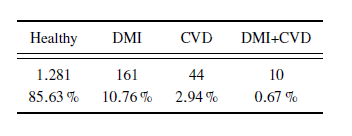
*A. San Antonio Heart Study*

The San Antonio Heart Study (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]. It includes 5,158 men and non-pregnant women of Mexican-American and non-Hispanic white residents of San Antonio, Texas. 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 levels were recorded during the oral glucose tolerance test (OGTT), which measures the body response to a standard 75 g dose of glucose after fasting overnight.

The OGTT was performed both at the baseline and at an average follow-up of 7.5 years. The participants were recruited in 2 cohorts, the first during the period 1979 to 1982, and the second from 1984 to 1988 [18]. The reassessment during the follow-up period took place during the years 1987 to 1990 for the first cohort, and 1991 to 1996 for the second cohort. In this study we use data from the second cohort with the plasma glucose and insulin levels of 1,496 healthy individuals recorded during the OGTT prior to in taking a standard dose of glucose and at 30, 60 and 120 minutes thereafter. Moreover, the socio-demographic information such as age, ethnicity and body-mass index (BMI) was also collected. Table 1 shows the distribution of the annotated classes at the follow-up.

Table 1. The distribution of the classes in the 1,496 individuals labeled at an average follow-up period of 7.5 years. T2DM: Type 2 diabetes mellitus, CVD: Cardiovascular disease, T2DM&CVD: Type 2 diabetes mellitus and CVD. Healthy: all the individuals that are not diagnosed with T2DM or with CVD.



T2DM was diagnosed by a WHO criteria, defining fasting glucose level ≥126 mg/dL or 2-hour glucose level ≥200 mg/dL [19]. Furthermore, all individuals taking anti-diabetic medications was also classified as diabetic. All individuals with self-reported cardiovascular event at follow-up, such as heart attack, stroke or angina, were labeled as CVD. Individuals without T2DM or self-reported CVD were labeled as healthy. During the course of the study, a total of 171 individuals developed T2DM with 10 individuals also reported at least one cardiovascular event. The incidence rate of T2DM for this population was 11.4%. Figure 1 illustrates the blood glucose and insulin values recorded at the baseline, and at 30, 60 and 120 minute during the OGTT.

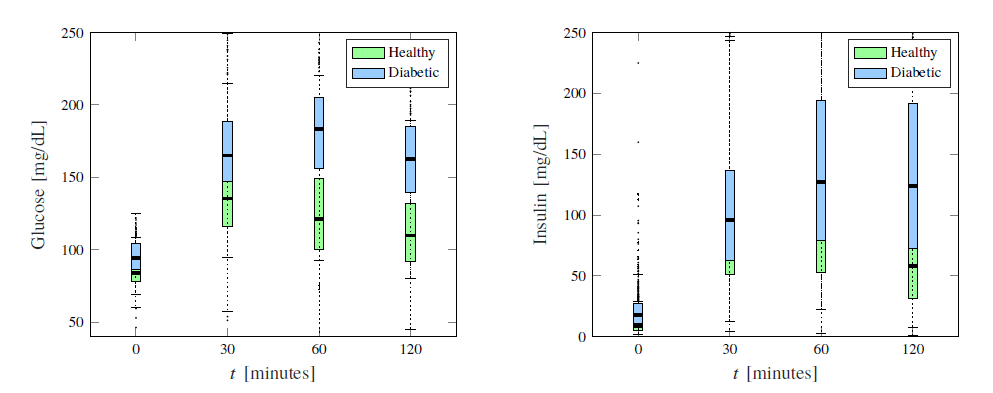


Figure 1. Box plots of glucose and insulin measurements for healthy and diabetic subjects

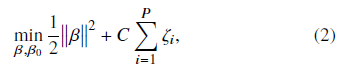
*B. Machine Learning Framework*

The machine learning (ML) framework used in this study is based upon support vector machine (SVM) utilizing the principle of structural risk minimization. The ultimate goal is to develop a model using the training data such that it generalizes well to new datasets and minimizes the empirical risk associated with misclassification of samples in the training set [21], [22]. For a binary classification problem, the model constructed by the SVM finds a decision boundary or a separating hyperplane aiming to minimize the overlapping between the two classes in the training set. SVM is particularly useful in the case of two classes that are not amenable to linear separation in which the feature space is first transformed by a kernel, and a linear boundary is then defined in this new space. This approach is not without potential problems although in general, it gives better training performance [23].

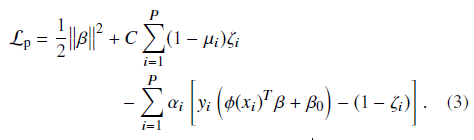
To formulate this approach, let us consider a training data ¹x1; y1º,¹x2; y3º,...,¹xk; yk º of k pairs containing xi 2 RN features and the binary classes yi 2 􀀀1; 1. The SVM approach transforms the input features using a nonlinear mapping \_ : x 7! \_¹xº into a higher dimension space RP, where in general P \_ N. Due to the transformation, the classes can then be separated using a linear decision boundary in the enlarged space. The non-linear SVM classifier F is expressed in terms of the higher dimensional hyperplane,



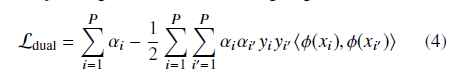
When the classes may not be completely separable, introducing a slack variable \_ in the higher dimension space RP is a common practice which allows for the classifier output in (1) to be on the incorrect side of the margin. Therefore, in order to find the optimal separating hyperplane that maximizes the distance M from the boundary for all the points, and bounds the value of Í i \_i, and in turn misclassification rate, we introduce the convex optimization problem,



with the nonlinear constraints yi¹\_¹xiºT \_+\_0º \_ 1􀀀\_i 8i and \_i \_ 0, and the coefficient C is termed as the cost parameter which decides the rigidity of the margin of the classifier. The solution of (2) can be computed using the Lagrange primal objective function [23],



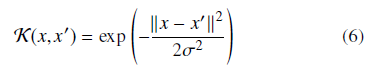
By minimizing Lprimal with respect to \_, \_0, and \_i , we get the corresponding dual form of the Lagrange function,



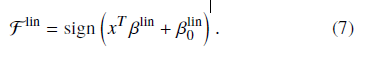
subject to 0 ≤ αi ≤ C and i \_i yi = 0 and the constraints, \_i; \_i \_ 08i. The nonzero coefficients \_i and \_0 are determined using (1). As the dimension of the input feature space goes up, the computation of the mapping \_ gets excessive in complexity. With the introduction of a kernel (5)



we can compute the inner product on (4) that alleviates the need to compute the mapping \_ [24], which becomes computationally expensive as the dimension of the input feature space increased. In this paper, we used the Gaussian radial basis function,



as the kernel where \_ is a free parameter. During the training, we tuned the values of the parameters C and through a grid search to obtain the optimal performance of the SVM. For a linear variant of the SVM, the classifier can expressed without any coordinate mappings,



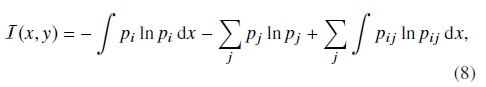
C. Experimental setup

*Feature extraction*

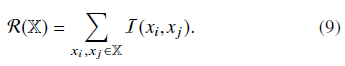
From the glucose and insulin measurements collected at the baseline, we computed the slope and area under the curve between all the possible combinations of a pair of readings. In addition, we also calculated parameters such as the insulinogenic (ratio of insulin and glucose slopes between any two time intervals) and Matsuda indices, as defined in [8], [20]. These variables have shown a good efficacy of diabetes prediction in previous studies [5], [8], since they are used to quantify the amount of insulin required by the body to maintain healthy glucose levels. In total we extracted 68 features.

*Feature Selection*

Before constructing the SVM model to predict a future diabetes event, we aim to find the most effective subset of the features in terms of the relevance to the classifier output. This process greatly reduces the computational cost during the model development by reducing the feature space dimension and also dispense useful scientific insight into the classification problem. We performed a two-step feature selection, where first ten features that correlated the most to the classifier target class were shortlisted. In the second step, we ranked the shortlisted features by evaluating the accuracy through SVM classification and selected only the four best features. In order to define the relevance between the feature and the class labels, consider a feature x in the input feature space RN as a continuous random variable and the class label y as a discrete random variable. Their relationship can be described in terms of the mutual information, defined as [25]:



where pi , and pj are the probabilities of the random variables x and y taking a particular value xi and yj 2 ¹􀀀1; 1º 8j respectively. The term pi j denotes the joint probability Pfx = xi; y = yj g. The three terms in (8) 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 individually yield the maximum I. However, a drawback of pursuing this approach is that the selected features may be mutually correlated, and having a redundant list of shortlisted features only adds to the computational cost of the classifier without necessarily improving its performance. Even more so, the addition of extra features commonly result in the deterioration of the classifier performance [26]. Therefore, an instinctive way forward is to keep only one feature from a correlated set of features that provides similar relevance information, and discard the remaining features from the set X. We follow the minimal redundancy-maximal-relevance (mRMR) algorithm [27], that selects the features, that not only yield the maximal mutual information (8) with respect to the class label, but also minimizes the mutual correlation among the features expressed in terms of redundancy R as:



where I follows its definition in (8). By minimizing R, the mRMR framework selects a set of mutually exclusive features that are most relevant to the class label. Here, we first shortlist a set of ten features that are strong predictors of the future development of type 2 diabetes, on the basis of yielding maximum I with respect to the diabetic class.

*Classification experiments*

We developed a supervised learning scheme using SAHS data and the labels (healthy, diabetic) obtained at the follow-up. The SVM was used for development of the models for diabetes prediction. As observed in Table I, the SAHS dataset is intrinsically imbalanced with the class distribution skewed toward the majority (healthy) class with a ratio of 7.5:1. We have excluded the 44 CVD entries for potentially wakening the labels as the only way of defining this class was based upon the self-reporting and not on quantitative assessment. for the classification problem and removed the rows that had missing entries or contained zero or infinite values. The dataset was then partitioned into training and validation sets. The minority (diabetic) class was defined as the positive class with a label of 1, whereas the majority (healthy) class was defined as the negative class marked by a label -1. In order to ensure that the model remained unbiased, robust, and generalize well to new data, we performed 10-fold cross-validation during the training, while the training performance was averaged over all 10-folds. All the experiments were carried out using the statistical and machine learning toolbox of Matlab (v9.2.0.556344) and the data was normalized prior to the training.

We have compared the performances of a linear and non-linear SVM. In both cases, the optimal hyperplane parameters defined in (2) and (6) were assessed through a grid search with a view to maximize the sensitivity. To probe the influence of data set balancing on the classification performance we have defined the following two experiments:

Experiment I: The dataset was balanced with random under-sampling of the majority class, and took 160 instances from each class for the training.

Experiment II: The dataset retained the unbalanced class distribution and randomly sampled 1,200 (out of 1281 instances) to generate the training set that contained 160 and randomly samples 1,200 instances of the diabetic and healthy classes respectively.