Predicting Diabetes in Healthy Population through Machine Learning

Lejla Alic[1], Hasan T. Abbas[1], Marelyn Rios[2], M. Abdul-Ghani[3] and Khalid Qaraqe[1]
[1]*Dept. of Electrical & Computer Engineering, Texas A&M University at Qatar, Doha, Qatar*
[2]*Dept. of Industrial & Systems Engineering, Texas A&M University,College Station, TX USA*
[3]*Division of Diabetes, University of Texas Health Science Center at San Antonio,TX USA*
LejlaResearch@gmail.com, {hasan.abbas, marelyn.rios, khalid.qaraqe}@qatar.tamu.edu, abdulghani@uthscsa.edu

Keywords: Type-2 Diabetes prediction, San Antonio heart study, machine learning, oral glucose tolerance test

Abstract: In this paper, we revisit the data of the San Antonio Heart study, and employ machine learning to predict the future development of type-2 diabetes. To build the prediction model, we use the support vector machines (SVMs) and ten features that are well-known in the literature as strong predictors of future diabetes. Due to the unbalanced nature of the dataset in terms of the class labels, we use 10-fold cross-validation to train the model and a hold-out set to validate it. The results of this study show a validation accuracy of 84.1 % with a recall rate of 81.1 % averaged over 100 iterations. The outcomes of this study can help in identifying the population that is at high risk of developing type-2 diabetes in future.

1 INTRODUCTION

The global incidence of diabetes was estimated at 422 million in the year 2014, and its prevalence among the adult population has seen in increase from 4.7 % in 1980 to 8.5 % in 2014 (Mathers and Loncar, 2006). In 2015 alone, an estimated 1.6 million deaths worldwide were directly attributed to diabetes. In addition, a diabetic patient is at a greater risk of developing cardiovascular disease, visual impairment and undergo limb amputations, as compared to a non-diabetic person. Due to the substantial socio-economic burdens not only to the effected families but the local healthcare system as well, the early detection, intervention and prevention of diabetes has become a paramount global concern related to health.

Impaired glucose tolerance (IGT) determines the abnormal insulin response in the body, and is considered one of the most important risk factors, both by the World Health Organization (WHO) (World Health Organization, 2006) and the American Diabetes Association (ADA) (American Diabetes Association, 2005), for detecting diabetes in its early stage, known as pre-diabetes. The IGT can be quantified by the glucose clamp technique, however, such an experiment is risky and requires highly qualified personnel, which limits its use in clinical practice or large epidemiological studies. A less invasive technique to quantify the IGT involves an oral glucose tolerance test (OGTT) in which the blood concentrations of glucose and insulin are assessed, in response to a standardized glucose dose taken orally before two hours of the measurement, after an overnight fasting state (Tschritter et al., 2003).

Studies have shown that only 50 % of the cases that exhibit the IGT go on to develop diabetes in future (Shaw et al., 1999; Unwin et al., 2002). On the other hand, 40 % diabetic subjects do not show any IGT in the initial screening. Previous studies have shown that extended the OGTT, that assesses the blood glucose and insulin intermittently during the 2 h time period can better predict the future risk of type-2 diabetes (Abdul-Ghani et al., 2007). In this paper, we extract the extended OGTT data from a population-based, epidemiological study, the San Antonio Heart Study (SAHS) (Burke et al., 1999; Lorenzo et al., 2006), and use a machine learning model to predict the future risk of diabetes. We use ten features that include a subject's demographic information and glucose characteristics derived from the OGTT measurements. The features are well-known as strong predictors of future diabetes in the literature. Here, we first describe the background and of the SAHS, after which we illustrate the machine learning technique used in this study. The results obtained during the training and validation phases are reported in terms of the accuracy, recall and specificity of the classifier models. Since the aim is to identify the high-risk subjects, we optimize the training models so that the recall (true positive rate) is maximized.

2 METHODOLOGY

2.1 San Antonio Heart Study

We extracted the dataset from an epidemiological population study of risk factors related to diabetes and cardiovascular diseases, known as the San Antonio Heart Study (SAHS) (Burke et al., 1999; Lorenzo et al., 2006). The study comprised of 5,158 men and non-pregnant women of Mexican-American and non-Hispanic white ethnicity, aged between 25 and 64 years and residing in San Antonio, Texas. All the protocols applied in the study were approved by the University of Texas Health Science Center, San Antonio institutional review board. 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 stages, the first from January 1979 to December 1982, and the second, from January 1984 to December 1988 (Haffner et al., 1986). 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 1,496 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 the T2D diagnosis, the WHO criteria, defining fasting glucose level $\geq$126 mg/dL or 2-hour glucose level $\geq$200 mg/dL was followed (Wei et al., 1998). Any participant reportedly taking anti-diabetic medications was also classified as diabetic. For the CVD classification, any cardiovascular event such as a heart attack, stroke or angina reported by the participant, was considered as an identifier. Table 1 outlines the distribution of patient classification used in this study. In order to construct a binary classifier, we have combined labels, T2D and both (a total of 171 participants) indicating diabetes.

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

Table 1: The classification of the SAHS data-set with a total of 1496 participants

2.2 Feature Selection

We selected 10 features for our prediction model consisting of socio-demographic variables such as age and ethnicity, and physiological factors that were either directly measured or derived from the OGTT. These features have individually been used in previous T2DM prediction studies (Abdul-Ghani et al., 2007; Abdul-Ghani and DeFronzo, 2009). A complete list of features used is shown in Table 2. Subjects having any missing feature values or labels were excluded before the model generation. We used Matlab to develop the machine learning routines and data processing. 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 (Matsuda and DeFronzo, 1999). The insulin sensitivity, $\Delta I/\Delta G_{0-120}$, where $x = 30, 120$ was calculated using the measured insulin and glucose values at time x during the OGTT.

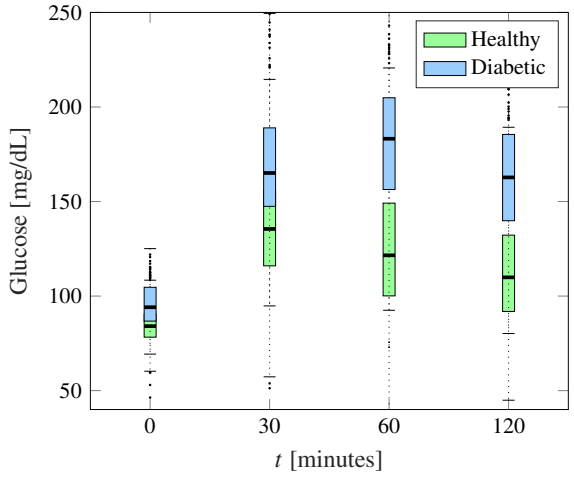
Figure 1 shows the box plots of the glucose and insulin values obtained during the OGTT at baseline, 30, 60 and 120 min, where the color associated with the boxes indicates the future label assigned at the the followup assessment. The box plot of the glucose values at 120 min provides a clear separation of the two classes. However, it must be noted that there are outliers either side of one standard deviation from the median which is depicted by a solid line in the box.

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

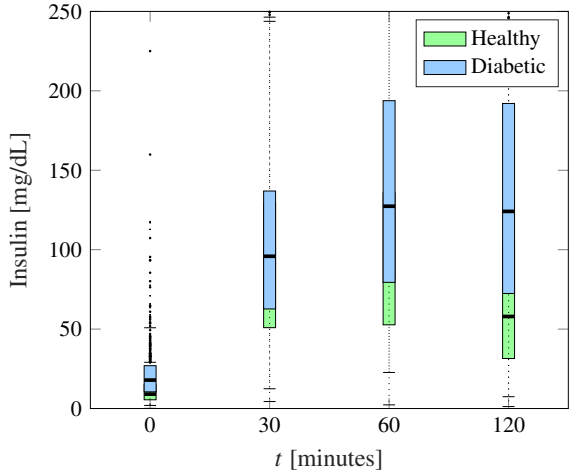
Table 2: Features used in this study

2.3 Machine Learning

We used a supervised learning technique in which the classifier labels were known *a priori* from the follow-up data. The data was partitioned for the training and validation subsets before the classification. This was done to minimize the empirical risk associated with the errors on the training set (Vapnik, 2006; Vapnik, 2000). 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



(a)



(b)

Figure 1: Box plots showing the distribution of plasma glucose (a), and insulin (b) measurements during the 2-hour OGTT

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 (Cortes and Vapnik, 1995):

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

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 1, 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.

3 RESULTS

The aim of this paper is to devise a machine learning scheme that can identify healthy subjects that are at an increased risk of developing type-2 diabetes. For this, the data used here is a subset of the SAHS that includes the OGTT data of 1,496 healthy subjects at baseline, out of which 171 were labeled as diabetic at the follow-up assessment and 1,281 maintained their healthy status. To determine the performance of our prediction models, we use accuracy, recall and specificity of the models. During the training, we emphasized on maximizing the recall of the classifier which in other words, maximizes the identification rate of high-risk diabetes. Using the strategy described in the previous section, we show the performance results that are averaged over 100 iterations.

3.1 Training

We trained ten prediction models with an increasing number of features. Each of the SVM classifiers was trained through a 10-fold cross-validation. The trained model was obtained by selecting the one that yielded the maximum accuracy averaged over 100 iterations. As an example, the feature, AuG_{0-120} provided a mean accuracy of 72 % which was greater than the accuracy given by all the other one feature models. The model obtained using a combination of two features (AuG_{0-120} and PG_{120}) generated 84 % accuracy. In the meantime, the recall increased from 94 % to 97 % by adding one feature. The maximum average accuracy during the training was obtained when four features (AuG_{0-120}, PG_{120}, age, and ethnicity) were used (see Table 3). The performance did not improve with further increments in the number of features. This suggests that the newly added features may not be independent to the existing ones.

Features	Accuracy	Specificity	Recall
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 3: The averaged performance of the trained models demonstrating maximum recall and their corresponding accuracy and specificity.

3.2 Validation

To validate the trained models, we used a holdout data set with the same unbalanced ratio of positive-to-negative class. Due to the small sample of the minority class and in order to avoid overlapping with the training set, only 11 diabetic samples were used. Figure 2 shows the box plots for the validation recall, accuracy and the specificity of the models that were trained to maximize the recall rate of the classifier. The same trends observed during the training were also seen in the validation phase. The combination of the four features that yielded the best training performance also produced the highest median recall rate. Adding more number of features resulted in slight improvement in the median accuracy. A worsening trend in the performance was observed when the number of features was more than seven. Figure 2 shows the validation performance of the models with maximized recall during the training.

4 DISCUSSION

Development of the classifiers on an unbalanced dataset poses a typical machine learning problem that results in he trained models being biased towards the majority class. In this study, we balanced the two classes with the aim to get unbiased models in the training. The classification threshold that controls the probability of a sample belonging to a certain class, can be varied to maximize the true positive rate (recall) of the classifier. Validation on the holdout data, on the other hand provides an independent assessment of the classifier.

5 CONCLUSIONS

Diabetes prediction models identify the high-risk population so that a timely population-based intervention could prevent future complications. In this paper, we used the linear support vector machines to construct a prediction model of future development of type-2 diabetes.

The outcomes of the study show that high values of glucose observed at the 2 h mark during the OGTT may strongly indicate the potential risk of future development of type-2 diabetes. In a possible extension of this study, the prediction models may be applied on other similar datasets that include the OGTT measurements.

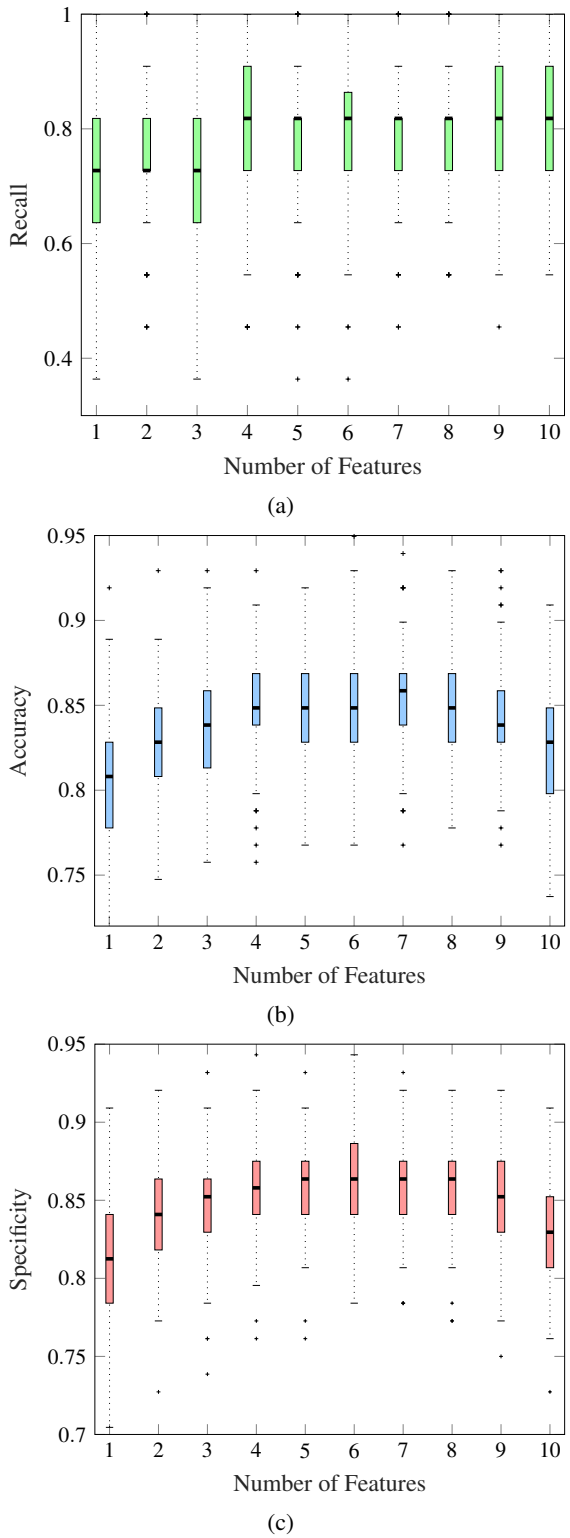


Figure 2: The validation performance of the models with maximized training recall. The box plots were obtained after 100 iterations of running the classifier.

ACKNOWLEDGEMENTS

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

Abdul-Ghani, M. A. and DeFronzo, R. A. (2009). Plasma Glucose Concentration and Prediction of Future Risk of Type 2 Diabetes. *Diabetes Care*, 32(suppl_2):S194–S198.

Abdul-Ghani, M. A., Williams, K., DeFronzo, R. A., and Stern, M. (2007). What Is the Best Predictor of Future Type 2 Diabetes? *Diabetes Care*, 30(6):1544–1548.

American Diabetes Association (2005). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 28(Supplement 1):S37–S42.

Burke, J. P., Williams, K., Gaskill, S. P., Hazuda, H. P., Haffner, S. M., and Stern, M. P. (1999). Rapid Rise in the Incidence of Type 2 Diabetes From 1987 to 1996: Results From the San Antonio Heart Study. *Archives of Internal Medicine*, 159(13):1450.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Haffner, S. M., Stern, M. P., Haztjda, H. P., Pugh, J. A., and Patterson, J. K. (1986). Hyperinsulinemia in a Population at High Risk for Non-Insulin-Dependent Diabetes Mellitus. *New England Journal of Medicine*, 315(4):220–224.

Lorenzo, C., Williams, K., Hunt, K. J., and Haffner, S. M. (2006). Trend in the Prevalence of the Metabolic Syndrome and Its Impact on Cardiovascular Disease Incidence: The San Antonio Heart Study. *Diabetes Care*, 29(3):625–630.

Mathers, C. D. and Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*, 3(11):e442.

Matsuda, M. and DeFronzo, R. A. (1999). Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care*, 22(9):1462–1470.

Shaw, J. E., Zimmet, P. Z., de Courten, M., Dowse, G. K., Chitson, P., Gareeboo, H., Hemraj, F., Fareed, D., Tuomilehto, J., and Alberti, K. G. (1999). Impaired fasting glucose or impaired glucose tolerance. What best predicts future diabetes in Mauritius? *Diabetes Care*, 22(3):399–402.

Tschritter, O., Fritsche, A., Shirkavand, F., Machicao, F., Haring, H., and Stumvoll, M. (2003). Assessing the Shape of the Glucose Curve During an Oral Glucose Tolerance Test. *Diabetes Care*, 26(4):1026–1033.

Unwin, N., Shaw, J., Zimmet, P., and Alberti, K. G. M. M. (2002). Impaired glucose tolerance and impaired fasting glycaemia: the current status on definition and intervention. *Diabetic Medicine*, 19(9):708–723.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, New York, 2nd ed edition.

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

Wei, M., Gaskill, S. P., Haffner, S. M., and Stern, M. P. (1998). Effects of Diabetes and Level of Glycemia on All-Cause and Cardiovascular Mortality: The San Antonio Heart Study. *Diabetes Care*, 21(7):1167–1172.

World Health Organization (2006). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation.