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 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, demographic 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 a support vector machine 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 that need to be collected for efficient diabetes prediction. Our approach show an average accuracy of 96.46% and a specificity of 80.45% 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 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. 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 [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 [\cite{HELIOVAARA1993181}]. Oral glucose tolerance test (OGTT) is used to diagnose diabetes [REF] and to provide a critical understanding of its future evolution [REF]. The glucose tolerance and insulin resistance are two parameters deduced from 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 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 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 [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. The blood glucose levels at baseline and 2h 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 [14]. However, investigations that are designed to identify individuals at high risk of developing T2DM in future 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 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 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, often the amount of data is highly skewed with the clinically relevant class in a outsized minority. There are various roundabout ways to obtain accurate classifier performance in this scenario that include the method of sampling [\cite{chawla2002smote}] in which the class distribution is artificially balanced by either under sampling the majority class, over-sampling the minority class of both. To introduce a certain bias, methods such as feature weighting schemes, assign distinct costs to training examples [\cite{Domingos}]. Other techniques 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 diabetes. 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 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 in terms of the four classes. 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.

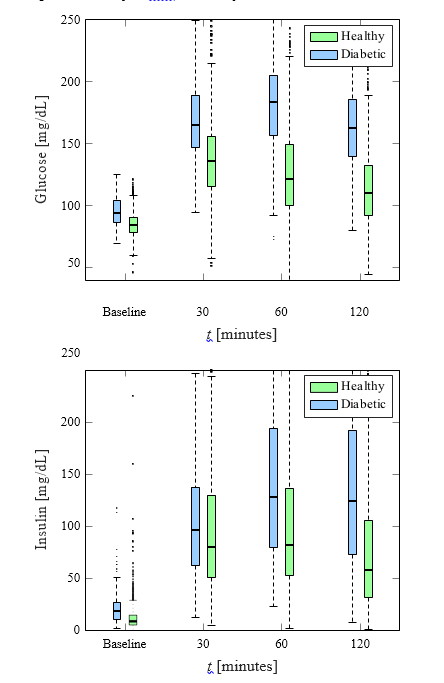


Figure 1: Box plots of glucose and insulin measurements for healthy and diabetic subjects.

### Machine Learning Framework

In this paper, 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 potentially increases computational complexity excessively with the increase of the dimensionality of the input feature space [24]. Introduction of a kernel alleviating 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 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 the slope and area under the curve between all the possible combinations of a pair of c. 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. sthe is denoted by the symbol Δ. The notation Thalf represents the linearly interpolated value between any interval

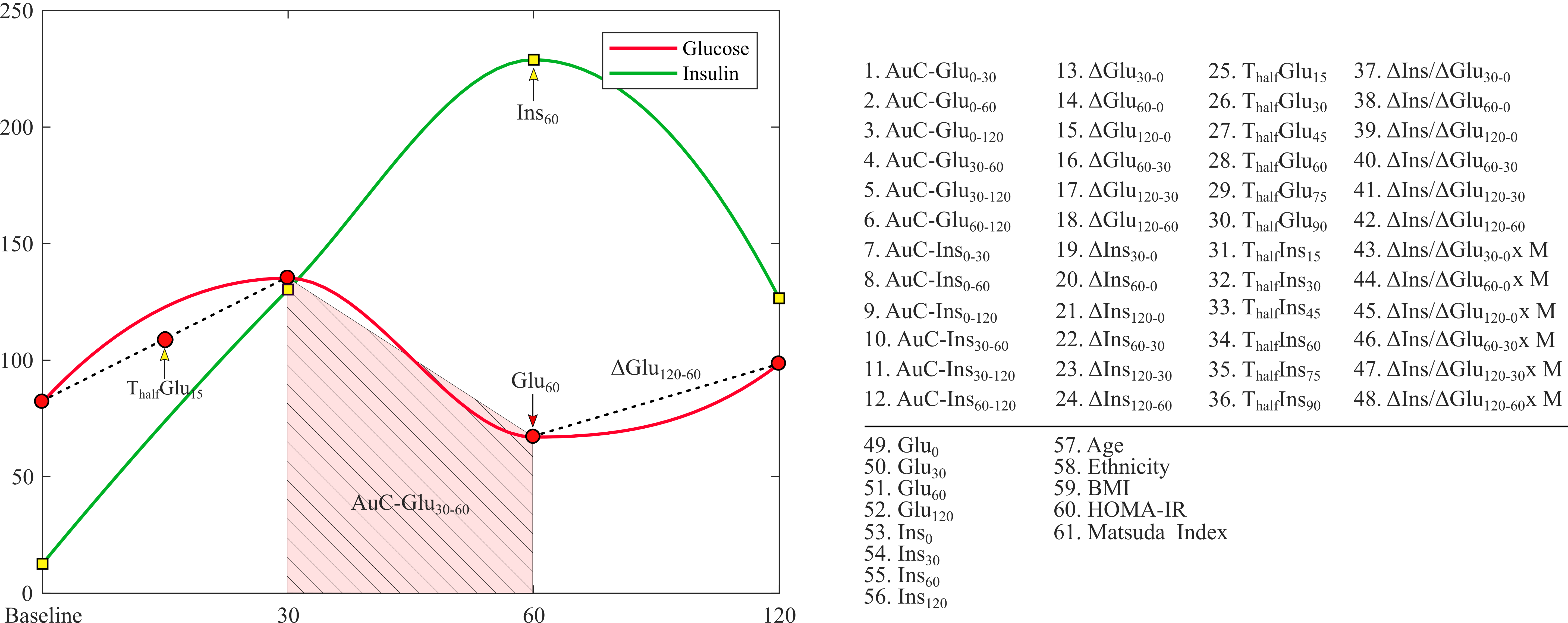
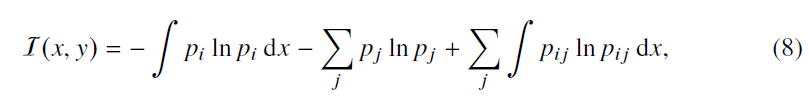
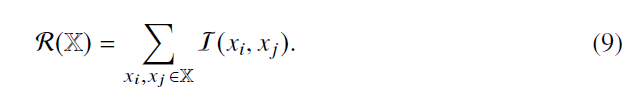


Figure 2: List of all 61 features extracted from the SAHS dataset.

### 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 features from the 61 available features using the minimal-redundancy- maximal-relevance (mRMR) algorithm [28], 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 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}. 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.

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. As shown in Table 1, the SAHS dataset is intrinsically unbalanced with the class distribution skewed toward the majority class with a ratio of 7.5:1. To ensure that a model was unbiased, robust, and generalized well to the new data, we performed 10-fold cross-validation (CV). To standardize the feature range prior to training, the feature space was scaled to unit variance around the mean for each feature respectively. For each combination of features, the data partitioning into training and validation sets were randomly performed 100 times. To analyze the influence of class imbalance on the classification performance we defined the following experiments:

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

Experiment II: The dataset retained the imbalanced class distribution. We randomly sampled 1,360 (out of 1,492 instances) to generate the training set that contained 160 minority instances (T2DM) and randomly sampled 1,200 majority instances (healthy) classes respectively.

Within each experiment, we have compared the performances of a linear and non-linear SVM. The training performance was averaged over all 10-folds. For both experiments, we used all 1023 combinations of ten most relevant features shortlisted during feature selection procedure. The optimal hyperplane parameters of the kernel were determined through a grid search. For selecting the best feature set, we chose geometric mean of specificity and sensitivity [\cite{ kubat1997addressing}] as the to evaluate model.

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, listed in Table 2having maximum dependency to the diabetic class. Notably, the highest ranking features involve the 2-hr values of the blood glucose and insulin measurements. As observed in Table 1, the ethnicity of the subjects is also strong predictor of diabetes.

We aimed to

Table 2: List of ten most relevant features ranked by the mRMR algorithm

|  |  |
| --- | --- |
| Rank | Feature |
| 1 | AuC-Glu0-120 |
| 2 | ΔGlu120-0 |
| 3 | ΔGlu120-60 |
| 4 | ETHN |
| 5 | ΔIns120-0 |
| 6 | ΔGlu60-0 |
| 7 | ΔGlu30-0 |
| 8 | ΔGlu60-30 |
| 9 | ΔIns120-60 |
| 10 | ΔIns60-0 |

Experiment I (balanced data)

The best averaged validation accuracy and sensitivity for each combination is shown in Fig. 3. The x-axis labels represent the feature combinations in which the numbers represent the feature rank as listed in Table 2.

The combination of four features namely, AuC-Glu0-120, ΔGlu120-0, ΔGlu120-60 and ΔGlu30-0 yielded the best performance of 0.78. We choose this combination to generate the classification models in the rest of this paper.

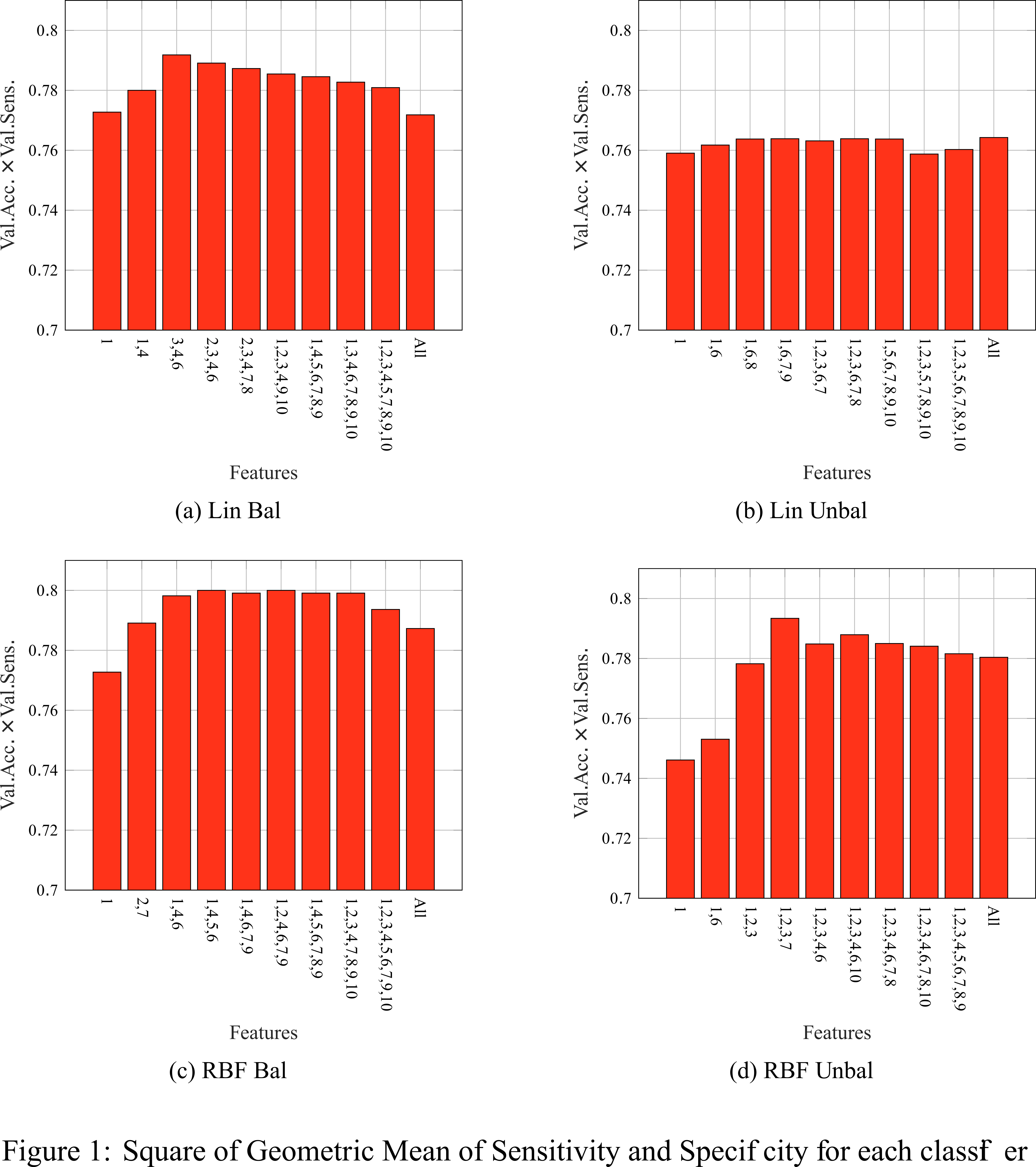


Figure 3: Average performance of the individual features gauged by the accuracy.

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. 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. Table III presents a comparison of the mean training performance of both the balanced and unbalanced cases of the linear and non-linear SVM classifiers averaged over 100 trials. Notably, the random undersampling of the majority class lowers the classifier stability as evident in a greater standard deviation (SD). An otherwise comparable performance suggests that the models are not skewed toward the majority class.

Table 3: Validation performance classifiers

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy ± SD | Sensitivity ± SD | Specificity ± SD |
| Linear SVM (Balanced) | 97.18 % ± 1.48 % | 75.91 % ± 12.68 % | 100 % |
| Linear SVM (Unbalanced) | 96.20 % ± 1.94 % | 77.00 % ± 12.08 % | 98.75 % ± 1.35 % |
| SVM-RBF (Balanced) | 97.73 % ± 1.41 % | 80.64 % ± 12.03 % | 100 % |
| SVM-RBF (Unbalanced) | 96.46 % ± 1.40 % | 80.45 % ± 11.50 % | 100 % |
| Two-step Approach [16] | - | 77.70 % | 77.40 % |
| SADPM [15] | 56.329 % | 88.80 % | 52.00 % |
| IGT (PG120 > 140 & <200 mg/dL) |  | 33.93 % |  |
| WHO Criteria (PG120 > 200 mg/dL) |  | 8.19 % |  |

Table III displays the mean validation performance of the classifiers. All the trained models 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 of the training data. The non-linear SVM classifier in which the class ratio was maintained as in the original dataset provided a sensitivity of 80.45%. Although the performance in the case when the classes are balanced is better, the minor improvement does not justify the artificial alteration of the class distribution of the original dataset. The obtained results were also compared to other T2DM prediction models that used the SAHS dataset. The San Antonio diabetes prediction model (SADPM) [15] used a logistic regression based on a person’s age, gender, ethnicity and the fasting glucose level. The two-step approach used the SADPM to first create a risk score, which was then combined with the 1-hr blood glucose concentration to identify high-risk individuals.

# Conclusion

In this paper, we developed a SVM prediction model to identify the individuals that are at an increased risk of developing T2DM in the future. We showed that a high prediction performance can be achieved by extracting information from blood glucose levels obtained during an oral glucose tolerance test. The accuracy, and more importantly sensitivity of the presented classifier, improved significantly compared to the previous prediction models that used the same dataset.

# 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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] ——, “The shape of plasma glucose concentration curve during OGTT predicts fu- ture risk of type 2 diabetes,” *Diabetes/Metabolism Research and Reviews*, vol. 26, no. 4, pp. 280–286, May 2010.

[12] M. A. Abdul-Ghani and R. A. DeFronzo, “Plasma Glucose Concentration and Prediction of Future Risk of Type 2 Diabetes,” *Diabetes Care*, vol. 32, no. suppl\_2, pp. S194–S198, Nov. 2009.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] S. M. Haffner, M. P. Stern, H. P. Haztjda, J. A. Pugh, and J. K. Patterson, “Hy- perinsulinemia in a Population at High Risk for Non-Insulin-Dependent Diabetes Mellitus,” *New England Journal of Medicine*, vol. 315, no. 4, pp. 220–224, Jul. 1986.

[20] M. Wei, S. P. Gaskill, S. M. Haffner, and M. P. Stern, “Effects of Diabetes and Level of Glycemia on All-Cause and Cardiovascular Mortality: The San Antonio Heart Study,” *Diabetes Care*, vol. 21, no. 7, pp. 1167–1172, Jul. 1998.

[21] M. Matsuda and R. A. DeFronzo, “Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp.” *Dia- betes Care*, vol. 22, no. 9, pp. 1462–1470, 1999.

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

[23] 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.

[24] 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.

[25] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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

[27] G. V. Trunk, “A problem of dimensionality: A simple example,” *IEEE Transac- tions on Pattern Analysis and Machine Intelligence*, no. 3, pp. 306–307, 1979.

[28] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans- actions on Pattern Analysis and Machine Intelligence*, pp. 1226–1238, 2005.