Predicting Future Type 2 Diabetes through Machine Learning

Hasan Abbas, Lejla Alic, Shafiqul Islam, Madhav Erraguntla, Qammer Abbasi, *Senior Member, IEEE,* and Marwa Qaraqe, *Member, IEEE*

***Abstract*—We present a prediction model that assesses the future risk of developing type 2 diabetes mellitus (T2DM). We use the oral glucose tolerance test (OGTT) data collected from a group of 1,551 healthy subjects to construct a machine learning model employing the support vector machines. We trained and validated the models on the data obtained from the second cohort of the San Antonio Heart Study, using a set of four features derived from the glucose measurements in the OGTT. The results of the proposed scheme show an average validation accuracy of 97.23% and recall of 77.27%. The results also show that the plasma glucose based features are the strongest predictors of the future development of T2DM.**

***Index Terms*—Type-2 Diabetes prediction, machine learning, disease risk assessment, San Antonio heart study.**

I. Introduction

THE global incidence of diabetes was estimated at 422 million in the year 2014 and its prevalence among the adult population has seen an increase from 4.7 % in the year 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

H. Abbas is with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874 Qatar; hasan.abbas@qatar.tamu.edu

L. Alic was with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874 Qatar; lejlaresearch@gmail.com

S. Islam and M. Qaraqe are with the College of Science and Engineering, Hamad Bin Khalifa University Doha, Qatar

M. Erraguntla is with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843

Q. Abbasi is with the School of Engineering, University of Glasgow, UK

Manuscript received April 19, 2005; revised August 26, 2015.

as age, ethnicity, and body mass index (BMI).

Does it improve the prediction performance over other models such as SADPM that do not require any invasive procedure. SADPM is based on demographics data and fasting plasma glucose.

II. Related Work

1. Abdul-Ghanis work on diabetes prediction model using statistical tools 2. Machine learning approaches primarily work on the aspects of diagnosis rather than prediction. BARAKAT rule based diagnosis 3. Ensemble based approach on diabetes diagnosis models

1. BARAKAT 2. ABDUl-GHANI 3. SAn Antonio Diabetes Prediction Model 4. Chinese paper using SVMS 5. Ensemble based approach for T2DM prediction

III. Materials and Methods

A. San Antonio Heart Study

We developed the diabetes prediction models using the data extracted from a population-based epidemiological study, the San Antonio Heart Study (SAHS). The aim of the SAHS was to assess the risk factors of diabetes and cardiovascular diseases [8], [9], for which 5,158 men and non-pregnant women of Mexican-American and non-Hispanic white residents of San Antonio, Texas were recruited. The age of the subjects at the time of recruitment was between 25 and 64 years. As a part of the data collection, the blood glucose and insulin levels were recorded during an oral glucose tolerance test (OGTT), which measures the subject's body response to a standard 75 g dose of glucose after fasting overnight. The OGTT was performed both at the baseline and follow-up phases of the study, which had an average span of 7.5 years. The SAHS subjects were recruited in 2 cohorts, the first during the period 1979 to 1982, and the second from 1984 to 1988 [10]. The reassessment during the follow-up period took place during the years 1987 to 1990 for the first cohort, and 1991 to 1996 for the second cohort. For the future T2DM prediction problem, we construct the machine learning model using the data from the second cohort, in which the plasma glucose and insulin levels of 1,496 of healthy subjects were recorded during the OGTT at times 0, 30, 60 and 120 minutes in the baseline evaluation. During the course of the study, a total of 171 subjects developed T2DM within which 10 subjects also reported of at least one cardiovascular event such as a heart attack, stroke or angina.

At the follow-up assessment, the 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 $\geq$126 mg/dL or 2-hour glucose level $\geq$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. Table I outlines the distribution of patient classification used in this study. In order to construct a binary classifier for this study, the subjects categorized under DMI and DMI+CVD were encoded by 1 whereas 0 was the label assigned to the healthy subjects. The subjects under the CVD catergory wrer not considered.

TABLE I: The classification of the 1,496 subjects used in this study

Healthy	DMI	CVD	DMI+CVD
1,281	161	44	10
85.63 %	10.76 %	2.94 %	0.67 %

B. Preparation of the Data

The dataset included the glucose and insulin values recorded at the baseline, 30, 60 and 120 minute intervals, and a distribution of these values marked by the follow-up labels of `healthy` and `diabetic` is shown in Fig. 1. Moreover, the socio-demographic information such as age, ethnicity and body-mass index (BMI) was also part of the dataset. From the glucose and insulin measurements, we computed the slope and area under the curve between all the possible combinations of a pair of readings. In addition, we also calculated parameters such as the insulinogenic (ratio of insulin and

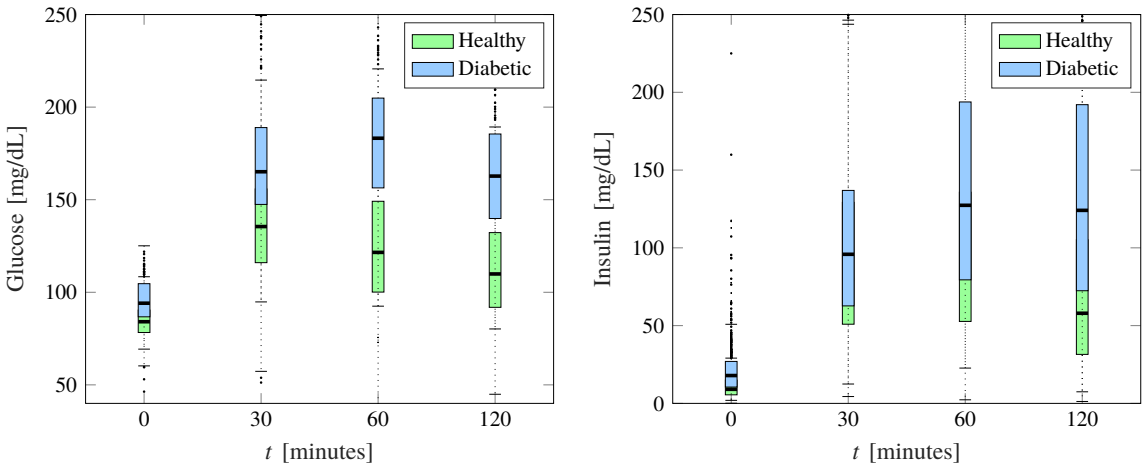

Fig. 1: Box plots of glucose and insulin measurements for healthy and diabetic subjects.

glucose slopes between any two time intervals) and Matsuda indices, as defined in [12], [13]. These variables have shown a good efficacy of diabetes prediction in previous studies [7], [13], since they are used to quantify the amount of insulin required by the body to maintain healthy glucose levels. In total we prepared 68 features for the classification problem and removed the rows that had missing entries or contained zero or infinite values. The dataset was then partitioned into training and validation sets. As can be observed from the Table I, the SAHS dataset is intrinsically imbalanced with the class distribution skewed toward the majority class with a ratio of 7.5:1. Considering the Therefore, we took two approaches for the training set, one in which the dataset was balanced by a randomized undersampling of the majority class, and second in which we persisted with natural class distribution. In both the cases, the validation set had the same class distribution representative of the original dataset. A total of 1360 samples were used for training that accounted for 91 % of the dataset, and 99 samples, consisting of 11 diabetic and 83 healthy subjects were reserved for the validation of the trained prediction models.

C. Machine Learning Framework

We used a supervised learning technique in which the classifier output labels were obtained from the follow-up data and a mapping function was developed using the support vector machines (SVM). This was done to minimize the empirical risk associated with the errors on the training set [**?**], [**?**]. We used the binary support vector machines (SVM) that have proven to be very effective in solving complex classification problems in many application domains, in which the data can not be easily separated into two classes. The SVM method aims to find an optimal hyperplane $\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0$, that maximizes the margin $d \sim 1/||\mathbf{w}||$, separating the two classes. The vector $\mathbf{x}$ denotes the training data and the vector $\mathbf{w}_0$ is a vector of weights expressed as a linear combination of the support vectors, z_i,

$$\mathbf{w}_0 = \sum_{i=1}^{N} c_i \mathbf{z_i}. \tag{1}$$

where N is the number of support vectors in the new feature space. Once the optimal hyperplane is obtained through the transformation, the linear decision classification function can be expressed as [**?**]:

1. Balanced Training -> Unbalanced Validation
2. Unbalanced Training -> Balanced Validation

$$I(\mathbf{x}) = \text{sign} \sum_{i=1}^{N} (c_i \mathbf{z_i} \cdot \mathbf{x} + b_0) \,. \tag{2}$$

In this study, we employed the linear SVM kernel by utilizing the Matlab's `svmtrain` function. The training data was first scaled to have a unit standard deviation. The misclassification cost was configured by setting the value of the `boxconstraint` parameter to a high value of 100, which would cause a stricter partitioning of the data with respect to the class labels.

To predict the future risk of type-2 diabetes, we defined a positive class (occurrence of diabetes at the follow-up) and a negative class (healthy). As illustrated in Table I, the OGTT data used in this study is heavily unbalanced. With 171 positive class instances as compared to 1281 that of the negative class, the size of class labels is unbalanced with the ratio of positive-to-negative instances of 1:8. To avoid the problem of overfitting to the majority class during the learning phase of the technique, we under-sampled the majority class (healthy) to the size of the minority class (diabetic) by a randomly selecting equal number of samples. During the prediction model generation, we employed 10-fold cross-validation framework in which 90 % of the training data, consisting of 360 samples was used for training and the remaining 10 % was used to test the model. To validate the trained models, we used a holdout data set with the same unbalanced ratio of negative-to-positive classes in the original data, i.e., 11 samples of the positive class, and 88 samples of the negative class. We started our experiments using one feature at a time, and then more number of features were incrementally added. This exercise assists in discovering any feature dependencies. In total, we performed 1,023 classification experiments. Each of these experiments was trained as a 10-fold cross-validation (CV) and, to minimize the effect of random selection of samples from the majority class, 100 iterations were performed for each experiment. Owing to the small sample size of the holdout dataset, this strategy ensures the unbiased reporting of the classifier performance. To maximize reliability of the model to predict diabetes events, we maximized the recall metric during the training phase, which is defined as,

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

where TP and FN are the true-positives and false-negatives respectively. During the validation phase, we tracked the confusion matrices for all the models yielding the maximum training recall for all the feature combinations.

D. Feature Selection

Before constructing the future diabetes prediction model based on the support vector machines, we aim to find the most effective subset of the features in terms of the relevance to the classifier output. This process greatly reduces the computational cost during the model development by reducing the feature space dimension and also dispense useful scientific insight in to the classification problem. We performed a two-step filter selection, first of which was the emphfilter method in which ten features that correlated the most to the classifier target class were shortlisted. In order to define the relevance between the featue and the class labels, consider a feature $\mathbf{X}$ from a feature set $\mathcal{X}$ as a continuous random variable and the class label $\mathbf{Y}$ as a discrete random variable. Their relationship can be described in terms of the mutual information, I defined as [14]:

$$I(\mathbf{X}, \mathbf{Y}) = -\int p_i \ln p_i - \sum_j p_j \ln p_j + \sum_j \int p_{ij} \ln p_{ij}, \tag{4}$$

where p_i is the probability of a random variable $\mathbf{X}$ taking a particular value x_i and so forth, and p_{ij} denotes the joint probability $P\{\mathbf{X} = x_i, \mathbf{Y} = y_j\}$. The three terms in (4) are the continuous, discrete and joint entropies of the random variables. The features that are most relevant to the class label are the ones that individually yield the maximum I. However, a drawback of pursuing this approach is that the selected features may be mutually correlated, and having a redundant list of shortlisted features only adds to the computational cost of the classifier without necessarily improving its performance. Even more so, the addition of extra features commonly

result in the deterioration of the classifier performance [15]. Therefore, an instinctive way forward is to keep only one feature from a correlated set of features that provide similar relevance information, and discard the remaining features from the set $\mathcal{X}$. We follow the minimal-redundancy-maximal-relevance () algorithm [16], that selects the features, that not only yield the maximal mutual information (4) with respect to the class label, but minimizes the mutual correlation among the features expressed in terms of redundancy R as:

$$R(\mathcal{X}) = \sum_{\mathbf{X}_i, \mathbf{X}_j \in \mathcal{X}} I(\mathbf{X}_i, \mathbf{X}_j). \tag{5}$$

where I is defined in (4). By minimizing R, the mRMR framework selects a set of mutually exclusive features that are most relevant to the class label. In this paper, we shortlist a set of ten features that are strong predictors of the future development of type-2 diabetes. With the application of the mRMR algorithm, Table II shows the list of the ten features ranked in order of their relevance to the class label, where the prefixes *au* and *sl* denote the area under the curve and slope respectively and the time interval is shown in the subscripts. In the second phase, we further pruned the number of features to four only by selecting the ones that provided the best performance in terms of validation accuracy using the SVM classification scheme. Table sdf shows the four best features obtained by averaging the validation accuracy over 100 iterations of the classification scheme.

TABLE II: List of ten most relevant features ranked by the mRMR algorithm

Rank	Feature
1	AuC-$Glu_{0\text{-}120}$
2	Sl-$Glu_{120\text{-}0}$
3	Sl-$Glu_{120\text{-}60}$
4	Sl-$Glu_{60\text{-}0}$
5	Sl-$Glu_{30\text{-}0}$
6	AuC-$Glu_{60\text{-}120}$
7	PG_0
8	PG_{120}
9	PG_{60}
10	AuC-$Glu_{0\text{-}30}$

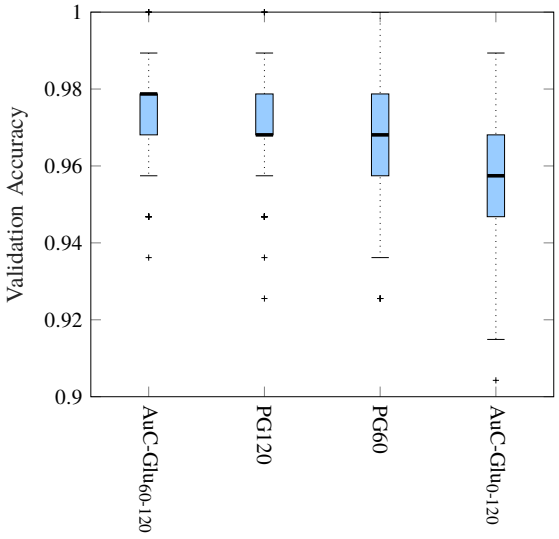

Fig. 2: Box plot of validation accuracy obtained after 100 iterations

IV. Conclusion

The conclusion goes here.

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. 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.

[13] 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.

[14] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, p. e87357, 2014.

[15] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on pattern analysis and machine intelligence*, no. 3, pp. 306–307, 1979.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1226–1238, 2005.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.