Predictive Modeling for Early Hyperglycemia Detection in Type 2 Diabetes

Pulkita Jain [*,1], Dr. Mudassir Rashid [*,2]

[*]Department of Chemical and Biological Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA

I.ABSTRACT

Type 2 Diabetes (T2D) is characterized by a combination of defects in insulin action and impairment in insulin secretion. Deficient insulin action causes people with Type 2 Diabetes to have difficulty controlling their blood glucose concentration (BGC) and experience periods of high (hyperglycemia) and low (hypoglycemia) BGC. Continuous glucose monitoring (CGM) sensors and machine learning algorithms can automate the process of meal size estimation, improve the accuracy of the carbohydrate estimations, and reduce the involvement of the subject. The aim of this project was to use dynamic Partial Least Squares (PLS) regression to model blood glucose data from CGM devices and optimize the model parameters for best generalized performance across all subjects. The mean square error for the modelled data was obtained to determine the accuracy of the predictive modelling. The parameters were optimized by explicit enumeration and grid search approach to minimize the mean square error (MSE), and the lowest MSE was obtained with the number of latent variables as 5 and past horizon as 5 (25 minutes). Future research will develop the logic inference using the first- and second-order derivatives of the prediction curve that will sound the alarms based on the predictions made in the current work.

II.INTRODUCTION

2.1 Diabetes

Type 2 diabetes (T2D) is a heterogeneous disease with a significant degree of interpersonal variability that affects an estimated 34 million Americans [4]. T2D is characterized by an increase in resistance to insulin, a decrease in insulin production and secretion, or some combination of these factors. This causes individuals with T2D to have difficulty controlling their blood glucose concentration (BGC) and experience periods of high (hyperglycemia) and low (hypoglycemia) BGC[4]. Ingesting food leads to an increase in blood glucose levels which the body processes signaling the pancreas to release insulin which simulates the adipose and muscle cells to absorb the glucose from the bloodstream. Essentially this is how the body regulates the glucose levels. There are other hormones like glucagon involved too, however, this paper will focus on the levels of insulin as the main factor. In T2D, as the body produces less insulin, less glucose is absorbed from the bloodstream leading to an overall increase in glucose levels which is termed as hyperglycemia [8]. Hyperglycemia can also occur when the body starts to resist the insulin produced which leads to high glucose levels. T2D is a chronic condition with no cure which demonstrates the responsibility of the patient to manage their own blood glucose levels. This can be done using continuous glucose monitoring sensors (CGM). Prolonged hyperglycemia can lead to chronic and severe health conditions over time such as heart and blood diseases, kidney failures, nerve damage in limbs and so on [9].

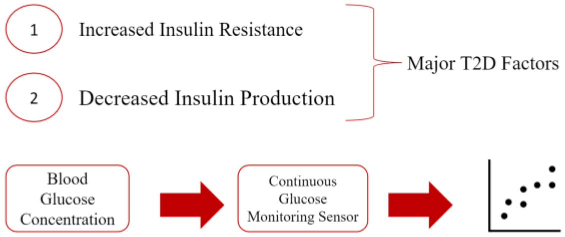
1
Increased Insulin Resistance
2
Decreased Insulin Production
Major T2D Factors
Blood Glucose Concentration
Continuous Glucose Monitoring Sensor

Figure 1: flowchart shows that the CGM sensor detects the blood glucose concentration and produces data points.

2.2 CGM Sensors (what are they?)

Blood Glucose monitoring has been revolutionized by the development of Continuous Glucose Monitoring (CGM) sensors [2]. These are wearable minimally-invasive devices that measure glucose concentration almost continuously (1–5 min sampling period) for several consecutive days/weeks. The first wearable CGM sensor prototype was introduced in 1999 and, since then, devices have evolved rapidly [2]. CGM readings may be sent to a phone with a sampling time of 5 minutes and accessed through phone applications. While more popular for individuals with T1D, CGM sensors provide valuable, real-time insight into T2D blood glucose dynamics. CGM sensors have greatly improved and are now able to incorporate filtering to provide a more accurate measurement [10]. These devices now incorporate several features such as the ability to make decisions regarding the amount of food eaten to balance hypoglycemia or the physical exercise done to balance hyperglycemia. CGM sensors can monitor the blood glucose data in the real time and general alerts for hypo/hyperglycemia, however they do not have the ability to make or analyze the future predictions for BGC.

2.3 Predictive Modeling in Diabetes (several citations)

Early hyperglycemia warning systems based on continuous glucose monitoring (CGM) sensors may provide a convenient solution for monitoring and reducing the severity of hyperglycemia episodes. Hyperglycemia prediction is an estimation of when a person's blood glucose concentration (BGC) will rise above a certain threshold soon. Typically, the threshold for hyperglycemia is considered 180 mg/dL and this value was initially used in this study. However, many people with Type 2 diabetes (T2D) may have a higher fasting BGC which would cause small amounts of carbohydrate consumption to push the person into hyperglycemia. Continuous glucose monitoring sensors and machine learning algorithms can automate the process of meal size estimation, improve the accuracy of the carbohydrate estimations, and reduce the involvement of the patient.

In figure 2, real time glucose data obtained from CGM sensors was graphed along with the threshold limit of hyperglycemia. The patients experience varied highs and lows during the 1500 minute sampling time. This depicts the need of predictive modelling in diabetes. Predictive modelling will be able to make accurate predictions about someone’s glucose level which can then be utilized by the patient to control their glucose levels more effectively before they ever reach the threshold.

System identification refers to the process of finding a mathematical model that describes a set of input-output data [7]. In this paper, the input data are past and current BGC measurements and the output data are the future BGC measurements. Identifying a model from this data therefore produces a predictive model that can be used to forecast future blood glucose values. The mathematical model of PLS Regression was utilized for this purpose [6]. Other mathematical models such as autoregressive-moving-average (ARMA) or Logistic Multiple Regression can also be used however, PLS was found to be the most effective and hence chosen as the predictive model.

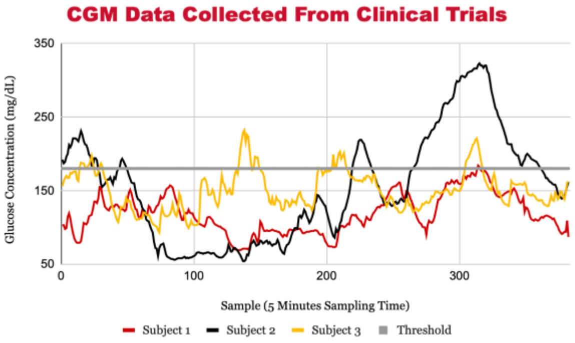

Figure 2: Real time data from 3 patients and the threshold limit of hyperglycemia at 180 mg/dL

III.METHODS

3.1 Data Used

The data set for this project was the real-time data obtained from 135 patients. It was cleaned and filtered to reduce noise and rectify missing measurements. The data was then modelled using PLS Regression technique. The MATLAB Statistics and Machine Learning toolbox software was used for this research project.

3.2 PLS Regression

3.2.1 What is PLS?

Partial least squares (PLS) regression allows for the coefficients for all outputs to be estimated simultaneously by projecting the input regressor data, X, and the output response or predictions, Y, onto orthogonal subspaces of A-pairs of latent variables[1]. Each pair of latent variables accounts for a certain percentage of the variance in the input regressor data and output response matrices. Mathematically, PLS regression consists of decomposing X and Y as the sum of the outer products of a score and loading vector as follows:

$$X = \sum_{a=1}^{A} t_a p_a{}^T + E = TP^T + E$$

$$Y = \sum_{a=1}^{A} u_a q_a{}^T + F = UQ^T + F$$

where t_a and u_a are the input and output scores representing the projections of the variables in X and Y on their subspaces, t_a and u_a define the orientation of the corresponding subspaces, the matrices,T, P, U, and Q contain their corresponding vectors, and E and F denotes residual matrices for the input regressor and output response data matrices. The noise reduction property of PLS regression stems from the idea that the fewer latent variables are typically a consequence of measurement noise and system irregularities and therefore can be discarded during the PLS regression algorithm.

Since the PLS algorithm is used to obtain a mathematical relationship between the original data matrices, the input regressor and output response matrices are related through an inner relation between the corresponding scores as follows:

$$u_a = b_a t_a + e_a \, , for \ a \in [1, A]$$

where b_a are the coefficients and e_a are the residuals of the inner relationship between the scores of X and Y [5]. In this work, the PLS regression parameters are computed using the nonlinear iterative partial least squares (NIPALS) algorithm where the subspace orientation and scores for both the regressor input and output response matrices are determined simultaneously to maximize the covariance between X and Y and obtain the optimal fit for the inner relationship.

In this work, dynamic PLS is used to model the time-varying correlations and lags between the past CGM data and the future CGM data to be predicted. Dynamic PLS is a powerful multivariate algorithm that builds efficient models for predicting the future values by maximizing the covariance between the past and future data.

PLS relates regressor and regressed variables by maximizing the covariances between them [3]. PLS builds linear relations between input data and output data and uses these relations to predict future values. PLS has latent variables that describe the important underlying features of the data.

3.2.2 Glucose Modelling Methodology

A hyperglycemia prediction algorithm based on PLS regression and qualitative trend analysis has been developed. A matrix of past CGM data is used to handle the CGM time series, and the data are split into training and testing data. The training set is used to identify the model parameters and algorithm hyperparameters before evaluating the prediction algorithm on the independent testing set. Usually, 80% of the dataset is designated for training and 20% is designated for testing so this allocation was used. The training set was normalized by calculating the Zscores for the data and then fitting the PLS model to it. Once the model was fit, mean square error was calculated to determine the prediction accuracy. Next, the testing data set was used with the now fit PLS model and the mean square error was calculated. The three important variables used in this project were that of past horizon, latent variables and future horizons. Latent variables are crucial to the PLS model and help find better relations to perform a better regression. Past horizons are the set of previous data points that are needed by the program to determine accurate predictions values. Future horizon is defined as how far ahead into the future are the predictions being made. A sample time of 5 minutes was set so the past horizon of 1 implied 5 minutes in real time. The program was developed to run with a range of past horizon and latent variables values so these two important parameters could be optimized. The PLS model was then run for 135 patients using the optimized parameters and for different future horizons and the results were tabulated.

IV.RESULTS

The following results were obtained when the PLS model was fit to the training and the testing data. The mean square error obtained for the training set and the testing set is in the table below. As it can be seen, the model (red) in the graph fits the actual glucose data really well. It shows that PLS is a great mathematical model to use for this type of predictive modelling. The testing data also worked really well with a very low mean square error for 15 minute ahead predictions.

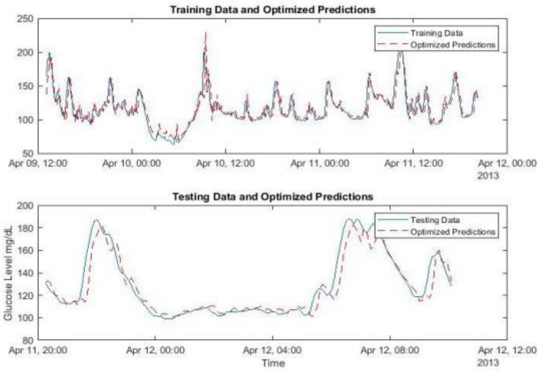

Figure 4: The top graph shows the training data in blue with its predictions in red. The bottom graph shows the testing data in blue with its predictions in red.

Table 1: The parameters for figure 3 and figure 4

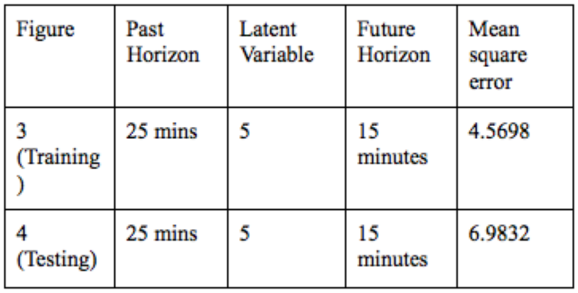

Figure	Past Horizon	Latent Variable	Future Horizon	Mean square error
3 (Training)	25 mins	5	15 minutes	4.5698
4 (Testing)	25 mins	5	15 minutes	6.9832

The next logical step was to verify the accuracy of the model for different future horizons. Three different future horizons of 15 minutes, 30 minutes and 45 minutes were picked for this case and the results were plotted for such. The figure below shows these graphs, and it can be seen from the figure that the mathematical model works really well for 15 minute ahead predictions. The predictions are fairly accurate for 30 minute ahead predictions, however the mean square error increases when moved on to the 45 minute ahead prediction horizon. This shows that the further into the future, the model tries to predict, the less accurate it becomes.

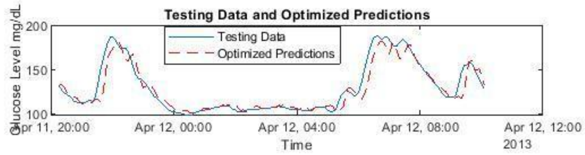

Figure 5: The graph shows optimized predictions for 15 minutes ahead into the future.

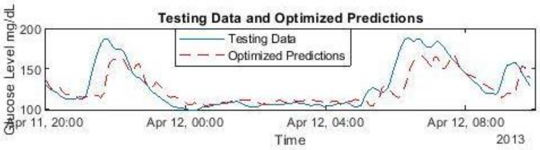

Figure 6: The graph shows optimized predictions for 30 minutes ahead into the future.

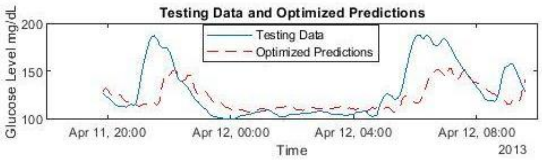

Figure 7: The graph shows optimized predictions for 45 minutes ahead into the future.

Table 2: Parameters for figure 5,6,7 and the associated mean square errors

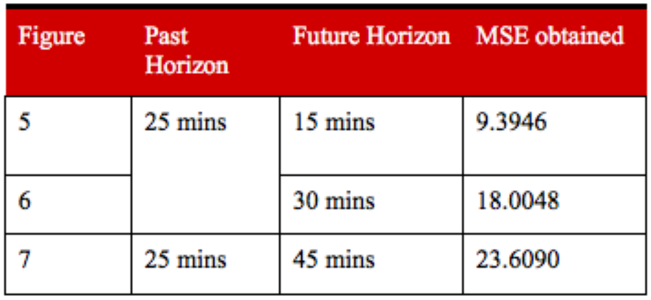

Figure	Past Horizon	Future Horizon	MSE obtained
5	25 mins	15 mins	9.3946
6		30 mins	18.0048
7	25 mins	45 mins	23.6090

Once the training and the testing data were modelled accurately and the results were satisfying, the next step was to optimize the user defined parameters of latent variable and past horizon. It is necessary to obtain these parameters since there is no set number available for this in the literature and it is crucial that the number of past data points being used is the same for each subject. Moreover, the lowest MSE obtained due to these optimized parameters makes the PLS model fit the data better. These parameters were optimized by explicit enumeration and grid search approach to minimize the mean square error (MSE).

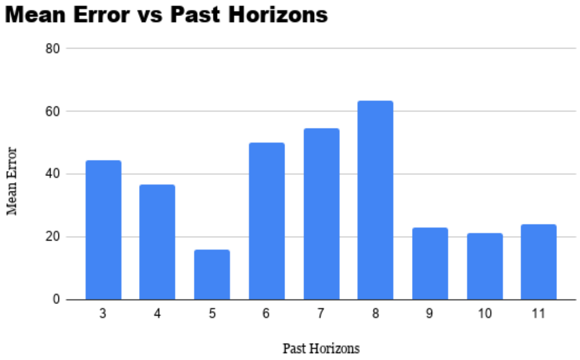

Figure 8: Mean Population Prediction Error vs Past Horizons.

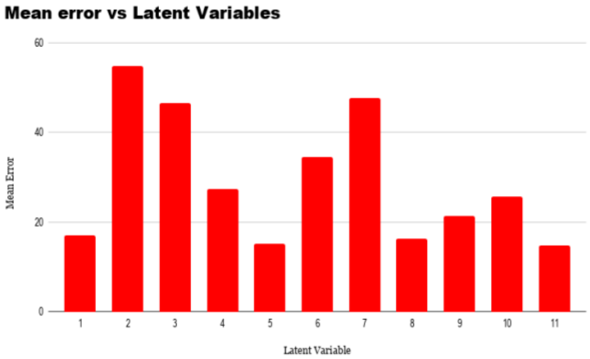

Figure 9: Mean Population Prediction Error vs Number of PLS Latent Variables.

The number of variables used was selected based on the number of prior data points used and the lowest error value. From the figures above, it can be seen that latent variables of 5 and past horizon of 5 (25 minutes ahead) give the lowest MSE for 135 subjects. Hence, these optimized parameters were selected and every figure in this paper has been created using these optimized parameters.

V.DISCUSSION

These findings suggest that the PLS model fits the real-time data well and demonstrates the associated errors with optimized parameters of past horizon and number of latent variables that will be used in the future research to develop this model further. Being able to make accurate predictions of blood glucose concentration enables the patient to further be able to manage their blood glucose levels well. The major significance of these findings can be utilized for future research work in developing an alarm system where the sensor can accurately predict the glucose levels and based on that sound an alarm if the patient will cross the hyperglycemia threshold limit in the future. By predicting accurately further into the future, proactive early warnings will be provided to the user on impending hyperglycemia events. Using the optimized parameters was successful since it helped eliminate large sources of error.

The data was obtained from 135 T2D patients, however, the demographics were not factored in, into this study. A good idea would be to include different types of parameters such as age, sex and other demographic information about the pateint to better see where the model can perhaps breakdown.

Future research will focus on developing the logic inference that will sound the alarms based on the predictions made in the current work. The logic inference will be developed using the first- and the second-order derivatives of the prediction curve. Qualitative trend analysis will also be utilized to help with the development of this alarm system. *Qualitative trend analysis* is used to extract information from time series data based on the overall behavior of the data. A polynomial is fit to a section of the data to provide a differentiable function. The first and second derivatives of this function are then found and evaluated. The sign of the derivatives after having been evaluated at a specific point then provides information about the data. The first derivative (dx) describes the direction and magnitude of the change in the data (i.e. the velocity) and can quantify whether it is increasing or decreasing in value. The second derivative (ddx) describes the rate of change of the velocity (i.e. the acceleration) and whether the data is heading towards a maximum or minimum, and change is decelerating or accelerating in either direction. Qualitative trend analysis is useful for hyperglycemia detection because the overall trend of blood glucose concentration is relatively well known. Postprandial hyperglycemia follows a roughly parabolic arc based on the carbohydrate content of the meal and physical characteristics of the person. The trends outlined by qualitative trend analysis can then be used to provide insight above the future trajectory of the BGC in real time. Using this, CGM sensors and machine learning algorithms will be able to automate the process of meal size estimation, improve the accuracy of the carbohydrate estimations, and help regulate hyperglycemia in people with T2D.

VI.CONCLUSION

Patients dealing with T2D regularly measure their blood glucose concentrations to live a healthy life. They do so by using the Continuous glucose monitoring sensor which records the BGC every 5 minutes and gives an estimate to the patient via digital applications. Current CGM sensors sound alarms when the hyperglycemia level has been crossed, however they cannot make predictions about such hyperglycemia events. This research paper takes the data obtained from CGM sensors and fits a mathematical predictive model to it. Thus, the mathematical model of Partial Least Squares Regression was used to model the continuous blood glucose concentration data obtained from the CGM sensors. After fitting the model, the prediction values of the BGC were calculated. The mean square error for all the future predictions was also calculated as well as the optimal parameters of past horizon and the number of latent variables were determined. Using the optimized parameters, predictions of BGC were also calculated for different future horizons. The error associated with these predictions also gave an estimate on how well the model works for different past and future horizons. This research will certainly enable patients suffering from T2D to better manage their BGC since they will have longer periods of time of knowing whether or not they will suffer a hyperglycemia event.

VII. ACKNOWLEDGEMENTS

Financial support for Pulkita Jain from the Program for Undergraduate Research Education (PURE) initiative of the Armour College of Engineering is gratefully acknowledged.

Contributions by Dr. Rashid and PhD candidate Andrew Shahidehpour are also gratefully acknowledged.

VIII. REFERENCES

1. Abdi, Hervé. "Partial least square regression (PLS regression)." *Encyclopedia for research methods for the social sciences* 6.4 (2003): 792-795.

2. Cappon, G.; Acciaroli, G.; Vettoretti, M.; Facchinetti, A.; Sparacino, G. Wearable Continuous Glucose Monitoring Sensors: A Revolution in Diabetes Treatment. Electronics 2017, 6, 65. https://doi.org/10.3390/electronics6030065

3. Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.

4. Defronzo - 2009 - From the triumvirate to the ominous octet A new paradigm for the treatment of type 2 diabetes mellitus

5. Geladi, Kowalski - 1986 - PARTIAL LEAST-SQUARES REGRESSION A TUTORIAL

6. Kaur, H. and Kumari, V. (2020), "Predictive modelling and analytics for diabetes using a machine

learning approach", *Applied Computing and Informatics*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1016/j.aci.2018.12.004

7. Muñoz, J. and Felicísimo, Á.M. (2004), Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science, 15: 285-292. https://doi.org/10.1111/j.1654-1103.2004.tb02263.x

8. Mellitus, Diabetes. "Diagnosis and classification of diabetes mellitus." *Diabetes care* 28.S37 (2005): S5-S10.

9. Paul Z Zimmet, Dianna J Magliano, William H Herman, Jonathan E Shaw, Diabetes: a 21st century challenge, The Lancet Diabetes & Endocrinology, Volume 2, Issue 1, 2014, Pages 56-64, ISSN 2213-8587, https://doi.org/10.1016/S2213-8587(13)70112-8. (https://www.sciencedirect.com/science/article/pii/S2213858713701128)

10. Sparacino, G.; Facchinetti, A.; Cobelli, C. "Smart" Continuous Glucose Monitoring Sensors: On-Line Signal Processing Issues. *Sensors* **2010**, *10*, 6751-6772. https://doi.org/10.3390/s100706751