Short-term vs. Long-term Analysis of Diabetes Data: Application of Machine Learning and Data Mining Techniques

Eleni I. Georga, Vasilios C. Protopappas, Stavroula G. Mougiakakou, *Member, IEEE*, Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*— Chronic care of diabetes comes with large amounts of data concerning the self- and clinical management of the disease. In this paper, we propose to treat that information from two different perspectives. Firstly, a predictive model of short-term glucose homeostasis relying on machine learning is presented with the aim of preventing hypoglycemic events and prolonged hyperglycemia on a daily basis. Second, data mining approaches are proposed as a tool for explaining and predicting the long-term glucose control and the incidence of diabetic complications.**

I. Introduction

DIABETES is a chronic disorder of the glucose-insulin metabolism characterized mainly by high blood glucose concentrations. This condition, known as hyperglycemia, is associated with long-term microvascular and macrovascular complications, which have a great impact on the health-related quality of life of diabetic individuals [1]. On the other hand, one of the most important short-term adverse effects of insulin therapy is hypoglycemia which prevention, especially during night, is considered paramount [2]. The care of both type 1 and type 2 diabetes takes place (i) at home where patient should properly administer medical and lifestyle parameters to control his glucose avoiding hypoglycemia and (ii) in hospital where physician should prescribe the best therapy scheme for controlling both diabetes and its complications.

The literature suggests that predictive modeling of glucose metabolism based on self-monitoring data has the potential to contribute to a more safe therapy scheme reducing the risk of hypoglycemia [3]. In particular, the technological progress in continuous glucose monitoring (CGM) has enabled the accurate short-term (up to 60 min) prediction of subcutaneous (s.c.) glucose concentration through the application of both linear [4]-[6] and non-linear data driven approaches [7]-[10]. It has been demonstrated that the inclusion of information on meals, insulin therapy and physical activity allows for improved accuracy [8]-[10] when compared to the autoregressive analysis of the glucose time series. In [11], the fusion of real-time adaptive models (recurrent neural networks and autoregressive models) resulted in 100% prediction accuracy of hypoglycemic events with 16.7 min time lag and 0.8 daily false alarms. Moreover, in [12] the problem of hypoglycemic event prediction was handled differently for nocturnal and diurnal periods as regards input variables.

The chronic nature of diabetes results also in massive amounts of clinical healthcare data. A recent review on data mining technologies for diabetes [13] highlighted their value mainly in (i) interpreting and predicting the long-term glycemic status of the patients and (ii) identifying important predictors of glucose control and diabetic complications. For instance, Breault *et al.* [14] employed a classification and regression tree (CART) to predict a binary target variable of HbA1c>9.5, while, in [15], CART was applied for the identification of risk factors for type 2 diabetes vascular complications. It is recognized that the outcomes of the current research studies are limited to the dataset used; however, they lead to better understanding of diabetes.

In this paper, we propose the (i) the short-term prediction of glucose concentration for providing patients advanced knowledge of hypoglycemia or hyperglycemia on a daily basis and (ii) the mining of long-term clinical diabetes data for the extraction of new knowledge and the provision of support to physicians in decision making about treatment and risk of complications.

II. Short-Term Predictive Modeling of Glucose Concentration

A. A machine learning approach

The s.c. glucose concentration at time $t+l$, assuming that t is the time at which the prediction is made and l is the prediction horizon, is predicted by a non-linear function $f:\Re^d \to \Re$ of the form:

$$f(x) = w^T \phi(x) + b,$$

where w is an M-dimensional weight vector, b is the bias

Manuscript received July 30, 2013. This work was supported in part by the research project "Development of an Information Environment for Diabetes Data Analysis and New Knowledge Mining" that has been co-financed by the European Union (European Regional Development Fund-ERDF) and Greek national funds through the Operational Program "THESSALY- MAINLAND GREECE AND EPIRUS-2007-2013" of the National Strategic Reference Framework (NSRF 2007-2013).

E.I. Georga, V.C. Protopappas and D.I. Fotiadis are with the Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, GR 45110 Greece (e-mail: egeorga@cs.uoi.gr, vprotop@cc.uoi.gr, corresponding author phone: +302651008824; fax: +302651008889; e-mail: fotiadis@cc.uoi.gr).

S.G. Mougiakakou is with the University of Bern, Bern, CH 3010 Switzerland (email: stavroula.mougiakakou@artorg.unibe.ch).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

parameter, ϕ is a vector of M fixed non-linear basis functions and $x \in \Re^d$ is the input vector. In our study, the function f is built using two kernel based algorithms i.e. support vector regression (SVR) or Gaussian processes (GP) regression [16].

The input x is comprised of six (6) input variables x_i, $i = 1, \cdots, 6$ each one of which is described as follows:

1. $\left[gl(t-30), ..., gl(t-5), gl(t)\right]$: s.c. glucose measurements within the last 30 min.
2. $\left[Ra(t+l-30), ..., Ra(t+l-5), Ra(t+l)\right]$: rate of meal glucose appearance [17] of within the last 30 min with respect to the time of prediction $t+l$.
3. $\left[I_p(t+l-30), ..., I_p(t+l-5), I_p(t+l)\right]$: plasma insulin concentration [18] within the last 30 min with respect to $t+l$.
4. $\left[SRa(t+l-80), ..., SRa(t+l-10), SRa(t+l)\right]$: meal-derived glucose inserted in plasma, with respect to $t+l$, calculated cumulatively every 10 min over the last 90 min where $SRa(t+l-(8-i)10) = \sum_{\tau=t+l-90}^{t+l-(80-10i)} Ra(\tau)$ for $i = 0, \ldots, 8$.
5. $\left[SEE(t-170), ..., SEE(t-10), SEE(t)\right]$: energy expenditure calculated cumulatively every 10 min over the last three (3) hrs where $SEE(t-(17-i)10) = \sum_{\tau=t-180}^{t-(170-10i)} EE(\tau)$ for $i = 0, \ldots, 17$. The term EE expresses the instantaneous energy expenditure provided by the SenseWear Armband.
6. *h*: hour of day identifying the 24 hourly intervals within a day.

B. Support Vector Regression

The SVR prediction of glucose concentration for a new point *x* is:

$$f(x) = \sum_{i=1}^{N} (a_i - a_i^*) k(x, x^i) + b,$$

where a_i, a_i^* are the Lagrange multipliers, *k is* a Gaussian RBF kernel function and x^i with $i = 1, \ldots, N$ the training set. The hyper parameters *C*, *ε* and the kernel parameter *γ* were individually optimized using the Differential Evolution algorithm in which the objective function was defined as the average *RMSE* of the 10-fold cross validation [10].

C. Gaussian Processes

The prediction for a new point *x* by GP is a Gaussian distribution with mean and covariance given by:

$$m(x) = \sum_{i=1}^{N} a_i k(x^i, x), \sigma^2(x) = k(x, x) + \beta^{-1} - \kappa^T C^{-1} \kappa,$$

where k is an exponential quadratic kernel, C is the covariance matrix, the vector κ has elements $k(x^i, x)$ for $i = 1, \cdots, N$, a_i is the i^{th} component of the $C^{-1}t$, β^{-1} is the noise precision and t is the target vector. The hyperparameters of GP (i.e. kernel parameters and β) are learnt for each patient separately through the minimization of the negative log likelihood function using the standard form for a multivariate Gaussian distribution.

D. Dataset

The dataset is comprised of 15 type 1 diabetic patients following multiple-dose insulin therapy and who were monitored under free living conditions within the European-Union co-funded research project METABO [10]. Patients were equipped with the Guardian Real-Time CGM system (Medtronic Minimed Inc.) and the SenseWear Armband (BodyMedia Inc.) wearable physical activity monitor and they were also recording information on food intake and insulin regime on a daily basis. The characteristics of that dataset are described in our previous study [10].

E. Results

The predictive performance of both SVR and GP techniques is evaluated by applying 10-fold cross validation individually for each patient. Table 1 reports the average value and the standard deviation of the RMSE and correlation coefficient *r* accompanying 30-min and 60-min predictions, as well as the percentages of the successful hypoglycemic (≤70 mg/dl) and hyperglycemic (≥180 mg/dl) predictions. It can be seen that both regression techniques produce an average prediction error less than 10 mg/dl and a degree of correlation close to 1 for all horizons. Moreover, in critical glucose ranges, hyperglycemic values are predicted with higher accuracy than hypoglycemic ones. Figs. 1 and 2 show the actual vs. the predicted daily glucose profile of a patient for 30-min and 60-min horizon, respectively.

III. Data Mining of Long Term Clinical Data

Data mining techniques are a valuable asset to the development of intelligent tools that support clinical decision making. Diabetes clinical data that we have at our disposal concern 100 patients who were monitored for at least 5 years. In particular, information on the blood glucose control, the combination therapy, the clinical and laboratory tests, and the medical surveillance of diabetic complications was recorded. Besides, enhancing the procedure of interpreting a patient's status, the exploitation of that information will eventually provide new knowledge

TABLE I
AVERAGE PREDICTION ACCURACY OF THE SVR AND GP TECHNIQUES

	Prediction Horizon							
	30 min				60 min			
	RMSE	r	% Hypo	% Hyper	RMSE	r	% Hypo	% Hyper
SVR	6.03 (1.67)	0.99 (0.00)	87%	96%	7.14 (1.84)	0.99 (0.01)	83%	94%
GP	6.36 (4.49)	0.99 (0.03)	88%	95%	8.64 (8.95)	0.96 (0.10)	85%	88%

The value in parenthesis indicate the standard deviation. The RMSE is expressed in mg/dl.

regarding the correlations between therapy, glucose control and complications. Moreover, as a step towards customized diabetes care, the long-term prediction of a patient's response to a therapy is investigated.

In particular, the proposed analysis advances the current clinical practice in the care of patients with diabetes as regards the following procedures:

- **Interpretation of the status of diabetic patients** by identifying patterns between the treatment regimen, the glucose levels and the results of clinical and laboratory tests for each patient. Given the sequential nature of diabetes data, association analysis techniques (i.e. Apriori) able to handle both co-occurrence and dynamic relationships are utilized.
- **Extraction of new medical knowledge** concerning the correlation and similarity of anthropometric, metabolic and biological parameters, whose role involved in the clinical manifestations and complications of diabetes. For this purpose, suitable clustering techniques are examined that may reveal significant similarities among patients.
- **Individualization of the treatment regimen** evaluating and predicting the response of a patient in the current and a new treatment, respectively, so that every patient achieves his personal glycemic goals. By classifying patients into classes which allow the indirect assessment of the effectiveness of previous treatments or the prediction of a patient's response (i.e. the prediction of HbA1c) to new treatments, the efficacy of a therapy scheme may be examined. In this direction, the risk of developing a diabetic complication is tested using classification approaches.

This overall analysis will directly promote research on the treatment of the disease and especially the personalization / adaptation of both the treatment itself and the goals for controlling glycemia and the complications of diabetes.

IV. DISCUSSION AND CONCLUSION

In this study, we proposed the intelligent analysis of short- and long-term diabetes data with the aim of enhancing both daily and clinical care of the disease. We focused on the (i) prediction of s.c. glucose concentration by presenting an already evaluated machine learning regression approach and (ii) on the extraction of new medical knowledge from clinical diabetes data.

Predictive modeling of glucose response to various environmental stimuli is expected to result in the prevention of hypoglycemic events and of prolonged postprandial hyperglycemic excursions on a daily basis. The CGM signal itself is indeed the most important predictor but its auto-correlation function vanishes at about 30 min and thus allows only for short-term predictions of low accuracy [4]-[7]. The literature suggests the use of multivariate datasets for capturing the implicit or explicit effect of patient's contextual information on the regulation of glucose [8]-[12]. Moreover, data-driven techniques that may accurately approximate both linear and non-linear functions of the input variables, while exhibiting a very good generalization performance, are preferred [8]-[12]. The results of the present study demonstrated that both SVR and GP kernel-based techniques may produce individualized glucose models of high prediction accuracy in both normal and critical glucose ranges, when inputted with the selected input. The comprehensive evaluation of these techniques with respect to hypoglycemia, which has recently been presented in [12], showed that the method's sensitivity to nocturnal hypoglycemic events increases when additional variables accounting for recurrent hypoglycemia due to antecedent hypoglycemia, exercise and sleep are utilized.

Moreover, the use of data mining techniques is becoming

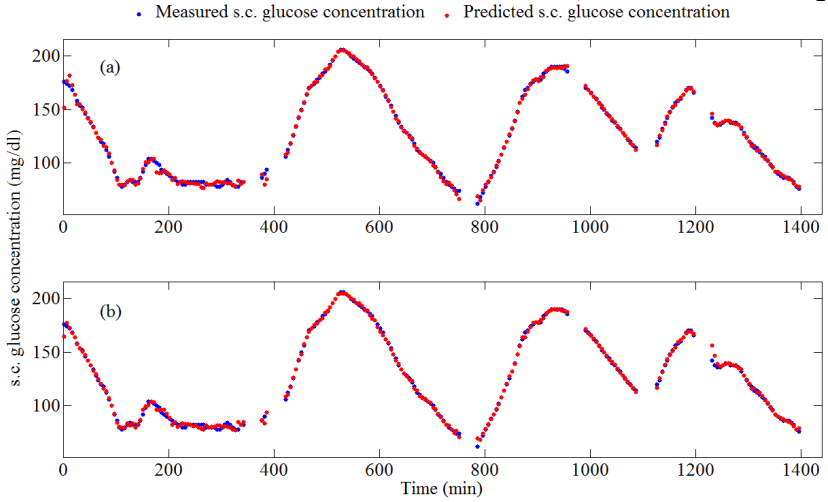

Fig. 1. Predicted vs measured s.c. glucose concentration of a typical subject during one day for 30-min horizon based on (a) SVR and (b) GP.

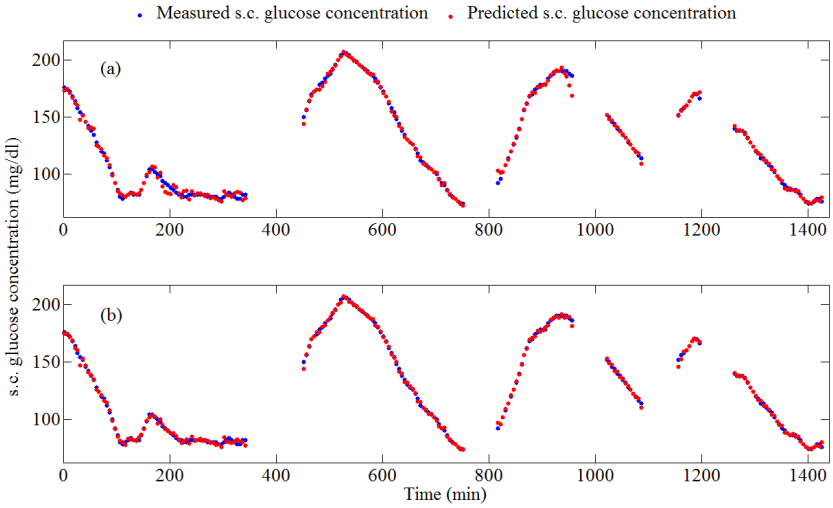

Fig. 2. Predicted vs measured s.c. glucose concentration of a typical subject during one day for 60-min horizon based on (a) SVR and (b) GP.

more prominent in the field of diabetes [13]. We proposed the provision of an integrated set of intelligent tools that facilitate the interpretation of the metabolic status of patients, the extraction of new medical knowledge and eventually the personalization of the treatment. These tools are based on the application of association analysis, clustering and machine-learning techniques for classification or regression on the large amounts of data resulting from the long-term clinical monitoring of diabetic patients. Firstly, the simultaneous analysis of a variety of medical, clinical and laboratory data and the creation of links between them helps to extract the additional information that cannot be acquired though an isolation process. In addition, the identification of groups of patients with high similarity between the studied parameters as well as the prediction of the long-term glycemic status contributes to a better treatment and thus evolution of the disease.

Concluding, the large amounts of data coming from the self-monitoring and the clinical monitoring of diabetes may reveal new knowledge about its short-term and long-term course, respectively. The capabilities of machine learning regression techniques and of data mining suggest promising future research direction in the field of diabetes.

REFERENCES

[1] American Diabetes Association, "Standards of medical care in diabetes", *Diabetes Care*, vol. 34, no. 1, pp. 11-61, Jan. 2011

[2] P. E. Cryer, "The barrier of hypoglycemia in diabetes", *Diabetes*, vol. 57, pp 3169-3176, 2008.

[3] C. Zecchin, A. Facchinetti , G. Sparacino and C. Cobelli, "Reduction of number and duration of hypoglycemic events by glucose prediction methods: a proof-of-concept in silico study", *Diabetes Technol Ther.*, vol. 15, no. 1, pp. 66-77, Jan. 2013.

[4] F. Zanderigo, G. Sparacino, B. Kovatchev, and C. Cobelli, "Glucose prediction algorithms from continuous monitoring data: Assessment of accuracy via continuous glucose error-grid analysis", *J Diabetes Sci Technol.*, vol. 1, no. 5, pp. 645-651, Sept, 2007.

[5] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, "Estimation of future glucose concentrations with subject-specific recursive linear models", *Diabetes Technol. Ther.*, vol. 11, no. 4, pp. 243-253, Apr. 2009.

[6] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: Data-driven glucose modelling", *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 246-254, Feb. 2009.

[7] C. Perez-Gandia, A. Facchinetti, G. Sparacino, C. Cobelli, E. J. Gomez, M. Rigla, A. de Leiva, and M. E. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring", *Diabetes Technol. Ther.*, vol. 12, no. 1, pp. 81-88, Jan. 2010.

[8] S. M. Pappada, B. D. Cameron, P. M. Rosman, A. E. Bourey, and T. J. Papadimos, "Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes", *Diabetes Technol. Ther.*, vol. 13, no. 2, pp. 135-141, Feb. 2011.

[9] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-term prediction of glucose concentration", *IEEE Trans Biomed Eng*,vol. 59, pp 1550-1560, 2012.

[10] E. I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression", *IEEE J Biomed Health Inform.*, vol. 17, pp. 71-81, 2013.

[11] E. Daskalaki, K. Nørgaard, T. Züger, A. Prountzou, P. Diem,and S. Mougiakakou, "An early warning system for hypoglycemic / hyperglycemic events based on fusion of adaptive prediction models", J Diabetes Sci Technol., vol. 7, no. 3, pp. 689-698, May 2013.

[12] E. I. Georga, V. C. Protopappas, D. Ardigo, D. Polyzos, and D. I. Fotiadis, "A Glucose Model Based on Support Vector Regression for the Prediction of Hypoglycemic Events Under Free-Living Conditions, *Diabetes Technol Ther.*, 2013, ahead of print.

[13] M. Marinov, A. S. Mosa, I. Yoo, and S. A. Boren, "Data-mining technologies for diabetes: a systematic review", *J Diabetes Sci Technol.*, vol. 1, no 6., pp. 1549-1556, Nov. 2011.

[14] J. L. Breault, C. R. Goodall, P. J. Fos, "Data Mining a Diabetic Data Warehouse", *Artif Intell Med.*, vol. 26, no. 1-2, pp. 37-54, Sep-Oct 2002.

[15] K. Miyaki, I. Takei, K. Watanabe, H. Nakashima, K. Watanabe,and K. Omae , "Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm", *J Epidemiol.*, vol. 12, no. 3, pp. 243-248, May 2002.

[16] C. M. Bishop, "Pattern recognition and machine learning", Springer 2006.

[17] E. D. Lehmann and T. Deutsch, "A physiological model of glucose-insulin interaction in type 1 diabetes mellitus", *J. Biomed Eng.*, vol. 14, no. 3, pp. 235-242, May 1992.

[18] C. Tarin, E. Teufel, J. Pico, and J. Bondia, and H. J. Pfleiderer, "A comprehensive pharmacokinetic model of insulin glargine and other insulin formulations", *IEEE Trans. Biomed. Eng.*, vol. 52, no. 12, pp. 1994-2005, Dec. 2005.