Predicting long-term Type 2 Diabetes with Support Vector Machine using Oral Glucose Tolerance Test

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

***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 the 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 1,492 healthy individuals collected during the San Antonio Heart Study. This study collected plasma glucose and insulin concentrations before glucose intake and at three time-points thereafter (30, 60 and 120 min). 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 best ranked features were used to generate SVM based prediction models. This research shows that an individual’s abnormal plasma glucose levels, and the information derived therefrom have the strongest predictive performance for the future development of T2DM. Significantly, 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 96.80 % and a specificity of 80.09 % obtained on a holdout set.**

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

# Introduction

The global incidence of diabetes was estimated at 422 million in 2014 and its prevalence among the adult population increased from 4.7 % in 1980 to 8.5 % in 2014 (Mathers and Loncar 2006). In 2015 alone, about 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 (Tuomilehto et al. 2001). The early identification of high risk individuals of diabetes is therefore, essential for targeted prevention strategies (Group and others 2015).

Even though the number of clinical studies aimed at diagnosing diabetes has been growing recently, studies predicting the risk of developing diabetes are limited. This subject has lately received an increased amount of research interest (Noble et al. 2011). 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 (Heikes et al. 2008; Glümer et al. 2004). 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 (Heliövaara et al. 1993). The oral glucose tolerance test OGTT is commonly used to screen diabetes (“Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus” 1997) and to provide a critical understanding of its future evolution (Stumvoll et al. 2000). In an OGTT, the plasma glucose and insulin levels are measured at regular intervals in a 2-hr period after orally administering a standard dose of glucose (Stumvoll et al. 2000). 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 T2DM.

A precursory stage of diabetes, commonly referred to as prediabetes, exists before overt T2DM, and is described by impaired fasting glucose (IFG), along with impaired glucose tolerance (IGT). According to the World Health Organization (WHO) diagnostic criteria, the IFG is defined as fasting plasma glucose level of 100mg/dL to 125mg/dL. The IGT which describes an abnormally raised glucose level is defined as the 2-hour plasma glucose level in the range of 140mg/dL to 199mg/dL, measured during the OGTT (World Health Organization and International Diabetes Federation 2006). Although prediabetes is considered as an intermediate stage in the natural progression of T2DM (DeFronzo and Abdul-Ghani 2011), it has been reported that only 50 % of the subjects diagnosed with IGT developed diabetes within 10 years (Shaw et al. 1999; Unwin et al. 2002). Moreover, long-term population studies have also shown that around 50 % of the diabetic patients did not exhibit IGT at any time prior to the diagnosis (Abdul-Ghani et al. 2007). This suggests that the fasting and 2-hour plasma glucose levels used in and of themselves cannot accurately predict the future development of type 2 diabetes mellitus (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 (Barakat, Bradley, and Barakat 2010). This approach uses SVM along with a rule-based explanation to provide a comprehensibility of the results to the clinicians. The plasma glucose levels at baseline and 2-hr were among the features used. Han et al employed an ensemble support vector machine (SVM) and random forest learning approaches to develop a decision making algorithm for the diagnosis of diabetes (Han et al. 2015). However, investigations that are designed to identify individuals at high risk of developing T2DM in the future are limited. The San Antonio diabetes prediction model (SADPM) (Stern, Williams, and Haffner 2002) 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 (Abdul-Ghani et al. 2007). In another study, multivariate logistic models using the plasma glucose values measured in the OGTT were used to predict the future risk of developing T2DM (Abdul-Ghani et al. 2011, 2009). The predictive power of different biomarkers such as the fasting plasma glucose level, BMI and hemoglobin A1C (HbA1c) for T2DM onset was assessed in (Ozery-Flato et al. 2013). This study focused on individuals with metabolic syndrome, a complex and serious health condition that greatly increases the risk of 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 diagnosis and prediction of a disease 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 (Chawla et al. 2002) 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 (Domingos 1999) in order to introduce a certain bias. Other techniques introduce evaluation metric such as the geometric mean (g-mean) (Kubat, Matwin, and others 1997), that concurrently optimizes the positive class accuracy (sensitivity) and the negative class accuracy (specificity) (Tang et al. 2009).

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) (Burke et al. 1999; Lorenzo et al. 2006).

# Materials and Methods

## 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 (Burke et al. 1999; Lorenzo et al. 2006). 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 individuals at the time of recruitment was between 25 and 64 years. As a part of the data collection, plasma glucose and serum insulin concentrations were collected during the OGTT 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 of 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 mg/dL or 2-hr glucose level 200 mg/dL (World Health Organization and International Diabetes Federation 2006). Furthermore, all individuals taking anti-diabetic medications were also classified as having T2DM. Individuals that reported by themselves any cardiovascular event such as a 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 this 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 [[tab:patients]](#tab:patients) shows the population distribution in terms of the four classes. The distribution in terms of the ethnicity shows the T2DM prevalence among the ma individuals more than double, as compared to the nhw population.

Table [tab:patients] The classification of the 1,492 subjects used in this study based on the ethnicity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 plasma glucose and serum insulin concentrations sampled at the baseline, and at 30, 60 and 120 min thereafter. The individuals are labeled at the SAHS follow-up using the current standard of care (Burke et al. 1999). Figure [[fig:OGTT\_mean]](#fig:ogtt_mean) shows the distributions of the data used in this study.

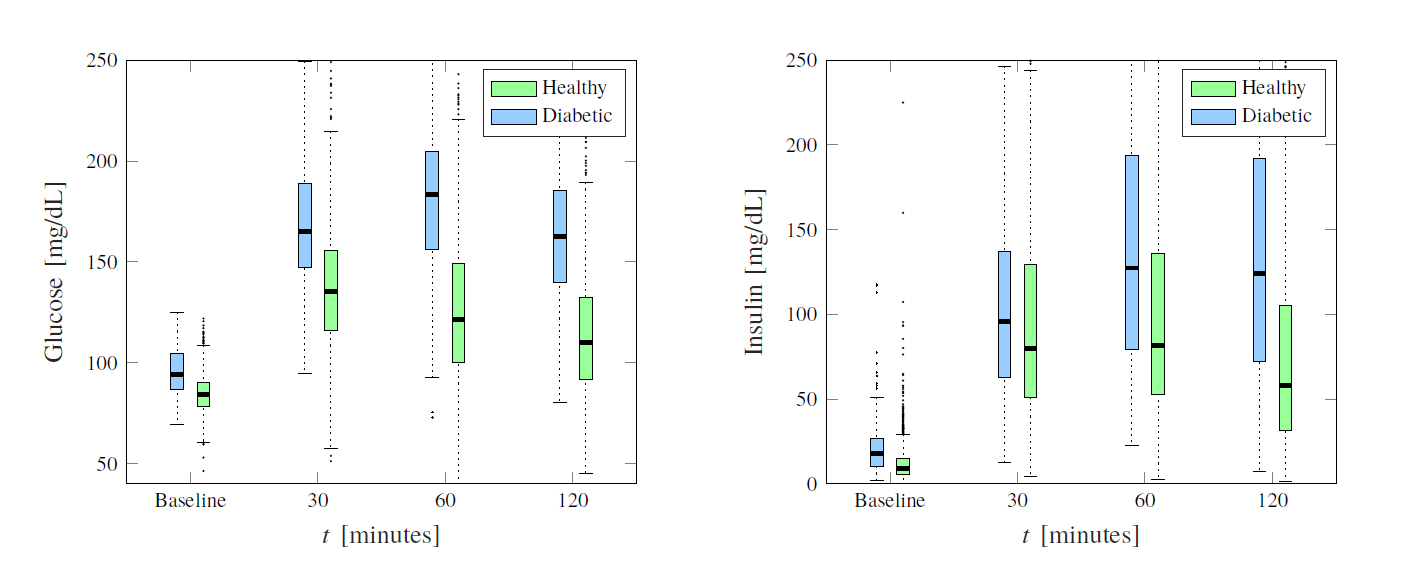


Fig. [[fig:OGTT\_mean]](#fig:ogtt_mean) Box plots of glucose and insulin levels for healthy and diabetic subjects measured at the baseline OGTT.

## Machine Learning Framework

In this paper, we implemented SVM to construct the models for the 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 (Vapnik 2000; Vapnik and Chervonenkis 2015). A model constructed by the SVM minimizes the overlap between 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 the 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 (Friedman, Hastie, and Tibshirani 2001). The 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 and specifying a cost that controls the rigidity of the class margin. This process is normally carried out through a grid search.

## Feature Extraction

We extracted all the features from the SAHS data acquired at the baseline. The dataset consists of plasma glucose and insulin concentrations recorded before glucose intake and at three time-points thereafter (30, 60, and 120min). The labels (healthy and diabetes) were generated at the 7.5 years follow-up using the current standard of care diagnostics (Burke et al. 1999). 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) (Seino et al. 1975), which is a direct measure of the insulin response to glucose. It is calculated as the ratio of the slope of the insulin curve to the slope of the glucose curve between any two time intervals in the OGTT. The second marker, Matsuda index (M) evaluates the insulin sensitivity from the OGTT using a product of the weighted averages of the glucose and insulin concentrations (Matsuda and DeFronzo 1999),

[[eq:Mat\_idx]](#eq:MI)

where the subscripts depict the time point of the OGTT. In case when the value at 90 min is not available, the average of and 90 min is used instead (Matsuda and DeFronzo 1999). The third marker, homeostatic model assessment - insulin resistance (HOMA-IR) (Matthews et al. 1985) evaluates the beta-cell function. It is defined as the product of fasting plasma 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 features (illustrated in Fig. [[fig:features]](#fig:features)) are used in this study. The prefix AuC denotes the area under the curve and the slope is denoted by the symbol . The term Thalf represents the linearly interpolated value between any two intervals.

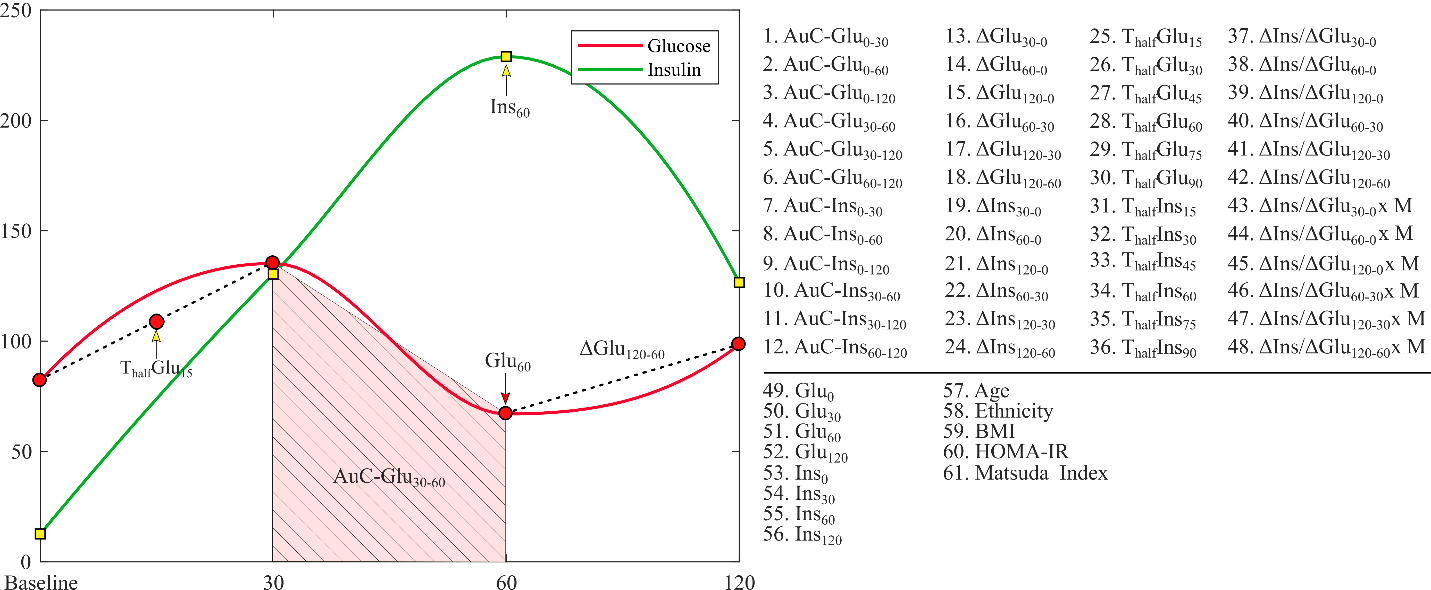


Fig. [[fig:features]](#fig:features): Illustration of all 61 features extracted from the SAHS dataset.

## 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 (Peng, Long, and Ding 2005), which selects the most relevant features with minimum correlation among them. The mRMR algorithm determines the relevance between a feature ( as continuous random variable) and the class label ( as discrete random variable) in terms of the mutual information defined as (Ross 2014),

[[eq:MI]](#eq:MI)

where , and are the probabilities of the random variables and taking a particular value and respectively. The term denotes the joint probability . The three terms in [[eq:MI]](#eq:MI) 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 . 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 ,

[[eq:R]](#eq:MI)

where follows its definition in [[eq:MI]](#eq:MI). This procedure yielding maximum with respect to the diabetic class, along with minimal , shortlists a set of ten features that are potentially strong predictors of the future development of T2DM.

## 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 [[tab:patients]](#tab:patients), 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 selected a hold-out set consisting of 11 minority and 83 majority instances. We evaluated each model 100 times, in which the data was randomly partitioned on each occasion. We compared the performances of linear and non-linear SVM for all possible 1,023 combinations of ten most relevant features by increasing the number of features incrementally from one to ten. 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 (Kubat, Matwin, and others 1997). All experiments were performed by an in-house developed software using Matlab(version 9.2.0 MathWorks Inc., Natick, Massachusetts, USA).

# Results and Discussion

The mRMR algorithm produces a sequential list of ten ranked features, shown in Table [[tab:ranked]](#tab:ranked). Besides ethnicity (ranked fourth), all other features are notably derived from OGTT measurements. The list contains six features derived from plasma glucose concentrations, while only three features are deduced from insulin concentrations.

Table [[tab:ranked]](#tab:ranked)List of ten most relevant features ranked by the mRMR algorithm

|  |  |
| --- | --- |
| Rank | Feature |
| 1 | AuC-G0-120 |
| 2 | G120-0 |
| 3 | G120-60 |
| 4 | Ethnicity |
| 5 | I120-0 |
| 6 | G60-0 |
| 7 | G30-0 |
| 8 | G60-30 |
| 9 | I120-60 |
| 10 | I60-0 |

[tab:ranked]

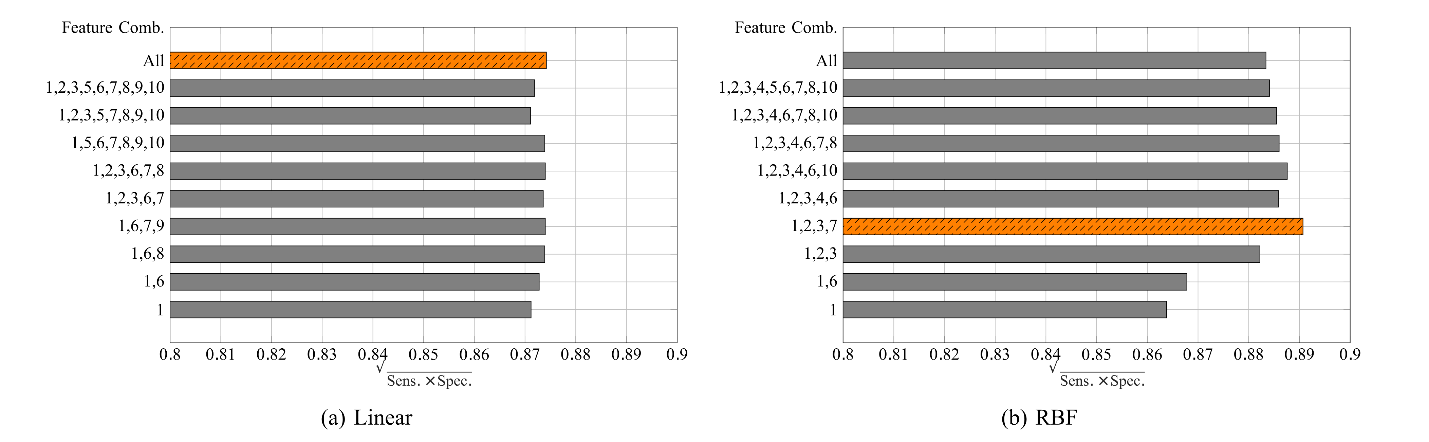


Fig. [fig:sn\_acc] The g-mean of sensitivity and specificity for (a) linear, and (b) RBF kernels. The maximum performance feature combination is depicted by a different color scheme.

In all the classification experiments, we aimed to maximize the ability to correctly predict the diabetic class without compromising the classifier accuracy. The bar plots in Fig. [[fig:gmean]](#fig:gmean) 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 iterations of the respective classifiers. The g-mean obtained from the linear SVM ranges from 0.8711 to 0.8742. As observed from Fig. [[fig:gmean\_lin\_unbal]](#fig:gmean_lin_unbal), the addition of more features does not result in a substantial performance improvement. However, the maximum g-mean of the sensitivity and specificity is obtained when all features are used. For the non-linear SVM with RBF kernel, the g-mean ranges from 0.8638 to 0.8903. The combination of the features namely, AuC-Glu0-120, G120-0, G120-60 and G30-0 yields the maximum performance. Notably, all four features are derived from the plasma glucose concentrations. We note that the glucose derived features are ranked the highest during feature selection. Moreover, a combination of glucose only features generate the best SVM models when less than four features are used.

Table [[tab: validation]](#tab:ranked) Comparison of validation performance of the best SVM classifiers with previous studies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy  SD | Specificity  SD | Specificity  SD | g-mean  SD |
| Linear SVM (10 features) | 95.55 0.24 % | 78.09 0.33  % | 97.87 0.3 % | 0.8742 ± 0.2100 |
| SVM-RBF (4 features) | 96.80 0.41 % | 80.09 1.42 % | 99.02 0.33 % | 0.8903 ± 1.5600 |
| SADPM (Stern, Williams, and Haffner 2002) | 56.329 % | 88.80 % | 52.00 % | 0.6795 |
| Two-step Approach (Abdul-Ghani et al. 2011) | 77.43 % | 77.70 % | 77.40 % | 0.7755 |

[tab:validation]

The accuracy and sensitivity of the same feature combinations are separately shown in Fig. [[fig:sn\_acc]](#fig:sn_acc). The best model obtained using a combination of four glucose derived features and RBF kernel has an accuracy of 96.80 %, and sensitivity of 80.09 %. It is worth mentioning that during all our classification experiments, maximum sensitivity of 81.27 % was obtained using only one ethnicity feature. However, the accuracy dropped to 41.33 %.

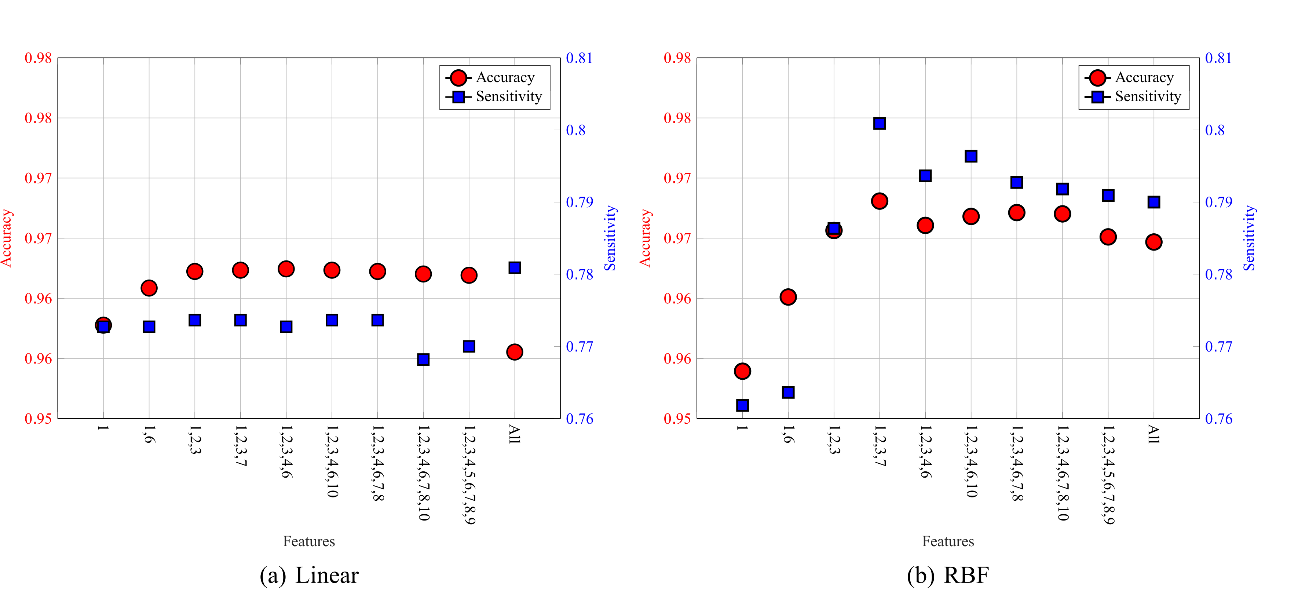


Fig. [[fig:sn\_acc]](#fig:sn_acc): The classifier performance in terms of accuracy and sensitivity for the best feature combinations.

Table [[tab:validation]](#tab:validation) presents a comparison of the generated SVM models to the results obtained in other studies using the SAHS dataset. We compared our results with the SADPM (Stern, Williams, and Haffner 2002), in which a person’s age, gender, ethnicity, fasting glucose level, family history, blood pressure, and cholesterol level were used to construct a logistic regression. It is notable that the SADPM has the highest sensitivity (88.80 %) however, the increased prediction performance comes along with a very low accuracy of 56.33 %. In (Abdul-Ghani et al. 2011), a two-step approach was introduced that first used the SADPM risk score and then augmented it with the 1-hour plasma glucose concentration measured in the OGTT. This strategy resulted in an improved accuracy but the sensitivity dropped to 77.70 %.

In the SAHS dataset, the prevalence of IFG and IGT was 8.91 % (133 instances) and 22.52 % (336 instances) respectively. Out of the 399 subjects diagnosed with prediabetes showing IFG or IGT at the baseline, only 120 (30.08 %) actually developed diabetes between the baseline and the follow-up. Furthermore, 120 (25.67 %) subjects diagnosed with diabetes at the follow-up did not show any symptoms of either IGT or IFG at the baseline.

Our investigation shows that features derived from insulin have less predictive value for T2DM as compared to glucose based features. Indices such as Matsuda and homair that are commonly used to assess the insulin function, also did not yield high correlation with the future development of T2DM.

# 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 plasma 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 plasma 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 the 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.

Abdul-Ghani, M. A., T. Abdul-Ghani, M. P. Stern, J. Karavic, T. Tuomi, I. Bo, R. A. DeFronzo, and L. Groop. 2011. “Two-Step Approach for the Prediction of Future Type 2 Diabetes Risk.” *Diabetes Care* 34 (9): 2108–12.

Abdul-Ghani, M. A., K. Williams, R. A. DeFronzo, and M. Stern. 2007. “What Is the Best Predictor of Future Type 2 Diabetes?” *Diabetes Care* 30 (6): 1544–8.

Abdul-Ghani, Muhammad A, Valeriya Lyssenko, Tiinamaija Tuomi, Ralph A DeFronzo, and Leif Groop. 2009. “Fasting Versus Postload Plasma Glucose Concentration and the Risk for Future Type 2 Diabetes: Results from the Botnia Study.” *Diabetes Care* 32 (2): 281–86.

Barakat, Nahla, Andrew P Bradley, and Mohamed Nabil H Barakat. 2010. “Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus.” *IEEE Transactions on Information Technology in Biomedicine* 14 (4): 1114–20.

Burke, James P., Ken Williams, Sharon P. Gaskill, Helen P. Hazuda, Steven M. Haffner, and Michael P. Stern. 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.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57.

DeFronzo, Ralph A, and Muhammad Abdul-Ghani. 2011. “Assessment and Treatment of Cardiovascular Risk in Prediabetes: Impaired Glucose Tolerance and Impaired Fasting Glucose.” *The American Journal of Cardiology* 108 (3): 3B–24B.

Domingos, Pedro. 1999. “MetaCost: A General Method for Making Classifiers Cost-Sensitive.” In *Proceedings of the Fifth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 155–64. KDD ’99. New York, NY, USA: ACM.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA:

Glümer, Charlotte, Bendix Carstensen, Annelli Sandbæk, Torsten Lauritzen, Torben Jørgensen, and Knut Borch-Johnsen. 2004. “A Danish Diabetes Risk Score for Targeted Screening.” *Diabetes Care* 27 (3): 727–33.

Group, Diabetes Prevention Program Research, and others. 2015. “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* 3 (11): 866–75.

Han, Longfei, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen. 2015. “Rule Extraction from Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes.” *IEEE Journal of Biomedical and Health Informatics* 19 (2): 728–34.

Heikes, Kenneth E., David M. Eddy, Bhakti Arondekar, and Leonard Schlessinger. 2008. “Diabetes Risk Calculator.” *Diabetes Care* 31 (5): 1040–5.

Heliövaara, M., A. Aromaa, T. Klaukka, P. Knekt, M. Joukamaa, and O. Impivaara. 1993. “Reliability and Validity of Interview Data on Chronic Diseases the Mini-Finland Health Survey.” *Journal of Clinical Epidemiology* 46 (2): 181–91.

Kubat, Miroslav, Stan Matwin, and others. 1997. “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection.” In *ICML*, 97:179–86. Nashville, USA.

Lorenzo, C., K. Williams, K. J. Hunt, and S. M. Haffner. 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): 625–30.

Mathers, Colin D, and Dejan Loncar. 2006. “Projections of Global Mortality and Burden of Disease from 2002 to 2030.” Edited by Jon Samet. *PLoS Medicine* 3 (11): e442.

Matsuda, M, and R A DeFronzo. 1999. “Insulin Sensitivity Indices Obtained from Oral Glucose Tolerance Testing: Comparison with the Euglycemic Insulin Clamp.” *Diabetes Care* 22 (9): 1462–70.

Matthews, DR, JP Hosker, AS Rudenski, BA Naylor, DF Treacher, and RC Turner. 1985. “Homeostasis Model Assessment: Insulin Resistance and -Cell Function from Fasting Plasma Glucose and Insulin Concentrations in Man.” *Diabetologia* 28 (7): 412–19.

Noble, Douglas, Rohini Mathur, Tom Dent, Catherine Meads, and Trisha Greenhalgh. 2011. “Risk Models and Scores for Type 2 Diabetes: Systematic Review.” *BMJ* 343: d7163.

Ozery-Flato, Michal, Naama Parush, Tal El-Hay, Žydrūnė Visockienė, Ligita Ryliškytė, Jolita Badarienė, Svetlana Solovjova, Milda Kovaitė, Rokas Navickas, and Aleksandras Laucevičius. 2013. “Predictive Models for Type 2 Diabetes Onset in Middle-Aged Subjects with the Metabolic Syndrome.” *Diabetology & Metabolic Syndrome* 5 (1): 36.

Peng, H., F. Long, and C. Ding. 2005. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1226–38.

“Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus.” 1997. *Diabetes Care* 20 (7): 1183–97.

Ross, Brian C. 2014. “Mutual Information Between Discrete and Continuous Data Sets.” *PloS One* 9 (2): e87357.

Seino, Y, M Ikeda, M Yawata, and H Imura. 1975. “The Insulinogenic Index in Secondary Diabetes.” *Hormone and Metabolic Research* 7 (02): 107–15.

Shaw, J. E., P. Z. Zimmet, M. de Courten, G. K. Dowse, P. Chitson, H. Gareeboo, F. Hemraj, D. Fareed, J. Tuomilehto, and K. G. Alberti. 1999. “Impaired Fasting Glucose or Impaired Glucose Tolerance. What Best Predicts Future Diabetes in Mauritius?” *Diabetes Care* 22 (3): 399–402.

Stern, Michael P, Ken Williams, and Steven M Haffner. 2002. “Identification of Persons at High Risk for Type 2 Diabetes Mellitus: Do We Need the Oral Glucose Tolerance Test?” *Annals of Internal Medicine* 136 (8): 575–81.

Stumvoll, Michael, Asimina Mitrakou, Walkyria Pimenta, Trond Jenssen, HANNELE Yki-Järvinen, Timon Van Haeften, WPHD Renn, and John Gerich. 2000. “Use of the Oral Glucose Tolerance Test to Assess Insulin Release and Insulin Sensitivity.” *Diabetes Care* 23 (3): 295–301.

Tang, Y., Y. Zhang, N. V. Chawla, and S. Krasser. 2009. “SVMs Modeling for Highly Imbalanced Classification.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39 (1): 281–88.

Tuomilehto, Jaakko, Jaana Lindström, Johan G Eriksson, Timo T Valle, Helena Hämäläinen, Pirjo Ilanne-Parikka, Sirkka Keinänen-Kiukaanniemi, et al. 2001. “Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle Among Subjects with Impaired Glucose Tolerance.” *New England Journal of Medicine* 344 (18): 1343–50.

Unwin, N., J. Shaw, P. Zimmet, and K. G. M. M. Alberti. 2002. “Impaired Glucose Tolerance and Impaired Fasting Glycaemia: The Current Status on Definition and Intervention.” *Diabetic Medicine* 19 (9): 708–23.

Vapnik, Vladimir Naumovich. 2000. *The Nature of Statistical Learning Theory*. 2nd ed. Statistics for Engineering and Information Science. New York: Springer.

Vapnik, Vladimir N, and A Ya Chervonenkis. 2015. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities.” In *Measures of Complexity*, 11–30. Springer.

World Health Organization, and International Diabetes Federation. 2006. “Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia: Report of a WHO/IDF Consultation.”