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

November 15, 2018

**Abstract**

Diabetes is a large healthcare burden worldwide. With staggering evidence that lifestyle modifications and drug intervention can prevent diabetes, early identification of high risk individuals is important for targeted prevention strategies. We present an automatic tool that uses machine learning techniques to predict development of type 2 diabetes mellitus (T2DM). Data generated from an oral glucose tolerance test (OGTT) was used to develop a predictive model based on the support vector machine (SVM). We trained and validated the models using the OGTT data and epidemiological 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, epidemiological information as age, ethnicity and BMI were also a part of the dataset. Using these 11 measurements, we have generated 61 deduced features out of which the 10 most relevant features were selected after ranking by the mRMR algorithm. Subsequently, all possible combinations of these 10 features were used to generate SVM prediction models. This research shows that insulin and epidemiological features are not relevant for prediction of diabetes. Furthermore, the models using the information extracted from blood glucose levels reached a high predictive performance. The results of our approach show an average accuracy of 96.46% and a specificity of 80.45% 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 with prevalence among the adult population increased from 4.7 % in 1980 to 8.5 % in 2014 [1]. In 2015 alone, 1.6 million deaths worldwide were estimated to be attributed to diabetes. In addition, compared to a non-diabetic individual, a diabetic patient is at a greater risk of developing cardiovascular diseases, visual impairment and limb amputations. Due to the substantial socio-economic burdens that are associated with diabetes, its early detection, intervention and prevention has become a worldwide top-level health concern. Currently 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, adopt exercise, and adhering to a pharmacological treatment [2]. The early identification of individuals prone to a high risk of diabetes is therefore, essential for designing targeted prevention strategies [3].

In the last two decades the number of clinical studies aimed at diagnosing diabetes have been growing steadily. However, there is a limited amount of clinical studies that predict the risk of developing diabetes. Although the clinical significance of such predictions largely depend on the type and quality of data collected, the diabetes prediction has received an increased amount of research interest in the last decade [4]. There are studies that assign a probability of having a future risk of diabetes using sociodemographic characteristics that are acquired through population surveys through [5, 6]. Due to the high possibility of inaccurate data entries, such approach can be misleading. The information collected from specimens such as blood on the other hand, provides an accurate first step of disease prognosis with a deeper clinical insight [REF]. In regard to diabetes, the oral glucose tolerance test (OGTT) not only is used to diagnose it, but provides a critical understanding of its future evolution through the beta-cell function and insulin resistance that are widely regarded as the major factors of the development of T2DM. In an OGTT, the blood glucose and insulin levels are measured at regular intervals in a 2-hr period after orally administering a standard dose of glucose [\cite stumvoll\_use\_2000}]. A precursory stage 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], which suggests that the fasting and 2-hour blood glucose levels used exclusively cannot accurately determine the future development of T2DM.

In contrast to traditional scientific approaches using population based statistics, machine learning (ML) methods develop models that are trained using ample data. ML has currently been proposed as an accurate method for diabetes screening. Barakat et al used socio-demographic information, and point of care measurement form blood and urine to develop diabetes risk models [13]. This approach uses ensemble approach based upon a combination of SVM and random forest. Han et al employed SVM to develop a decision making algorithm for the diagnosis of diabetes [14] which provided an added comprehensibility of the classification mechanism. This approach uses socio-demographic information and glucose levels at baseline and 2h thereafter during an Previous investigations designed to identify individuals at high risk of developing T2DM in future included San Antonio Diabetes prediction model (SADPM) [15] where a logistic regression model was constructed using 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].

We hypothesized that features extracted from OGTT data will be able to predict future onset of diabetes. In this paper, we develop a diabetes prediction model by identifying the features, computed from the OGTT data that strongly correlate to future diabetes. Furthermore, we develop a support vector machine prediction model using these features. For this purpose, we use the OGTT data generated from the population-based, epidemiological study, the San Antonio Heart Study (SAHS) [17, 18].

# Materials and Methods

### 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.

T2DM was diagnosed at an average follow-up of 7.5 years using 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 for the case of this study. 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 after the intake of standardized glucose concentration during the OGTT. Out of 5,158 entries available in the SAHS dataset, we have excluded xxx for incomplete entries. Table 1 …

Table 1: The classification of the 1,496 subjects used in this study

|  |  |  |  |
| --- | --- | --- | --- |
| Healthy | DMI | CVD | DMI+CVD |
| 1.281 | 161 | 44 | 10 |
| 85.63 % | 10.76 % | 2.94 % | 0.67 % |

.

### Machine Learning Framework

We developed a supervised learning scheme in which the classifier output labels were obtained from the follow-up data, and the support vector machine (SVM) technique was used to construct the future diabetes prediction framework. The SVM develops models from a given training dataset such that it generalizes well to new datasets and minimizes the empirical risk associated with misclassification of samples in the training set [22,23]. For a binary classification problem, the model constructed by the SVM finds a decision boundary or a separating hyperplane which aims to minimize the overlapping between the two classes in the training set. 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 is first transformed to a higher dimension and then a linear boundary is determined, which generally gives better training performance [24].

Let us consider a training data (*x*1, *y*1), (*x*2, *y*3),..., (*xk*, *yk* ) of *k* pairs containing *xi* ∈ R*N* features and the binary classes *yi* ∈ −1, 1. The SVM approach transforms the input features using a nonlinear mapping Φ: *x* f→ φ(*x*) into a higher dimension space R*P* , 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 ζ in the higher dimension space R*P* is a common practice which allows for the classifier

250

Healthy Diabetic

200

150

Glucose [mg/dL]

100

50

250

Baseline 30 60 120

*t* [minutes]

200



Healthy Diabetic

150

Insulin [mg/dL]

100

50

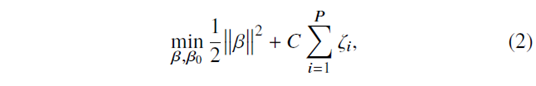
0

Baseline 30 60 120

*t* [minutes]

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

output in (1) to be on the incorrect side of the margin. To define the optimal separating hyperplane that maximizes the distance *M* from the boundary for all data points, and bounds the value of I,*i* ζ*i*, we introduce the convex optimization problem,

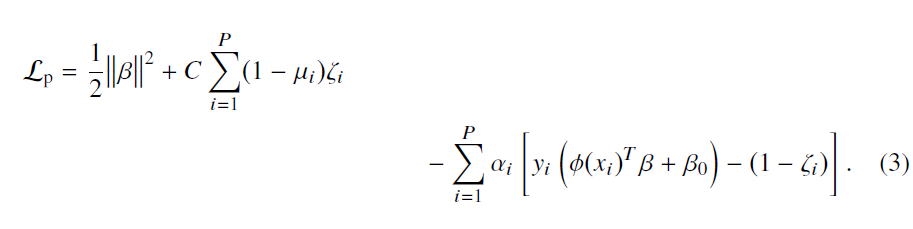


with the nonlinear constraints *yi* (φ(*xi*

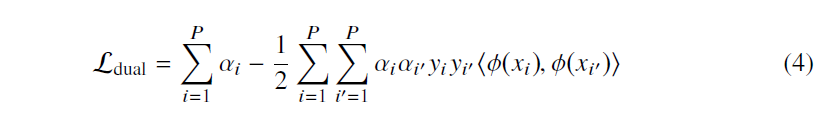
)

β + β0) ≥ 1 − ζ*i* ∀*i* and ζ*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 [24],



By minimizing Lprimal with respect to β, β0, and ζ*i* , we get the corresponding dual form of the Lagrange function,

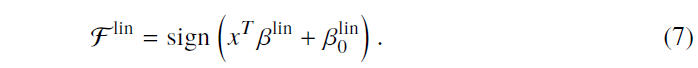


subject to 0 ≤ α*i* ≤ *C* and α*i yi* = 0 and the constraints, ζ*i*, µ*i* ≥ 0∀*i*. The nonzero coefficients α*i* and β0 are determined using (1). The computation complexity to compute the mapping Φ increases excessively with the increasing dimension of the input feature space With the introduction of a kernel,

K(*x*, *x*,) = <φ(*x*), φ(*x*,)>, (5)

the inner product on (4) can be computed without the mapping Φ [25] . In this paper, we used the Gaussian radial basis function, as the kernel where σ is a free parameter. 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.

For a linear SVM, the classifier can be expressed without coordinate mapping as follows:,



### Feature Extraction

We extracted all the features from the SAHS data acquired at the baseline. The dataset consists of the blood glucose and insulin values 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 [\cite{rapid\_rise\_1984\_stern}]. Furthermore, the socio-demographic information such as age, ethnicity and body-mass index (BMI) was also part of the dataset. From the glucose and insulin measurements, 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 index [12], and Matsuda index [21]. We added these parameters as they have shown a good efficacy for diabetes prediction in previous studies [9, 12]. A total of 61 features (illustrated in Figure 2) are used in this study

• Raw measurements of blood glucose and insulin concentrations (N = 8)

• Deduced features from measured blood glucose and insulin defined as the slope and area under the curve N = 48)

• Features previously used to predict diabetes in the healthy population (N=3):

* IInsulinogenic index [8]: ratio of insulin and glucose slopes between any two time intervals,
* Matsuda index [20]:
* Homeostatic model assessment of insulin resistance (HOMA-IR) [REF]: product of fasting blood glucose and fasting blood insulin divided by 22.5.

• Sociodemographic characteristics: age, ethnicity, and BMI

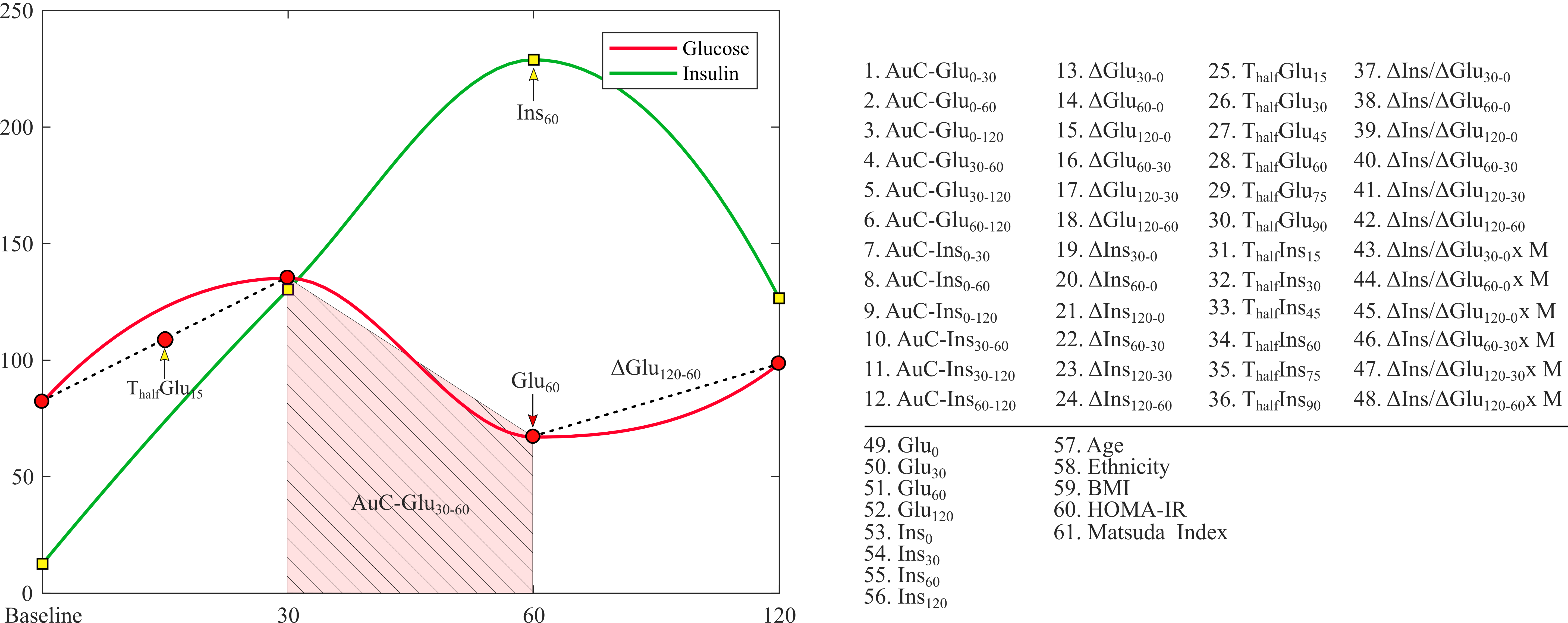
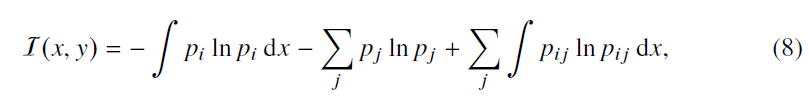


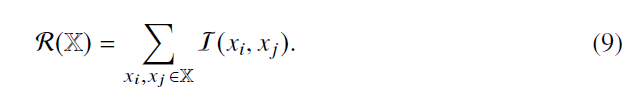
Figure 2: List of all 61 features extracted from the OGTT.

### 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. Therefore we reduced and ranked 61 features to ten most relevant features. To select the ten most relevant features, we used the whole dataset available. This step decreases the computational cost during the model development by reducing the dimensionality of the feature space. To prevent the selection of highly correlated features, which adds to the computational cost of the classifier without necessarily improving the performance [26], we minimize the correlation among the features by minimal redundancy-maximal-relevance (mRMR) algorithm [27]. This algorithm utilize relevance between feature (x as continuous random variable) and the class labels (y as discrete random variable) 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 ∈ (−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 that are most relevant to the class label are the ones that individually yield the maximum of 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 [27]. 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 [28], which selects the features that not only yield the maximal mutual information (8) with respect to the class label, but 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 T2DM, on the basis of yielding maximum I with respect to the diabetic class.



# Classification

We developed a supervised learning scheme using baseline SAHS data and the labels (healthy, diabetic) obtained at the follow-up after an average of 7.5 years. The SVM was used for. In each experiment, we used a kernel-based SVM method [REF] to train, test and validate the performance of the diabetes prediction models. T the 44 CVD entries were excluded for potentially wakening the labels as the only way of defining this class was based upon the self-reporting and not on quantitative assessment. Furthermore, we also removed all entries with missing information.

The minority class of diabetic subjects was defined 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 label of -1. 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 a model that is unbiased, robust, and generalize well to the new data, we performed 10-fold cross-validation (CV). To standardize the feature range

prior to training, we have scaled the feature space to unit variance around the mean for each feature respectively. Within each 10-fold CV 11 positive labels were held out for validation. This class was matched with balanced or unbalanced number of negative labels, respectively. For each combination of features, the data partitioning into training and validation sets were randomly performed 100 times. 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 1,281 instances) to generate the training set that contained 160 and randomly samples 1,200 instances of the diabetic and healthy classes respectively.

Within each experiment, we have compared the performances of a linear and non-linear SVM, and the training performance was averaged over all 10-folds. The optimal hyperplane parameters were assessed through a grid search with a view to maximize the sensitivity. For both experiments, we used all 1023 combinations of ten most relevant features shortlisted during feature selection procedure. The optimal hyperplane parameters *C* and σ in and were determined through a grid search with a view to maximize the classifier sensitivity [REF]. For selecting the best feature set, we chose the criteria of maximizing the product of validation accuracy and sensitivity.

In each experiment and for both linear and non-linear SVM, the number of features was increased incrementally from single feature to a combination of ten features ranked during the feature selection. All the 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 the ranked features, listed in Table 2, that are ordered in decreasing relevance. . The prefixes *AuC* and *Sl* denote the area under the curve and slope respectively. The OGTT time interval corresponding to the feature appears in the subscripts.

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.

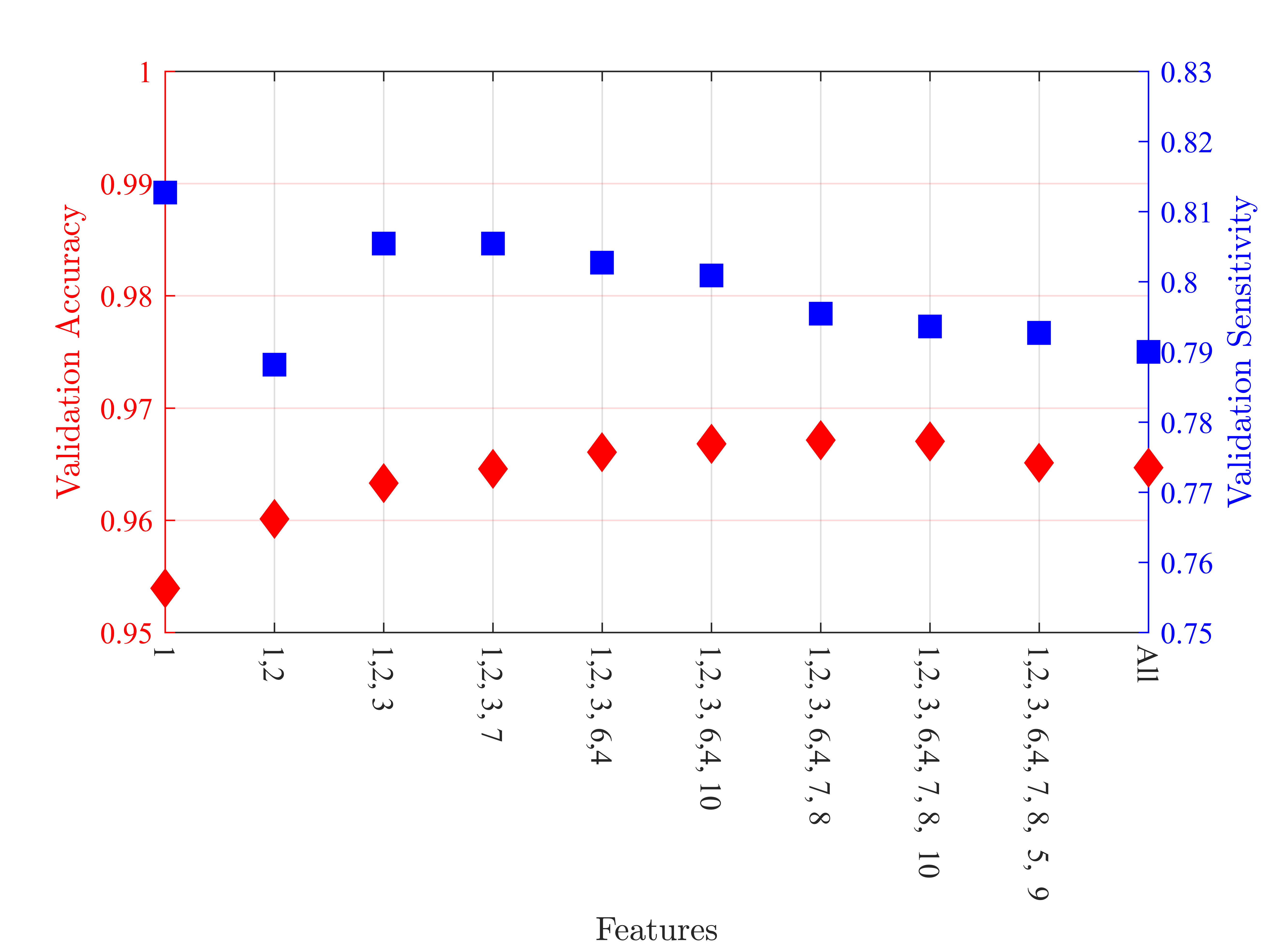


Figure 3: Average performance of the individual features gauged by the accuracy.

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 IV 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. The optimal hyperplane parameters corresponding to the Matlab arguments ‘BoxConstraint’ and ‘Gamma’, were respectively assigned the values of 1.0 and 6.0. For the linear version of the SVM, an average of 250 and 1,114 support vectors were

Table 3: Mean Training performance of the classifiers

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Accuracy ± SD | Sensitivity ± SD | Specificity ± SD |
| Linear SVM (Balanced) | 78.46 % ± 1.47 % | 77.58 % ± 1.37 % | 79.34 % ± 2.22 % |
| Linear SVM (Unbalanced) | 79.66 % ± 0.45 % | 77.77 % ± 0.84 % | 79.91 % ± 0.45 % |
| SVM-RBF (Balanced) | 78.33 % ± 1.70 % | 78.44 % ± 1.84 % | 78.22 % ± 2.37 % |
| SVM-RBF (Unbalanced) | 78.90 % ± 0.53 % | 78.97 % ± 0.51 % | 78.36 % ± 0.77 % |

Table 4: 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 % |

used to construct the hyperplane for the balanced and unbalanced datasets respectively. On the other hand, the corresponding values were 278 and 1,198 for the nonlinear SVM with the RBF as the kernel. It should be noted that the difference in the dimensionality of the hyperplanes between the two variants of the SVM is not large, which indicates that the discriminating power of the features used.

Table 4 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 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.