A Novel Machine Learning Approach for Long-Term HbA1c Prediction

Almiqdad Elzein [1] Marwa Qaraqe [1] Goran Petrovski [2]

[1]College of Science and Engineering at Hamad Bin Khalifa University [2]Sidra Medicine

Abstract

Glycated hemoglobin (HbA1c) is an important indicator for Diabetes management. Hemoglobin is a protein in red blood cells that bonds with glucose molecules, of which Higher levels indicate higher blood glucose levels in the prior 90-120 days. Studies have shown an association between elevated HbA1c and diabetes-related comorbidities. Predicting future levels of HbA1c would enable physicians to make changes to treatment plans to avoid elevated HbA1c levels and irreversible health complica- tions. To this end, a novel Machine Learning method is developed in order to predict HbA1c levels 76 days in advance given a patient's blood glucose data from the previous 14 days. The blood glucose data used is collected by the FreeStyle Libre sensor, which takes a measurement of the blood glucose every 15 minutes. HbA1c levels are divided into four classes, each corresponding to an HbA1c range that reflects the quality of diabetes management. The proposed method converts a patients' blood glucose time-series into greyscale images and uses a series of binary Support Vector Machine (SVM) models in a multi-stage classification scheme to predict the future HbA1c class, or range, of a patient. The proposed method achieves a 79% accuracy on data provided by Sidra Medicine.

Motivation

Diabetes is a chronic metabolic disorder that has been growing at an alarming rate throughout the world. One of a widely used biomarkers for diabetes diagnosis and management is glycated hemoglobin (HbA1c). HbA1c is a blood test that measures the percentage of hemoglobin proteins in the blood that are coated with sugar. The proper management of diabetes depends on the periodic measurement, often every 3 months, of HbA1c levels in order to track the severity of a patient's diabetes. Assessing a patient's current HbA1c levels allows medical professionals to make necessary adjustments to the patient's lifestyle and dietary habits in order to avoid health complications that can come as a result of poor diabetes management. Instead of having to wait 3 months in order to assess a patient's HbA1c, we propose a method for predicting a patient's HbA1c at the end of a 3-month period based on their blood glucose data from the first 14 days of the 3-month period. This way, diabetes management can take a much more proactive form, allowing patients and physicians to make changes to treatment plans, lifestyle, and diet in the present to avoid elevated HbA1c levels and their associated consequences in the future.

Methodology

The proposed method for HbA1c level prediction assumes that the level of control that a patient has on their diabetes can be divided into four categories depending on which range a patient's HbA1c falls within. The four chosen control levels are demonstrated below:

Control level	HbA1c (%)
Good	HbA1c≤6
Medium	6<HbA1c≤7.5
Poor	7.5<HbA1c≤8.5
Terrible	HbA1c>8.5

The goal of the proposed method is to use the blood glucose data, given by the FreeStyle Libre sensor, from the first 14-days of a 3-month period to predict if the patient's HbA1c at the end of the 3-month period will indicate "good", "medium", "poor", or "Terrible" control of their diabetes.

The proposed methodology assumes that a blood glucose reading is taken every 15 minutes. One issue with using the data from the FreeStyle Libre sensor is that some data might be missing due to a variety of factors, including the patient taking off the sensor for one reason or another. As a result, a Machine Learning-based method is adopted to estimate the missing data points. This method assumes that the blood glucose readings taken at the same time on different days are correlated and that the readings taken at different times on the same day are correlated. Thus, a Machine Learning model is trained to estimate each of the available blood glucose readings given their 8 closest "neighbors". After training, this model is then used to fill-in any missing data for that patient. Note that a model is trained for each patient to learn that patient's blood glucose trends and used to fill-in any missing data they might have. The process for filling-in a patient's missing data is outlined below:

Day	12:00AM	12:15AM	12:30AM	12:45AM
1	74	71	68	74
2	149.5	172	166	162
3	141.5	140	137	138
4	290	**m1**	289	286
5	217	200	198	199
6	60	82	90	93
7	136	156	157	158
8	104	122	**m2**	135
9	173	163	169	172
10	228	213	199	178
11	194	191	201.5	197

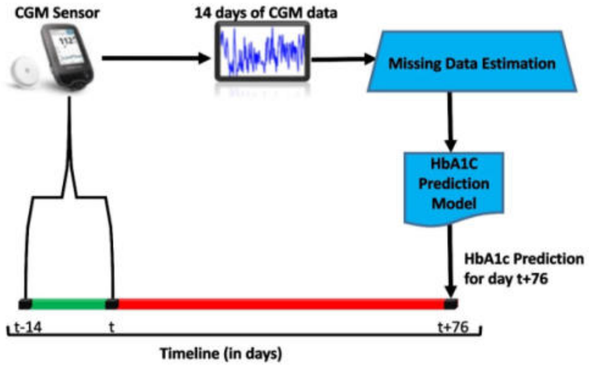
Training Set

N1	N2	N3	N4	N5	N6	N7	N8	Y
74	71	68	166	137	140	141.5	149.5	172
71	68	74	162	138	137	140	172	166
173	163	169	199	201.5	191	194	228	213

N1	N2	N3	N4	N5	N6	N7	N8	Missing
141.5	140	137	289	198	200	217	290	**m1**
156	157	158	135	172	169	163	122	**m2**

ML Model → m1 and m2 estimates

After the missing points of a patient have been filled-in, the HbA1c range of that patient at the end of the 90-day period can be estimated based on the blood glucose readings from the first 14 days of the 90-day period, as shown below:



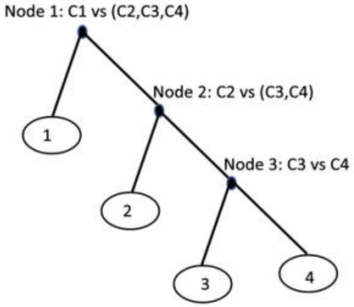
Instead of learning the HbA1c range from a time-series representation of the 14 days of blood glucose data, we perform a pre-processing step where we transform the time-series data into a greyscale image that consists of bars of varying widths and colors. Each bar in the image represents a blood glucose reading, where consecutive bars represent consecutive blood glucose readings. The maximum width K of any bar is set to 10 pixels while the height of the image is set to a constant 1344 pixels. The color C and width W or a bar representing a blood glucose reading BG are given by:

$$C = 255\,\frac{BG}{500}$$

$$W = K\,\frac{BG}{500}$$

Note that $\frac{BG}{500} \leq 1$ since the maximum blood glucose reading is 500. As a result, the color of a bar is within the range [0,255] where a bar with color 0 is black and a bar with color 255 is white. An example of a greyscale image representing a 14-day period of blood glucose readings is shown below:

Note that our fundamental task is classifying greyscale images such as the one above into one of four classes. We found that the most effective way to perform this classification is to use multi-stage classification stead of multi-class classification. In multi-class classification, a single model would take an image as input and outputs one of four classes as the prediction. On the other hand, a multi-stage classification scheme resembles a decision tree where each node of the tree answers the binary question "class c or not" where class c is any of the classes in question. If the answer at the tree node is "yes", the classification task is done. Otherwise, the input image continues down the tree. The general scheme of a mutli-stage classifier for a four-class classification task can be represented as follows:



Control level	# of Time Series
Good	64
Medium	72
Poor	99
Terrible	99

Note that there are $4! = 24$ different tree models that can be used for a four-class multi-stage classification scheme. Further note that the above tree requires three binary models, one at each node. The models used in our proposed methodology are binary Support Vector Machine (SVM) models that use the pixel values of the greyscale image representations as featurues.
In order to train binary models for a multi-stage classification scheme, a sufficient number of data is required for each of the four HbA1c classes. Patient data for each of the four classes were provided by Sidra Medicine. Blood glucose time-seires data for 90-day periods as well as the HbA1c value measured on the 90th day were provided. The breakdown of the number of time-series for each class was as follows:

Results

We evaluate the performance of the multi-stage approach by randomly selecting 20% of the given data for validation. Then, the accuracy of the model is measured as follows:

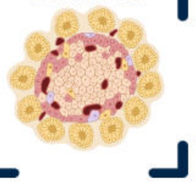
$$Accuracy = \frac{D_C}{D_T} \times 100$$

where D_C is the number of correctly classified instances and D_T is the number of total instances. Overall, the model was able to correctly classify 211/267 time-series, giving an accuracy score of 79%.

Conclusions

In the proposed algorithm, we transform blood glucose time-series data into greyscale images and use the greyscale images in a multi-stage classification scheme for HbA1C prediction. This approach allows us to predict general HbA1c levels 2.5 months in advance, a novel contribution that is absent from the literature. This approach promises to revolutionize diabetes management by making it more proactive.

SCAN ME



Welcome to our tech-focused conference! Our poster presentation features an innovative AR experience that you can access by scanning the image below. See the concepts come to life and get a deeper understanding of the topic at hand.

Don't forget to turn up the volume on your device for an explanation. During the networking breaks, feel free to connect with the presenters to discuss the topic further.

Thanks for joining us!