Support Vector Machine to Predict Type 2 Diabetes using Oral Glucose Tolerance Test

Hasan T. Abbas, Lejla Alic, Madhav Erraguntla, Muhammad Abdul-Ghani, Jim X. Ji, *Senior Member, IEEE,* Qammer H. Abbasi, *Senior Member, IEEE,* and Marwa Qaraqe, *Member, IEEE*

***Abstract*—Diabetes is a large healthcare burden worldwide. There is substantial evidence that lifestyle modifications and drug intervention can prevent diabetes, therefore, an early identification of high risk individuals is important to design targeted prevention strategies. In this paper, we present an automatic tool that uses machine learning techniques to predict development of type 2 diabetes mellitus (T2DM). Data generated from an oral glucose tolerance test (OGTT) was used to develop a predictive model based on the support vector machine (SVM). We trained and validated the models using the OGTT and demographic data of 1,496 healthy individuals collected during the San Antonio Heart Study. This study collected blood glucose and insulin concentrations before glucose intake and at three time-points thereafter (30, 60, and 120 minutes). Furthermore, personal information as age, ethnicity and body-mass index was also a part of the dataset. Using 11 blood measurements, we have deduced 61 features, which are then assign a rank and the top ten features are shortlisted using Minimum Redundancy Maximum Relevance feature selection algorithm. All possible combinations of the 10 best ranked features were used to generate the SVM based prediction models. This research shows that an individual's abnormal blood glucose levels, and the information derived therefrom have the strongest predictive performance. Significantly, the insulin and demographic features do not provide additional performance improvement for diabetes prediction. The results of this work identify the parsimonious clinical data needed to be collected for an efficient prediction of T2DM. Our approach shows an average accuracy of 96.81% and a specificity of 80.45% obtained on a holdout set.**

***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 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 to the high mortality rate, an individual with diabetes is at a greater risk of developing cardiovascular diseases, visual impairment and limb amputations, as compared to a non-diabetic individual. Due to the substantial socio-economic burdens that are associated with diabetes, its early detection, prevention, and management has become a worldwide top-level health concern. There is experimental evidence that the development of diabetes can be delayed or even prevented provided an individual undertakes a lifestyle change that includes diet management, adopting exercise, and adhering to a pharmacological treatment [2]. The early identification of high risk individuals of diabetes is therefore, essential for targeted prevention strategies [3].

Even though the number of clinical studies aimed at diagnosing diabetes have been growing in the last two decades, studies predicting the risk of developing diabetes are limited. This subject has lately received an increased amount of research interest [4]. However, the clinical significance of such predictions largely depend on the type and quality of data collected. There are studies that assign a probability to the future risk of diabetes using socio-demographic characteristics such as age, ethnicity, body-mass index (BMI) and genealogical information collected through population [5], [6]. Due to the unreliable data collection, such techniques can be misleading. The collection of blood samples, on the other hand, provides more reliable data and is a first step towards the disease prognosis with a deeper clinical insight [7]. oral glucose tolerance test (OGTT) is commonly used to screen diabetes [8] and to provide a critical understanding of its future evolution [9]. In an OGTT, the blood glucose and insulin levels are measured at regular intervals in a 2-hr period after orally administering a standard dose of glucose [9]. The glucose tolerance and insulin resistance are two of the most significant parameters deduced from the OGTT that are widely regarded as the major factors in the development of type 2 diabetes mellitus (T2DM).

A precursory stage of diabetes, commonly referred to as prediabetes, exists before overt T2DM, and is described by an impaired glucose tolerance (IGT). According to the World Health Organization (WHO) diagnostic criteria, the IGT is defined as fasting blood glucose level of >126 mg/dL and a 2-hour blood glucose level in the range of 140 mg/dL to 199 mg/dL, measured during the OGTT [10]. Although prediabetes is considered as an intermediate stage in the natural progression of T2DM [11], it has been reported that only 50 % of the subjects diagnosed with IGT developed diabetes within 10 years [12], [13]. Moreover, long-term population studies have also shown that around 50 % of the diabetic patients

H. T. Abbas and J. X. Ji are with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874 Qatar; hasan.abbas;jim.ji@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

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

M. Abdul-Ghani is with the UT Health, San Antonio, TX 78229; abdulghani@uthscsa.edu

Q. H. Abbasi is with the School of Engineering, University of Glasgow, UK; qammer.abbasi@glasgow.ac.uk

M. Qaraqe is with the College of Science and Engineering, Hamad Bin Khalifa University Doha, Qatar; mqaraqe@hbku.edu.qa

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

did not exhibit IGT at any time prior to the diagnosis [14]. This suggests that the fasting and 2-hour blood glucose levels used in and of themselves cannot accurately predict the future development of T2DM.

Machine learning (ML) has been proposed as a viable instrument for diabetes screening. In contrast to traditional diagnostic techniques employing population based statistics, ML methods develop models that are trained using large amounts of data. Barakat et al used socio-demographic information, and point of care testing from blood and urine to develop diagnostic models of diabetes [15]. This approach uses support vector machine (SVM) along with a rule-based explanation to provide a comprehensibility of the results to the clinicians. The blood glucose levels at baseline and 2-hr were among the features used. Han et al employed an ensemble SVM and random forest learning approaches to develop a decision making algorithm for the diagnosis of diabetes [16]. However, investigations that are designed to identify individuals at high risk of developing T2DM in future are limited. The San Antonio diabetes prediction model (SADPM) [17] uses a logistic regression supported by physiological parameters such as systolic blood pressure and cholesterol level. The underlying causes of T2DM in the form insulin resistance and insulin secretion were studied to develop a prediction model in [14]. In another study, multivariate logistic models using the blood glucose values measured in the OGTT were used to predict the future risk of developing T2DM [18], [19]. The predictive power of different biomarkers such as the fasting blood glucose level, BMI and hemoglobin A1C (HbA1c) for T2DM onset was assessed in [**?**]. This study focused on individuals with metabolic syndrome, a complex and serious health condition that greatly increases the risk of cardiovascular disease (CVD) and diabetes.

The standard ML algorithms are designed to yield optimal performance in terms of accuracy over the full dataset. However, medical applications such as disease diagnosis and prediction require a biased decision-making mechanism that favors one of the classes. This approach inherently maximizes the performance of the class that is more relevant in clinic terms. Therefore, the objective in such applications is to design a classifier that improves the accuracy of the class that is clinically more relevant. Additionally, often the amount of data is highly skewed with the clinically relevant class in an outsized minority. There are various roundabout ways to obtain accurate classifier performance in this scenario that include the method of sampling [20] in which the class distribution is artificially balanced by either under sampling the majority class, over-sampling the minority class or both. Furthermore, feature weighting schemes assign distinct costs to training examples [21] in order to introduce a certain bias. Other techniques introduce evaluation metric such as the geometric mean (g-mean) [22], that concurrently optimizes the positive class accuracy (sensitivity) and the negative class accuracy (specificity) [23].

We hypothesized that the features extracted from the OGTT will be able to predict the future onset of T2DM. In this paper, we therefore propose a screening tool that identifies the most relevant features extracted from the OGTT data that strongly correlate with the future development of T2DM. We then use SVM to develop a prediction model by utilizing these relevant features estimated from the longitudinal cohort study, the San Antonio Heart Study (SAHS) [24], [25].

II. Materials and Methods

A. San Antonio Heart Study

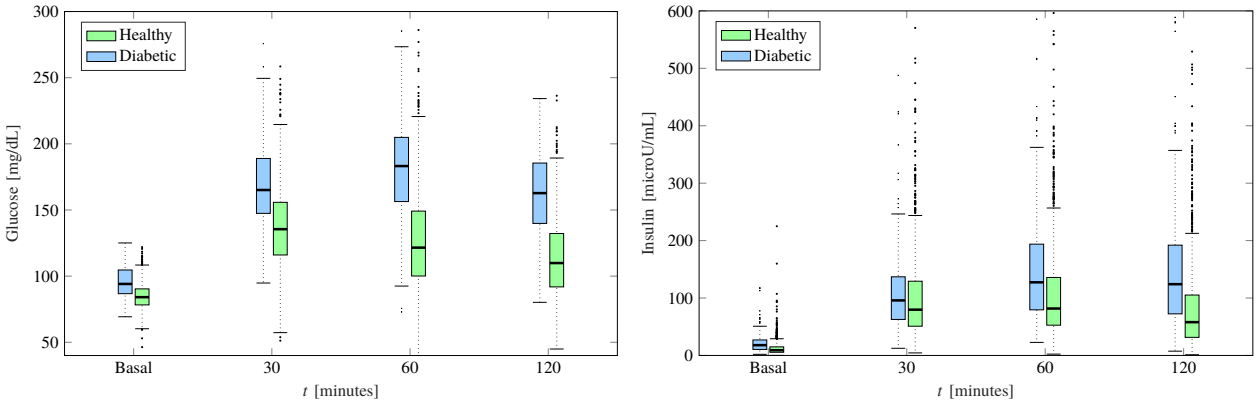
The SAHS is a population-based epidemiological study that was conducted to assess the risk factors of diabetes and cardiovascular diseases in healthy population [24], [25]. In total, 5,158 men and non-pregnant women of Mexican American (MA) and Non-Hispanic White (NHW) residents of San Antonio, Texas participated in the study in two cohorts. The age of the individuals at the time of recruitment was between 25 and 64 years. As a part of the data collection, the blood glucose and insulin concentrations were collected during the oral glucose tolerance test at the baseline and after an average follow-up of 7.5 years. The BMI was also recorded for each individual at the baseline. In this study, we analyzed only the data generated from the second cohort of the SAHS which comprised 1,492 subjects from the second cohort of the SAHS.

T2DM was diagnosed at the follow-up using the WHO criteria, i.e. fasting glucose level $>126\,\mathrm{mg/dL}$ or 2-hr glucose level $\geq 200\,\mathrm{mg/dL}$ [10]. Furthermore, all individuals taking anti-diabetic medications were also classified as having T2DM. Individuals that reported by themselves any cardiovascular event such as heart attack, stroke or angina, were labeled as having CVD at the follow-up. All other participants without T2DM or self-reported CVD were labeled as healthy for the case of this study. During the course of the longitudinal study, a total of 171 individuals developed T2DM with 10 individuals also reporting at least one cardiovascular event. The incidence rate of T2DM in the second cohort of the SAHS population was 10.79 %. Table I shows the population distribution in terms of the four classes. The distribution in terms of the ethnicity shows a more than double T2DM prevalence among the MA individuals, as compared to the NHW population.

TABLE I: The classification of the 1,492 subjects used in this study

	Healthy	T2DM	CVD	T2DM+CVD
Total	1,277	161	44	10
	85.56 %	10.79 %	2.95 %	0.67 %
MA	836	131	24	7
	83.77 %	13.13 %	2.40 %	0.70 %
NHW	441	30	20	3
	89.27 %	6.07 %	4.05 %	0.61 %

The data used in this study consists of the blood glucose and insulin concentrations sampled at baseline, and at 30, 60 and 120 min thereafter. The individuals are labeled at the SAHS follow-up using the current standard of care [24]. Figure 1 shows the distributions of the data used in this study.



(a) (b)

Fig. 1: Box plots of glucose and insulin levels for healthy and diabetic subjects measured at the baseline OGTT.

B. Machine Learning Framework

In this paper, we implemented SVM to construct the models for prediction of future T2DM. The SVM develops models from a given training dataset such that it generalizes well to a new dataset and minimizes the empirical risk associated with misclassification of samples in the training set [26], [27]. A model constructed by the SVM minimizes the overlap between the classes in the training set by optimizing the separating hyperplane. For problems that may not be amenable to linear separation between the two classes, the SVM technique is very attractive due to fact that the input feature space can be transformed to a higher dimension space, and a linear boundary can then be determined. This approach generally provides a better training performance but potentially increases computational complexity excessively with the increase of the dimensionality of the input feature space [28]. Introduction of a kernel alleviates the need to determine the transformation by calculating the inner product between the coordinates of the input feature space instead. In this paper, we used the Gaussian radial basis function (RBF), as the kernel. The performance of the SVM can be optimized by tuning the free parameter of the kernel σ and specifying a cost that controls the rigidity of the class margin. This process is normally carried out through a grid search.

C. Feature Extraction

We extracted all the features from the SAHS data acquired at the baseline. The dataset consists of the blood glucose and insulin concentrations recorded before glucose intake and at three time-points thereafter (30, 60, and 120 min). The labels (healthy and diabetes) were generated at the 7.5 years follow-up using the current standard of care diagnostics [24]. From the glucose and insulin concentrations, we computed the slope and area under the curve between all the possible combinations of a pair of measurements. In addition, we also calculated three empirical markers that describe the relationship between the glucose intake and insulin response. The first is the insulinogenic index (IGI) [29], which is a direct measure of the insulin response to glucose. It is calculated as the ratio of the slope of insulin curve to the slope of glucose curve between any two progressive time intervals in the OGTT. The second marker, Matsuda index (MI) M, evaluates the insulin sensitivity from the OGTT using a product of the weighted averages of the glucose and insulin concentrations [30],

$$\mathrm{M} = \frac{10,000}{\sqrt{Glu_0 \times Ins_0 \times \frac{15Glu_0+30Glu_{30}+30Glu_{60}+30Glu_{90}+15Glu_{120}}{120} \times \frac{15Ins_0+30Ins}}} \tag{1}$$

where the subscripts depict the time point of the OGTT. In case when the value at 90 minutes is not available, the average of 60 and 120 minutes is used instead [30]. The third marker, homeostatic model assessment - insulin resistance (HOMA-IR) [31] evaluates the beta-cell function. It is defined as the product of fasting blood glucose concentration and fasting blood insulin concentration divided by 22.5. These markers have been used to estimate abnormalities in the insulin sensitivity. A total of 61 features (illustrated in Fig. 2) are used in this study. The prefix AuC denotes the area under the curve and the slope is denoted by the symbol Δ. The term $\mathrm{T}_{\mathrm{half}}$ represents the linearly interpolated value between any two intervals. The OGTT time interval in minutes corresponding to the feature appears in the subscripts.

D. Feature Selection

Before constructing the SVM model to predict a future diabetes occurrence, we search for the most effective subset of features in terms of relevance to the classifier output, i.e. incidence of T2DM at the follow-up. As a first step, we selected the ten most relevant features from the 61 available features using the minimum redundancy maximum relevance (mRMR) algorithm [32], which selects the most relevant features with minimum correlation among them. The mRMR algorithm determines the relevance between a feature (x as continuous random variable) and the class label (y as discrete random variable) in terms of the mutual information I defined

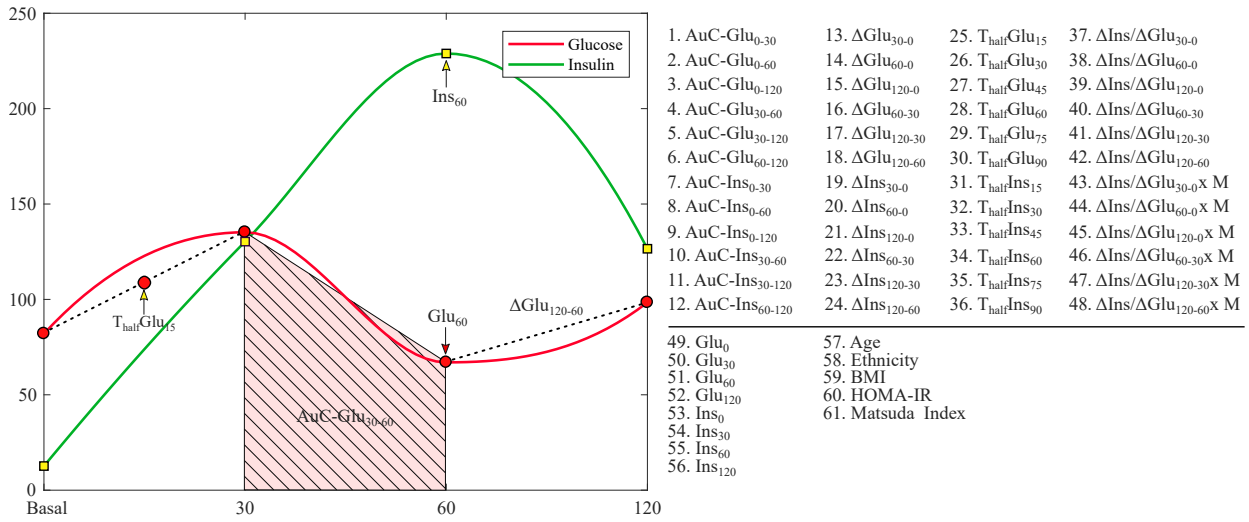


Fig. 2: Illustration of all 61 features extracted from the SAHS dataset.

as [33],

$$\mathcal{I}(x,y) = -\int p_i \ln p_i \,\mathrm{d}x - \sum_j p_j \ln p_j + \sum_j \int p_{ij} \ln p_{ij} \,\mathrm{d}x, \tag{2}$$

where p_i, and p_j are the probabilities of the random variables x and y taking a particular value x_i and $y_j \in (-1, 1)\,\forall j$ respectively. The term p_{ij} denotes the joint probability $P\{x = x_i, y = y_j\}$. The three terms in (2) represent the continuous, discrete and joint entropies of the random variables in the respective order. The features that are most relevant to the class label are the ones that maximize $\mathcal{I}$. A heuristic approach is to keep only one a single feature from a correlated set of features that provides similar relevance information, and discard the remaining features. In order to ensure this, the mRMR algorithm minimizes the mutual correlation among the features expressed in terms of redundancy $\mathcal{R}$,

$$\mathcal{R}(\mathbb{X}) = \sum_{x_i, x_j \in \mathbb{X}} \mathcal{I}(x_i, x_j). \tag{3}$$

where $\mathcal{I}$ follows its definition in (2). This procedure yielding maximum $\mathcal{I}$ with respect to the diabetic class, along with minimal $\mathcal{R}$, shortlists a set of ten features that are potentially strong predictors of the future development of T2DM.

E. Classification

We developed a supervised learning scheme using the baseline SAHS dataset and the labels (healthy, T2DM) obtained at the follow-up after an average of 7.5 years. In each experiment, we used a kernel-based binary SVM method to train, test and validate the performance of the diabetes prediction models. We excluded the 44 CVD entries as the only way of defining this class was based upon self-reporting and not on quantitative assessment. Furthermore, we also removed all entries with any information missing. That resulted in a total of 1,492 instances that were used in this study, out of which 171 were from the minority class and 1,321 were majority instances. As shown in Table I, the SAHS dataset is intrinsically unbalanced with the class distribution skewed toward the majority class with a ratio of 7.5:1. We considered the minority class of diabetic subjects as the positive class with a label of 1, whereas the majority class consisting of healthy persons was termed as the negative class marked by a '-1' label. To standardize the feature range prior to training, the feature space was scaled to unit variance around the respective mean for each feature respectively. To ensure that a model was unbiased, robust, and generalized well to the new data, we performed 10-fold cross-validation (CV).

For each CV, we first randomly select a hold-out set consisting of 11 minority and 83 majority instances and then randomly sampled the remaining data into 100 different train and test sets. In each of these 10 attempts we have compared the performances of a linear and non-linear SVM for all 1,023 possible combinations of ten most relevant features by increasing the number of features incrementally from single feature to a combination of all ten features. The optimal hyperplane parameters of the kernel were determined through a grid search. To select the best feature set, we have used the geometric mean of specificity and sensitivity [22]. All experiments were performed by an in-house developed software using Matlab ®(version 9.2.0 MathWorks Inc., Natick, Massachusetts, USA).

III. Results and Discussion

The mRMR algorithm produces a sequential list of ten ranked features, shown in Table II. Besides ethnicity (ranked fourth), all other features are notably derived form the OGTT measurements. The list contains six features derived from blood glucose concentrations, while only three features are deduced from insulin concentrations. In all the classification experiments, we aimed to maximize the ability to correctly predict the diabetic class. The bar plots in Fig. 3 show the g-mean of the sensitivity and specificity obtained from the linear and RBF kernels. For each number of features used, we selected the combination that generated the maximum g-mean. All the results presented here are averaged over 100 iterations of the respective classifiers. The accuracy and specificity of the

same feature combinations are illustrated separately in Fig. 4.

As illustrated in Fig. 3, the g-mean for the linear SVM ranges from 0.7587 to 0.7643, and for non-linear SVM from 0.7461 to 0.7934.

More observations from Fig. 4 A combination of four features, namely AuC-Glu0-120, Glu120-0, Glu120-60 and Glu30-0 provided the best classification performance using a non-linear SVM: g-mean of 0.89, accuracy of 96.8

In the second phase, we further refined the number of variables to four by only selecting the ones which provided the best performance using the SVM classification scheme with the parameters C and γ preconfigured to a value of 1. For this purpose, we employed the accuracy achieved in the validation set as the evaluation criterion. Table III shows the mean validation accuracies of the variables which is obtained by performing 100 iterations of the SVM classifier supplied with only one variable at a time.

IV. Data Experiments

In this paper, we employed the non-linear SVM (**??**) in the form of radial basis functions (RBF) (**??**) since the classes can not be linearly separated directly as observed in Fig. 1. We also used the linear variant of the SVM (**??**) to compare the classifier performances. Moreover, due to the unbalanced nature of the dataset, we conducted two experiments in the preprocessing phase. In the first the dataset was balanced where we randomly undersampled the majority class, and took 160 instances from each class for the training. In the second experiment, we retained the class ratio of the data set and took 1,360 samples to generate the training set that contained 160 and 1,200 instances of the diabetic and healthy classes respectively. In order to ensure that the model remained unbiased and generalized well to new data, we performed 10-fold cross-validation during the training and the performance obtained was averaged over all the 10 folds. All the experiments were carried out using the statistical and machine learning toolbox of Matlab and the data was normalized prior to the training. The optimal hyperplane parameters C and σ in (**??**) and (**??**) respectively were determined through a grid search with a view to maximize the classifier sensitivity defined as,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

where TP and FN refer to the number of correctly and incorrectly classified diabetic subjects respectively.

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

Rank	Feature
1	AuC-Glu$_{0\text{-}120}$
2	ΔGlu$_{120\text{-}0}$
3	ΔGlu$_{120\text{-}60}$
4	Ethnicity
5	ΔIns$_{120\text{-}0}$
6	ΔGlu$_{60\text{-}0}$
7	ΔGlu$_{30\text{-}0}$
8	ΔGlu$_{60\text{-}30}$
9	ΔIns$_{120\text{-}60}$
10	ΔIns$_{60\text{-}0}$

TABLE III: Average performance of the individual features gauged by the accuracy

Features	Mean Accuracy (SD)
AuC-Glu$_{60\text{-}120}$	0.973 (0.013)
PG$_{120}$	0.971 (0.015)
PG$_{60}$	0.967 (0.022)
AuC-Glu$_{0\text{-}120}$	0.958 (0.019)
AuC-Glu$_{30\text{-}120}$	0.950 (0.018)
Sl-Glu$_{60\text{-}0}$	0.946 (0.025)
Sl-Glu$_{120\text{-}0}$	0.931 (0.026)
PG$_{0}$	0.816 (0.039)
Sl-Glu$_{120\text{-}60}$	0.763 (0.050)
Sl-Glu$_{30\text{-}0}$	0.745 (0.044)

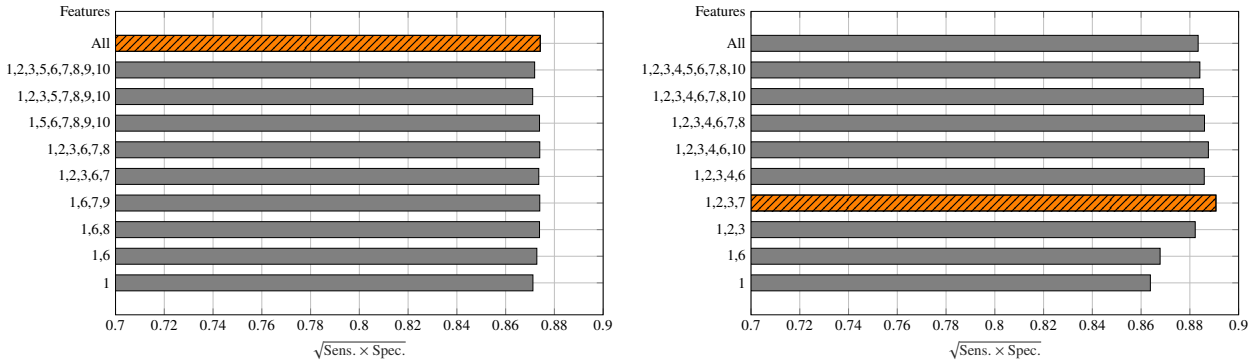
TABLE IV: Mean Training performance of the classifiers

	Accuracy ± SD	Sensitivity ± SD	Specificity ± SD
Linear SVM (Balanced)	78.73 % ± 1.40 %	77.77 % ± 1.20 %	79.69 % ± 2.30 %
Linear SVM (Unbalanced)	79.74 % ± 0.43 %	77.81 % ± 0.90 %	80.00 % ± 0.40 %
SVM-RBF (Balanced)	78.95 % ± 1.90 %	79.07 % ± 2.20 %	78.84 % ± 2.60 %
SVM-RBF (Unbalanced)	78.51 % ± 0.80 %	78.13 % ± 1.30 %	78.57 % ± 1.00 %

V. Results and Discussion

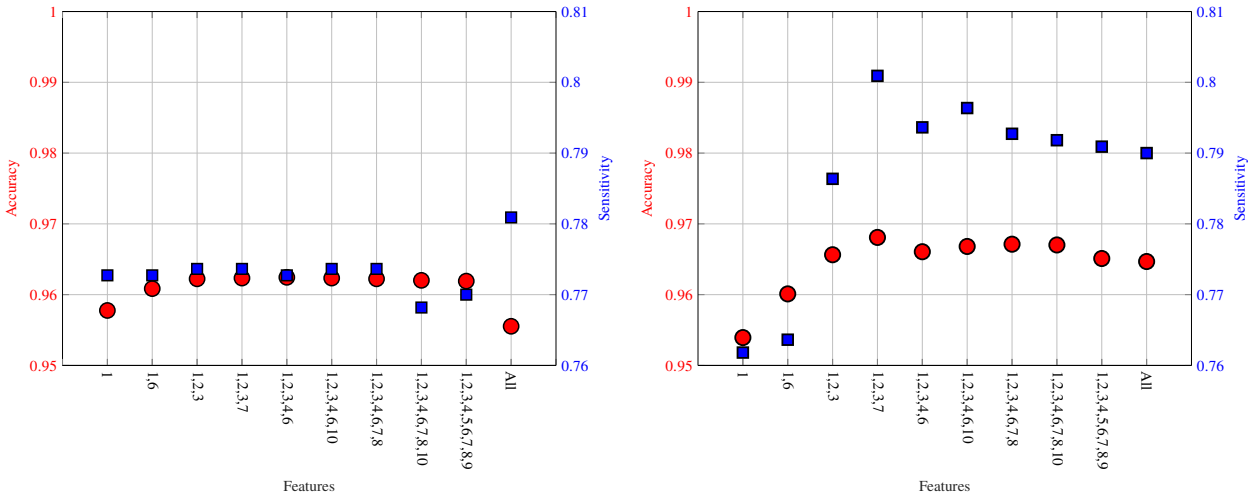
In order to correctly predict the future diabetes subjects, the model was trained to maximize the sensitivity. To train the predictor model, we used four features, all of which were derived from the blood glucose measurements. Table IV presents the mean training performance of the linear and non-linear SVM classifiers obtained over 100 trials. We used the definition of accuracy as the ratio of number of correctly classified subjects to the total number of subjects, whereas the specificity was the ratio of the correctly classified healthy subjects to the total number of healthy subjects. Most notably, the similarity in the performance for balanced and unbalanced training routines demonstrates that the model in the latter case unbiased toward the majority class. The optimal hyperplane parameters corresponding to the Matlab arguments 'BoxConstraint' and 'Gamma', were respectively assigned the values of 1.0 and 5.0. For the linear version of the SVM, an average of 250 and 1,114 support vectors were used to construct the hyperplane for the balanced and unbalanced datasets respectively. On the other hand, the corresponding values were 278 and 1,198 for the nonlinear SVM with the RBF as the kernel. It should be noted that the difference in the dimensionality of the hyperplanes between the two variants of the SVM is not large, which indicates that the discriminating power of the features used.

Table V displays the validation performance of the classifiers. All the model were validated on a hold-out set 100 times set, in which each iteration resulted in a randomly generated set of 11 diabetic and 83 healthy samples, that were not part



(a) Linear (b) RBF

Fig. 3: The geometric mean of sensitivity and specificity for (a) linear, and (b) RBF kernels.



(a) Linear (b) RBF

Fig. 4: The classifier performance in terms of accuracy and sensitivity for the best feature combinations.

TABLE V: Validation performance classifiers

	Accuracy ± SD	Specificity ± SD	Specificity ± SD
Linear SVM (Balanced)	97.29 % ± 1.40 %	76.82 % ± 11.90 %	100 %
Linear SVM (Unbalanced)	97.19 % ± 1.50 %	76.82 % ± 12.70 %	99.89 % ± 0.40 %
SVM-RBF (Balanced)	97.62 % ± 1.50 %	79.64 % ± 12.70 %	100 %
SVM-RBF (Unbalanced)	98.51 % ± 1.40 %	87.27 % ± 11.50 %	100 %
Two-step Approach [18]	77.43 %	77.70 %	77.40 %
SADPM [17]	56.329 %	88.80 %	52.00 %

of the training data. The best mean performance of 98.51 % accuracy and specificity of 87.27 % was obtained from the nonlinear SVM with the RBF kernel. The standard deviation of the two metrics along 100 iterations was 1.40 % and 11.50 % respectively. We also compared the results of our approach with two other techniques that used the SAHS dataset. The logistic regression based SADPM based on

VI. Conclusion

In this paper, we present a most promising set of features that are used to develop a non-linear SVM based future T2DM prediction model. The features were derived from the OGTT data and were augmented by personal information such as age, ethnicity and BMI. Using a feature selection algorithm, we demonstrate that the features deduced from the blood glucose concentrations provide the optimal feature subset and have the strongest predictive power for the future development of T2DM. Moreover, the performance of the presented prediction model is significantly better in terms of accuracy and sensitivity combined, as compared to other T2DM prediction schemes. In order to address the unbalanced nature of the SAHS dataset, we chose the g-mean of sensitivity and specificity as the performance evaluation criteria.

The principal contribution of this study includes a T2DM prediction model based on the features derived only from the blood glucose concentrations measured during an OGTT. The findings of this paper provide a complementary and cost-effective tool for the clinicians to screen individuals that are at an increased risk of developing T2DM in future.

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] J. Tuomilehto, J. Lindström, J. G. Eriksson, T. T. Valle, H. Hämäläinen, P. Ilanne-Parikka, S. Keinänen-Kiukaanniemi, M. Laakso, A. Louheranta, M. Rastas *et al.*, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," *New England Journal of Medicine*, vol. 344, no. 18, pp. 1343–1350, 2001.

[3] D. P. P. R. Group *et al.*, "Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the diabetes prevention program outcomes study," *The Lancet Diabetes & Endocrinology*, vol. 3, no. 11, pp. 866–875, 2015.

[4] D. Noble, R. Mathur, T. Dent, C. Meads, and T. Greenhalgh, "Risk models and scores for type 2 diabetes: systematic review," *BMJ*, vol. 343, p. d7163, 2011.

[5] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes risk calculator," *Diabetes Care*, vol. 31, no. 5, pp. 1040–1045, 2008.

[6] C. Glümer, B. Carstensen, A. Sandbæk, T. Lauritzen, T. Jørgensen, and K. Borch-Johnsen, "A danish diabetes risk score for targeted screening," *Diabetes Care*, vol. 27, no. 3, pp. 727–733, 2004.

[7] M. Heliövaara, A. Aromaa, T. Klaukka, P. Knekt, M. Joukamaa, and O. Impivaara, "Reliability and validity of interview data on chronic diseases the mini-finland health survey," *Journal of Clinical Epidemiology*, vol. 46, no. 2, pp. 181 – 191, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0895435693900567

[8] , "Report of the expert committee on the diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 20, no. 7, pp. 1183–1197, 1997. [Online]. Available: http://care.diabetesjournals.org/content/20/7/1183

[9] M. Stumvoll, A. Mitrakou, W. Pimenta, T. Jenssen, H. Yki-Järvinen, T. Van Haeften, W. Renn, and J. Gerich, "Use of the oral glucose tolerance test to assess insulin release and insulin sensitivity." *Diabetes Care*, vol. 23, no. 3, pp. 295–301, 2000.

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

[11] R. A. DeFronzo and M. Abdul-Ghani, "Assessment and treatment of cardiovascular risk in prediabetes: Impaired glucose tolerance and impaired fasting glucose," *The American Journal of Cardiology*, vol. 108, no. 3, pp. 3B–24B, 2011.

[12] 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.

[13] 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.

[14] 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.

[15] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[16] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 728–734, Mar. 2015.

[17] M. P. Stern, K. Williams, and S. M. Haffner, "Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test?" *Annals of Internal Medicine*, vol. 136, no. 8, pp. 575–581, 2002.

[18] M. A. Abdul-Ghani, T. Abdul-Ghani, M. P. Stern, J. Karavic, T. Tuomi, I. Bo, R. A. DeFronzo, and L. Groop, "Two-Step Approach for the Prediction of Future Type 2 Diabetes Risk," *Diabetes Care*, vol. 34, no. 9, pp. 2108–2112, Sep. 2011.

[19] M. A. Abdul-Ghani, V. Lyssenko, T. Tuomi, R. A. DeFronzo, and L. Groop, "Fasting versus postload plasma glucose concentration and the risk for future type 2 diabetes: results from the botnia study," *Diabetes Care*, vol. 32, no. 2, pp. 281–286, 2009.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[21] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99. New York, NY, USA: ACM, 1999, pp. 155–164. [Online]. Available: http://doi.acm.org/10.1145/312129.312220

[22] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[23] Y. Tang, Y. Zhang, N. V. Chawla, and S. Krasser, "Svms modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, Feb 2009.

[24] 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.

[25] 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.

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

[27] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.

[28] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, NY, USA:, 2001, vol. 1, no. 10.

[29] Y. Seino, M. Ikeda, M. Yawata, and H. Imura, "The insulinogenic index in secondary diabetes," *Hormone and Metabolic Research*, vol. 7, no. 02, pp. 107–115, 1975.

[30] 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.

[31] D. Matthews, J. Hosker, A. Rudenski, B. Naylor, D. Treacher, and R. Turner, "Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man," *Diabetologia*, vol. 28, no. 7, pp. 412–419, 1985.

[32] 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.

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

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.