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

November 15, 2018

**Abstract**

Diabetes is a large healthcare burden worldwide. With substantial evidence that lifestyle modifications and drug intervention can prevent diabetes, the early identification of high risk individuals is important to design targeted prevention strategies. We present an automatic tool that uses machine learning techniques to predict development of type 2 diabetes mellitus. 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 (i.e. age, ethnicity and body-mass index) of 1,496 healthy individuals from the second cohort of San Antonio Heart Study (SAHS). This study collected blood glucose and insulin concentrations before a standardized glucose intake and at three time-points thereafter (30, 60, and 120 minutes). In addition to this baseline data collection, SAHS also collects the same data at an average follow-up of 7.5 years. This study uses 11 baseline blood measurements and class labels extracted from follow-up data using standard of care for diagnostics of type 2 diabetes mellitus (T2DM). From the available data, we have used a total of 61 features and ranked the top ten features using Minimum Redundancy Maximum Relevance feature selection algorithm. All possible combinations of the 10 best ranked features were used to generate support vector machine (SVM) based T2DM prediction models with hold-out This study shows that the features extracted from blood glucose levels have the strongest predictive performance. Conversely, the insulin concentrations and demographic features do not provide additional performance improvement for diabetes prediction. The results of this work identify the parsimonious clinical data that need to be collected for efficient diabetes prediction.

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

# 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 the 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 risk of developing diabetes are limited. Nevertheless, 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 sociodemographic characteristics such as age, ethnicity, body mass index (BMI), and genealogical information collected through population studies [5, 6]. Due to a non-quantitative character of the sociodemographic data, the predictions techniques may produce misleading results. 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 [\cite{HELIOVAARA1993181}]. Oral glucose tolerance test (OGTT) is used to diagnose type 2 diabetes mellitus (T2DM) [REF] and to provide a critical understanding of its future development [REF]. concentrationsou The glucose tolerance and insulin resistance are two parameters deduced from 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 often described by an impaired glucose tolerance (IGT). According to the World Health Organization 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 140-200 mg/dL, measured during the OGTT [\cite{organization\_definition\_2006}]. Although prediabetes is considered as an intermediate stage in the natural progression of T2DM [\cite{defronzo2011assessment}], it has been reported that only 50 % of the subjects diagnosed with IGT developed diabetes within 10 years [7, 8]. 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 [9]. This suggests that the fasting and 2-hour blood glucose levels cannot accurately predict the future development of T2DM.

Machine learning (ML) has been proposed as a viable instrument for diabetes screening [REF]. 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 [13]. This approach uses support vector machine (SVM) along with a rule-based explanation to provide a comprehensibility of the results to the clinicians. Han et al employed an ensemble SVM and random forest learning approaches to develop a decision making algorithm for the diagnosis of diabetes [14]. However, investigations that are designed to predict development of T2DM are limited. San Antonio diabetes prediction model (SADPM) [15] uses a logistic regression supported by physiological parameters such as systolic blood pressure and cholesterol level. The underlying causes of T2DM in the form of insulin resistance and insulin secretion were studied to develop a prediction model in [9]. 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 [10, 16].

The standard machine learning algorithms are designed to yield optimal performance in terms of accuracy over the full dataset. However, a medical application such as disease diagnosis and prediction require a biased decision-making mechanism that favors one of the classes. This approach would inherently maximize the performance for a clinically more relevant class. Therefore, the objective in such applications is to design a classifier that improves the accuracy of clinically relevant class. Additionally, the amount of clinically relevant data is often an outsized minority. There are various ways to boost classifier performance that include sampling. This produce an artificially balanced class distribution by either under-sampling the majority class, over-sampling the minority class, or a carefully weighted combination of under-sampling and over-sampling [\cite{chawla2002smote}]. To introduce a certain bias, methods such as feature weighting schemes, assign distinct costs to training examples [\cite{Domingos}]. Other techniques do not balance the data but introduce evaluation metric such as the geometric mean (G-mean) [\cite{kubat1997addressing}], that optimize the positive class accuracy (sensitivity) and the negative class accuracy (specificity) at the same time [\cite{Tang\_SVM}].

We hypothesized that the features extracted from the OGTT will be able to predict future onset of T2DM. In this paper, we therefore propose an approach that identifies the most relevant features extracted from the OGTT data that strongly correlate with the onset of future 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) [17, 18]. The following aspects set this work apart from previous studies:

1. Unlike other T2DM prediction studies that defined the overall accuracy as the classifier evaluation criteria, we use the G-mean metric that optimizes the sensitivity while addressing the imbalanced nature of the dataset.
2. A set of features is deduced from the OGTT data that yields the best future T2DM prediction performance. Significantly, the set only contains features derived from the glucose measurements.

# 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 [16], [17]. In total, 5,158 men and non-pregnant women of Mexican-American (MA) and non-Hispanic white (NHW) residents of San Antonio (Texas, USA) 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 sampled during the oral glucose tolerance test at the baseline and a 7.5 years follow-up.. The body mass index (BMI) at baseline was also recorded for each individual. In this study, we analyzed the data of the 1,492 subjects from the second cohort of the SAHS.

T2DM was diagnosed at the follow-up using a World Health Organization criteria, i.e. fasting glucose level ≥126 mg/dL or 2-hour glucose level ≥200 mg/dL [19]. 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 cardiovascular disease (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 1 shows the population distribution. The distribution in terms of the ethnicity shows that the prevalence of T2DM among the MA individuals more than double than the NHW population.

Table 1: 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 % | * 1. % |

### The data used in this study consists of the blood glucose and insulin concentrations sampled at baseline, and at 30, 60 and 120 minute thereafter. The individuals are labeled at the SAHS follow-up using the current standard of care [\cite{rapid\_rise\_1984\_stern}]. Fig. 1 shows the distributions of the data used in this study.

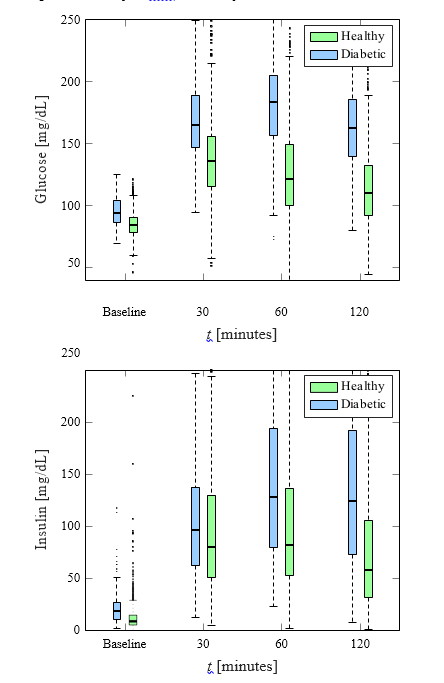


Figure 1: Box plots of glucose and insulin measurements for healthy and diabetic subjects.

### Machine Learning Framework

In this study, we implemented SVM to construct the models for prediction of a future T2DM onset. 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 [22,23]. A model constructed by the SVM minimizes the overlap between the classes in the training set by optimizing the separation 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 and a linear boundary can then be determined in the transformed space. This approach generally provides a better training performance but, with the increase of the dimensionality of the input feature space, potentially increases computational complexity excessively [24]. Introduction of a kernel calculating the inner product between the coordinates of the input feature space, alleviates the need to determine the transformation. In this paper, we used the Gaussian radial basis function as the kernel. During the training, we tuned the values of the parameters *C* and γ through a grid search to obtain the optimal performance of the SVM.

### 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 minutes). The labels (healthy and diabetes) were generated at the 7.5 years follow-up using the current standard of care diagnostics [\cite{rapid\_rise\_1984\_stern}]. From the glucose and insulin measurements, we computed a total of 48 features: slope and area under the curve (see Fig. 1) . 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) [\cite{seino1975insulinogenic}], 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 time intervals in the OGTT. The second marker, Matsuda index [21], evaluates the insulin sensitivity using the OGTT data. The third marker, homeostatic model, assesses t insulin resistance (HOMA-IR) [\cite{matthews1985homeostasis}] by 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 shown a good efficacy for diabetes prediction in previous studies [9, 12]. A total of 61 features (illustrated in Fig. 2) are used in this study: i.e. 8 glucose and insulin concentrations, 3 socio-demographic features, 48 deduced features, and 3 features currently used to predict the future onset of T2DM.

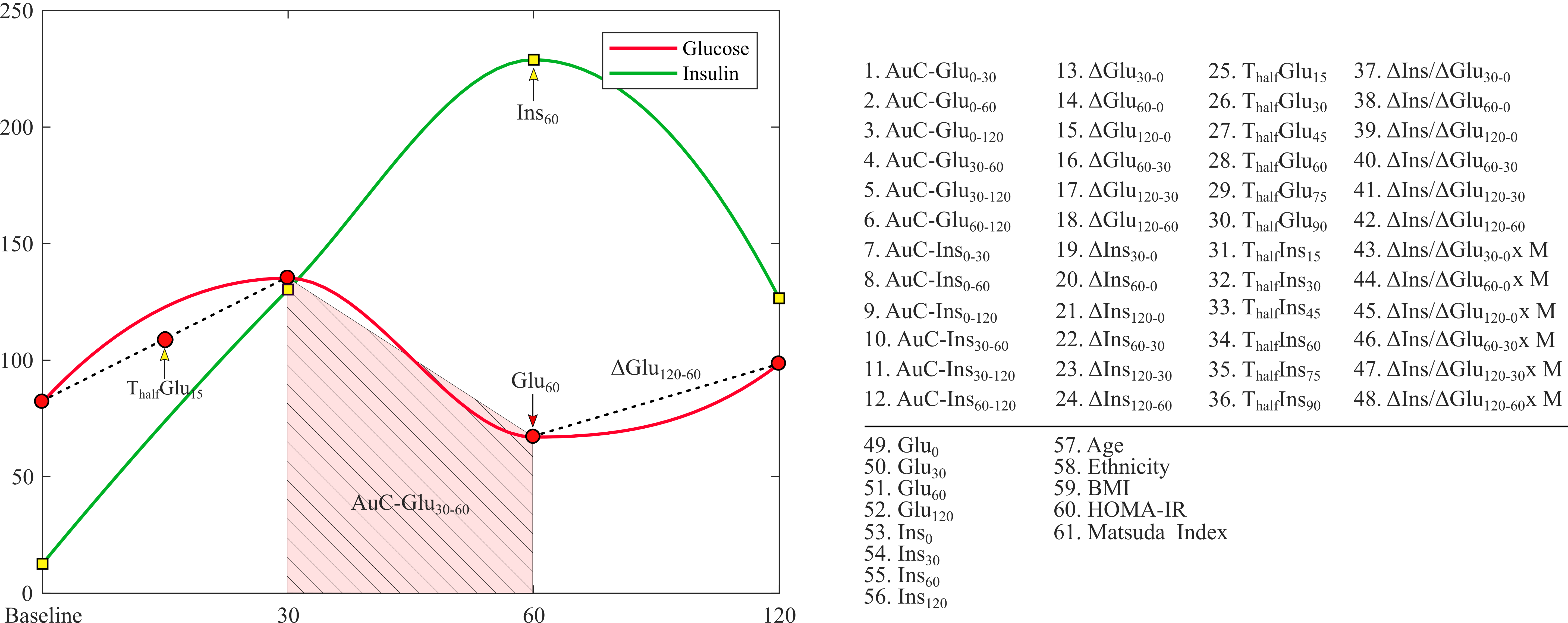
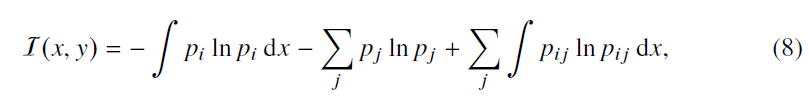
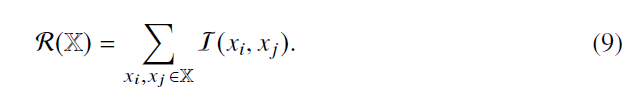


Figure 2: List of all 61 features extracted from the SAHS dataset.

### Feature Selection

Before constructing the SVM model to predict a future diabetes event, we search for the most effective subset of features in terms of relevance to the classifier output, i.e. incidence of diabetes at the follow-up. As a first step, we selected the ten most relevant, from the 61 available, features using the minimal-redundancy- maximal-relevance (mRMR) algorithm [28], which selects the most relevant features with minimum correlation. 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 as [25]:



where pi , and pj are the probabilities of the random variables x and y taking a particular value xi and yj ∈ (−1, 1) ∀ j respectively. The term pij denotes the joint probability, P{x = xi, y = yj}, between two random variables. The three terms in (8) represent the continuous, discrete and joint entropies of the random variables in the respective order. The features most relevant to the class label are the ones that maximize I keeping only one 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 R,

where I follows its definition in (8). This procedure, yielding maximum I with respect to the diabetic class along with minimal R, 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 yielded 1492 instances with used in this study: 171 minority and 1321 majority instances. 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) to train and test the model. For each CV, we first randomly sample a hold-out set consisting of 11 minority and 83 majority samples and 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 1023 possible combinations of ten most relevant features by increasing all tenThe 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 [\cite{ kubat1997addressing}]. All experiments were performed by an in-house developed software using Matlab® (version 9.2.0 MathWorks Inc., Natick, Massachusetts, USA).