B. (2005). Measurement of circulating glucose concentrations: the time is now for consistency among methods and types of samples.*Clinical Chemistry*, 51 (9), 1569-1570.

[2] National Diabetes Information Clearinghouse. (2011). National diabetes statistics.

[3] American Diabetes Association. (2004). Accuracy of the GlucoWatch G2 Biographer and the continuous glucose monitoring system during hypoglycemia: experience of the Diabetes Research in Children Network. *Diabetes Care*, 27(3), 722-726.

[4] Sabokdast, M., et al., (2015). Protection by beta-Hydroxybutyric acid against insulin glycation, lipid peroxidation and microglial cell apoptosis. *DARU Journal of Pharmaceutical Sciences*, 23(1), 42.

[5] Coster, S., Gulliford, M. C., Seed, P. T., Powrie, J. K., & Swaminathan, R. (2000). Monitoring blood glucose control in diabetes mellitus: a systematic review. *British Journal Of Clinical Governance-Bradford*, 5(4), 225-227.

[6] American Diabetes Association. (2014). Diagnosis and classification of diabetes mellitus. *Diabetes care*, 37(Supplement 1), S81-S90.

[7] Bratlie, K. M., York, R. L., Invernale, M. A., Langer, R., & Anderson, D. G. (2012). Materials for diabetes therapeutics. *Advanced healthcare materials*, 1(3), 267-284.

[8] Amir, O., et al., (2007). Continuous noninvasive glucose monitoring technology based on "occlusion spectroscopy".*Journal of Diabetes Science and Technology*, 1(4), 463-469.

[9] Yadav, J., Rani, A., Singh, V., & Murari, B. M. (2015). Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy. *Biomedical signal processing and control*, 18, 214-227.

[10] Pandey, R., et al., (2017). Noninvasive monitoring of blood glucose with raman spectroscopy. *Accounts of chemical research*, 50(2), 264-272.

[11] Spegazzini, N., et al., (2014). Spectroscopic approach for dynamic bioanalyte tracking with minimal concentration information. *Scientific reports*, 4, 7013.

[12] Malin, S. F., Ruchti, T. L., Blank, T. B., Thennadil, S. N., & Monfre, S. L. (1999). Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy. *Clinical chemistry*, 45(9), 1651-1658.

[13] Khalil, O. S. (2004). Non-invasive glucose measurement technologies: an update from 1999 to the dawn of the new millennium. *Diabetes technology &therapeutics,* 6(5), 660-697.

[14] https://www.neospectra.com/wp-content/uploads/2018/01/Neospectra-SWS62231-Datasheet.-11-12-17.pdf