Predicting Diabetes in Healthy Population through Machine Learning

Lejla Alic*, Hasan Abbas*, Marelyn Rios†, Abdul Ghani§, and Khalid Qaraqe*
*Dept. of Electrical & Computer Engineering, Texas A&M University at Qatar, Doha, Qatar 23874
†Dept. of Industrial & Systems Engineering, Texas A&M University, College Station, TX 77843-3128
§Division of Diabetes, University of Texas Health Science Center at San Antonio,Texas
Email: LejlaResearch@gmail.com

***Abstract*—In this paper, we revisit the data generated by a population-based, epidemiological study, the San Antonio Heart Study, and use a machine-learning model to predict the future development of diabetes. To build this model, we used support vector machines (SVMs), ten features used in literature, and 171 diabetes patients and 1281 healthy participants. An optimal prediction model is trained in a 10-fold cross-validation (with 20 attempts), and validated using a hold-out set.**

***Index Terms*—OGTT, San Antonio heart study, diabetes prediction, machine learning, pre-diabetes**

I. Introduction

The global incidence of diabetes is estimated at 422 million in the year 2014 and its prevalence among the adult population has increased from 4.7 % in 1980 to 8.5 % in 2014 [1]. In 2015 alone, an estimated 1.6 million deaths worldwide were attributed to diabetes. In addition, a diabetic patient is at a greater risk of developing cardiovascular diseases, visual impairment and limb amputations, as compared to a non-diabetic person. Due to the substantial socio-economic burdens that are associated with diabetes, its early detection, intervention and prevention has become a worldwide top-level health concern.

Impaired glucose tolerance (IGT), defined by World Health Organization (WHO) [2] and the American Diabetes Association (ADA) [3], that is used to detect diabetes in its early stage, known as pre-diabetes, which identifies the impaired insulin response. Glucose clamp techniques can quantify the IGT. However, such techniques are labor-intensive and complicated for clinical practice or large epidemiological studies. A less invasive technique to quantify IGT involves an oral glucose tolerance test (OGTT) that samples the blood concentration of glucose and insulin over 2 hours after a standardized glucose dose [4]. However, the studies have shown that only 50 % of such cases actually develop diabetes within a span of 10 years [5], [6]. Moreover, 40 % diabetic subjects do not show any IGT in the initial screening. Previous studies have shown that extended OGTT, that assesses the blood glucose and insulin in the period prior to 2 h limit, can predict diabetes onset more reliably [7].

In this paper, we revisit the data generated by a population-based, epidemiological study, the San Antonio Heart Study [8], [9], and use a machine learning model to predict the onset of diabetes by using predefined OGTT features used earlier in the literature. On top of a range of glucose and insulin concentrations and their derivatives, our approach also takes into account physiological factors such as age, ethnicity, and body mass index (BMI).

II. Materials and Methods

A. San Antonio Heart Study

The data-set used in this paper is extracted from an epidemiological population study of risk factors related to diabetes and cardiovascular diseases, known as the San Antonio Heart Study (SAHS) [8], [9]. The study consisted of a total of 5158 men and non-pregnant women of Mexican-American and non-Hispanic white ethnicities, aged between 25 and 64 years and residing in San Antonio, Texas. All protocols applied in the study were approved by the institutional review board at the University of Texas Health Science Center, San Antonio. Blood samples of all the participants that went through an overnight fast, were drawn after orally administering a 75 g dose of glucose. After an average follow-up period of 7.5 years, the same participants were subjected to another round of OGTT. The participants in the SAHS study were enrolled in 2 phases, the first from January 1979 to December 1982, and the second, from January 1984 to December 1988 [10]. The reassessment during the follow-up period took place from October 1987 to November 1990 for the first phase, and October 1991 to October 1996 for the second phase. For this paper, we analyzed a subset of data from the second phase, with plasma glucose and insulin levels of 1496 participants measured at 0, 30, 60 and 120 minutes at baseline. At the follow-up assessment (average follow-up time of 7.5 years), the participants were classified as having type 2 diabetes (T2D), cardiovascular disease (CVD) or normal. For T2D diagnosis, the WHO criteria, defining fasting glucose level ≥126 mg/dL or 2-hour glucose level ≥200 mg/dL was followed [11]. Any participant reportedly taking anti-diabetic medications was also classified as diabetic. For CVD classification, any cardiovascular event such as a heart attack, stroke or angina reported by the participant, was considered as an identifier [8]. Table I outlines the distribution of patient classification used in this study. In order to construct a binary classifier, we have combined labels, T2D and both (total of 171 participants) indicating diabetes.

Healthy	DMI	CVD	Both
1281	161	44	10
85.63 %	10.76 %	2.94 %	0.67 %

TABLE I
The classification of the SAHS data-set with a total of 1496 participants

B. Data Processing

Processing and analysis was performed by in-house developed software (using MATLAB Version 9.2.0.556344), MathWorks, Inc., Natick, MA). We excluded the participants with any missing labels or measurements. Two socio-demographic factors (age and ethnicity), and eight physiological factors either measured during the OGTT or derived later, were used in this study, listed in Table II. These features have previously been used in diabetes prediction studies [7], [12]. The area under the 2 h glucose curve (AuG_{0-120}) was calculated using the trapezoidal rule, while the Matsuda index (M) was used as defined in [13]. $\Delta I/\Delta G_{0-120}$, where $x = 30, 120$ was calculated using the measured insulin and glucose values at time x during the OGTT.

Socio-demographic	Physiological	
	Measured	Derived
Age	BMI	AuG_{0-120}
Ethnicity	PG_0	Matsuda Index (M)
	PG_{120}	$\Delta I/\Delta G_{0-120}$
		$\Delta I/\Delta G_{0-30} \times M$
		$\Delta I/\Delta G_{0-120} \times M$

TABLE II
Features used in this study

C. Machine Learning

A supervised classification algorithm was used for the training and validation of the classification model. This minimizes the empirical risk associated with the errors on the training set [14], [15]. Support vector machines (SVM) have proven to be very effective in solving complex classification problems in many different application domains. SVM minimizes the marginal distance between the separating hyperplane and the closest data point of the multiple groups instead of minimizing the empirical error as used in other methods.

To predict diabetes onset, we defined a negative class (occurrence of diabetes at the follow-up) and a positive class (all other classes combined). As illustrated in Table I, the OGTT data used in this study is heavily unbalanced. With 171 negative indices and 1281 positive indices, the size of class labels is unbalanced with the ratio of negative-to-positive instances of 1:8. To validate the trained models, we used a holdout data set with the same unbalanced ratio of negative-to-positive class: 11 samples of negative class, and 88 samples of positive class. The remaining data was used to train and test the prediction models with a linear-kernel SVM model

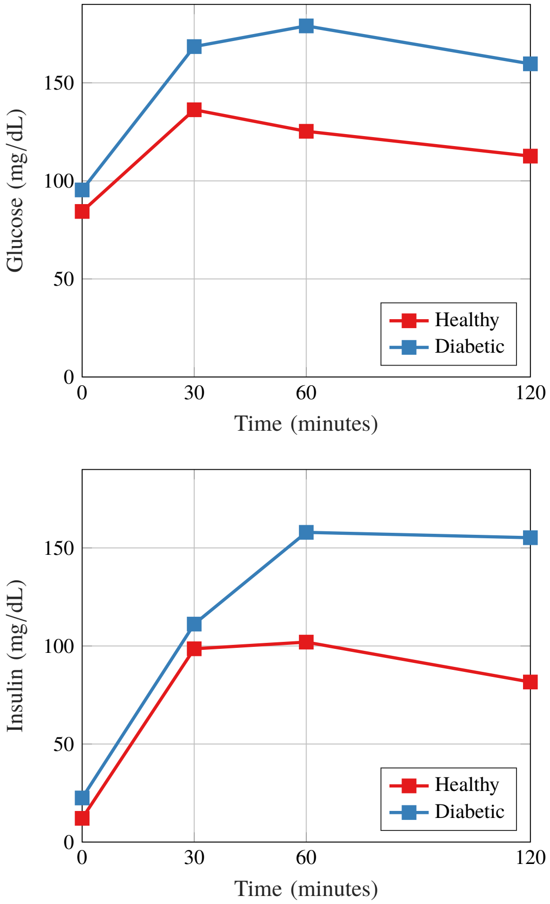

Fig. 1. Mean glucose and insulin curves of healthy and diabetic patients

in a 10-fold cross-validation framework, based on 90 % of the population in each fold. To prevent the over-fitting towards the majority class, we randomly under-sample the majority class to match the size of the minority class in the training set. Furthermore, to prevent violating the assumption of feature independency, we reduced the feature vector space by starting with classifying one feature at a time and linearly add features. We performed a total of 1023 classification experiments. Each of these experiments was trained as a 10-fold cross-validation (CV) with 20 attempts. Each of the attempts used randomly sampled data. In order to maximize reliability of the model to predict diabetes events, we maximized the specificity during the training phase. During the validation phase, we tracked the confusion matrices for all models with maximum training specificity representing all feature combinations separately.

III. Results

The data used is a subset of the SAHS study and include OGTT data of 1281 healthy subjects and 171 diabetes patients. Figure 1 illustrated the glucose and insulin curves, assessed at the baseline and labeled at the follow-up, for both groups separately.

A. Training

We trained ten prediction models with increasing number of features. Each of the SVM classifiers, was trained in a 10-fold CV by using in 20 attempts. The mean training accuracy, for one feature used, was 72%, and has increased to 89% with four features used. Simultaneously, sensitivity ranged from 50% (for one feature), trough 78% (for four features), to 81% (for eight features). Table III shows the training performance for all combinations of features separately.

Features	Accuracy	Sensitivity	Specificity
1	0.72	0.50	0.94
2	0.84	0.75	0.97
3	0.86	0.75	1
4	0.89	0.78	1
5	0.86	0.72	1
6	0.86	0.75	1
7	0.89	0.78	0.97
8	0.89	0.81	0.97
9	0.86	0.78	0.91
10	0.81	0.78	0.81

TABLE III
The averaged performance of the trained models demonstrating maximum specificity and their corresponding accuracy and sensitivity.

B. Validation

To validate the trained models, we used a holdout data set with the same unbalanced ratio of negative-to-positive class. Figure 2 represents the averaged validation performance for the models with maximized specificity. The performance is assessed by sensitivity, specificity and accuracy at the validation. The validation accuracy ranged from 48% (for one feature) to 80% (for eight features), while specificity ranged from 68% (for eight feature) to 91% (for one features). Figure 2 shows the validation performance of the models with maximized specificity during the training. All tree matrices are performing less compared to the training performance, with significant difference in performance (between training and validation) for low number of features. However, the higher number of features performed slightly less that training, but the difference was not significant.

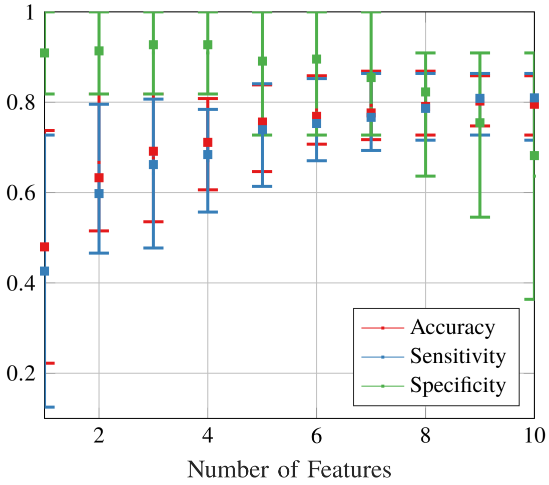

Fig. 2. The validation performance of the models with maximized training specificity. The error bars show average, minimum, and maximum values obtained in all 20 attempts.

IV. Conclusion and Future Perspectives

Diabetes prediction models potentially identify high-risk populations so that a timely population-based intervention could prevent future complications. In this paper, a machine learning approach was used to construct a prediction model of future development of diabetes.

The decreasing validation specificity for the increasing number of features indicates the potential over-fitting. Moreover, the features are clearly pendent, as they are generated from a limited number of measurements.

In conclusion, this study indicates that adding more features increases the performance of the models predicting diabetes in healthy population. However, in order to demonstrate the performance of our model, data sets with diverse ethnicities and different demographics are required.

V. Acknowledgment

This publication was made possible by NPRP grant number NPRP 10-1231-160071 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] C. D. Mathers and D. Loncar, “Projections of Global Mortality and Burden of Disease from 2002 to 2030,” *PLoS Medicine*, vol. 3, no. 11, p. e442, Nov. 2006.

[2] W. H. Organization and others, “Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation,” 2006.

[3] American Diabetes Association, “Diagnosis and Classification of Diabetes Mellitus,” *Diabetes Care*, vol. 28, no. Supplement 1, pp. S37–S42, Jan. 2005.

[4] O. Tschritter, A. Fritsche, F. Shirkavand, F. Machicao, H. Haring, and M. Stumvoll, “Assessing the Shape of the Glucose Curve During an Oral Glucose Tolerance Test,” *Diabetes Care*, vol. 26, no. 4, pp. 1026–1033, Apr. 2003.

[5] J. E. Shaw, P. Z. Zimmet, M. de Courten, G. K. Dowse, P. Chitson, H. Gareeboo, F. Hemraj, D. Fareed, J. Tuomilehto, and K. G. Alberti, “Impaired fasting glucose or impaired glucose tolerance. What best predicts future diabetes in Mauritius?” *Diabetes Care*, vol. 22, no. 3, pp. 399–402, Mar. 1999.

[6] N. Unwin, J. Shaw, P. Zimmet, and K. G. M. M. Alberti, “Impaired glucose tolerance and impaired fasting glycaemia: the current status on definition and intervention,” *Diabetic Medicine*, vol. 19, no. 9, pp. 708–723, Sep. 2002.

[7] M. A. Abdul-Ghani, K. Williams, R. A. DeFronzo, and M. Stern, “What Is the Best Predictor of Future Type 2 Diabetes?” *Diabetes Care*, vol. 30, no. 6, pp. 1544–1548, Jun. 2007.

[8] J. P. Burke, K. Williams, S. P. Gaskill, H. P. Hazuda, S. M. Haffner, and M. P. Stern, “Rapid Rise in the Incidence of Type 2 Diabetes From 1987 to 1996: Results From the San Antonio Heart Study,” *Archives of Internal Medicine*, vol. 159, no. 13, p. 1450, Jul. 1999.

[9] C. Lorenzo, K. Williams, K. J. Hunt, and S. M. Haffner, “Trend in the Prevalence of the Metabolic Syndrome and Its Impact on Cardiovascular Disease Incidence: The San Antonio Heart Study,” *Diabetes Care*, vol. 29, no. 3, pp. 625–630, Mar. 2006.

[10] S. M. Haffner, M. P. Stern, H. P. Haztjda, J. A. Pugh, and J. K. Patterson, “Hyperinsulinemia in a Population at High Risk for Non-Insulin-Dependent Diabetes Mellitus,” *New England Journal of Medicine*, vol. 315, no. 4, pp. 220–224, Jul. 1986.

[11] M. Wei, S. P. Gaskill, S. M. Haffner, and M. P. Stern, "Effects of Diabetes and Level of Glycemia on All-Cause and Cardiovascular Mortality: The San Antonio Heart Study," *Diabetes Care*, vol. 21, no. 7, pp. 1167–1172, Jul. 1998.

[12] M. A. Abdul-Ghani and R. A. DeFronzo, "Plasma Glucose Concentration and Prediction of Future Risk of Type 2 Diabetes," *Diabetes Care*, vol. 32, no. suppl_2, pp. S194–S198, Nov. 2009.

[13] M. Matsuda and R. A. DeFronzo, "Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp." *Diabetes Care*, vol. 22, no. 9, pp. 1462–1470, 1999.

[14] V. N. Vapnik, *Estimation of dependences based on empirical data: Empirical inference science: afterword of 2006*, 2nd ed., ser. Information science and statistics. New York, N.Y: Springer, 2006.

[15] ——, *The nature of statistical learning theory*, 2nd ed., ser. Statistics for engineering and information science. New York: Springer, 2000.