A Type 2 Diabetes Prognostic Model based on Support Vector Machine

Hasan Abbas, Lejla Alic, Shafiqul Islam, Madhav Erraguntla, Muhammad Abdul-Ghani, Qammer Abbasi, *Senior Member, IEEE,* and Marwa Qaraqe, *Member, IEEE*

***Abstract*—We present a prediction model that assesses the future risk of developing type 2 diabetes mellitus (T2DM). We use the oral glucose tolerance test (OGTT) data collected from a group of 1,551 healthy subjects to construct a machine learning model employing the support vector machines. We trained and validated the models on the data obtained from the second cohort of the San Antonio Heart Study, using a set of four features derived from the glucose measurements in the OGTT. The results of the proposed scheme show an average validation accuracy of 97.23% and recall of 77.27%. The results also show that the plasma glucose based features are the strongest predictors of the future development of T2DM.**

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

I. Introduction

THE global incidence of diabetes was estimated at 422 million in the year 2014 and its prevalence among the adult population has seen an increase from 4.7 % in the year 1980 to 8.5 % in 2014 [1]. In 2015 alone, an estimated 1.6 million deaths worldwide were attributed to diabetes. In addition, a diabetic patient is at a greater risk of developing cardiovascular diseases, visual impairment and limb amputations, as compared to a non-diabetic person. 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.

The presumption of delaying or even preventing the future development of future diabetes is backed by experimental evidence [2], provided the person undergoes a lifestyle change that includes managing diet and incorporating exercise, and adheres to a medical treatment. Moreover, one of the first signs of diabetes is the impaired glucose tolerance (IGT) in which the blood sugar level rises beyond the normal levels defined by the World Health Organization (WHO) and the American Diabetes Association (AMA). The oral glucose tolerance test (OGTT) is commonly used to screen the individuals that have IGT, and are at an increased risk of developing diabetes in the future. In an OGTT, the blood glucose and insulin levels are periodically recorded in a 2 h period from a person that has undergone an overnight fast and administered a standard concentration of glucose. The outcomes of the OGTT provide useful information on the IGT, as well as any impairment in the insulin function of the body.

On the contrary, studies have indicated that only 50 % of subjects with IGT went on to develop diabetes within a span of 10 years [3], [4]. Furthermore, long-term population studies have shown that around 50 % future diabetic subjects did not exhibit IGT at all [5]. Previously, it has been shown that as compared to the IGT, the glucose concentrations at the 1 h and 2 h intervals during the OGTT, correlate more to the future diabetes risk [6]–[8].

Research studies that assess diabetes can be broadly categorized into two themes, first of which deals with the objective to detect any undiagnosed state of diabetes, and the second that aims to identify the person that are high risk of developing diabetes in the future [**?**]. The clinical significance of such investigations depends upon the data collection methods. Certain studies rely on collecting socio-demographic characteristics such as age, ethnicity, body mass index (BMI) and genealogical information through conducting population surveys, and then assign a probability to individuals of having diabetes [**?**], [9]. However, such self-assessment techniques can often be misleading and can not be relied upon. On the other hand, the outcomes of the diabetes related studies that involve physiological data such as blood samples collected in a laboratory environment provide an accurate clinical insight. We can further divide these types of enquiries into two types, namely, the screening of undiagnosed diabetes, and the future prediction of diabetes. The former category has seen an increased amount of research interest in the last ten years. Using statistical and machine learning techniques, various researchers have developed *risk models* for diabetes screening. In [10], support vector machine (SVM) framework was employed for the diagnosis of diabetes that also incorporated a tree-based decision making algorithm. A rule-based ensemble method combining SVM and random forest (RF) classifiers was used to detect diabetes in [11] which provided an added comprehensibility of the classification mechanism.

Previous investigations designed to identify individuals at high risk of developing type 2 diabetes in future included San Antonio Diabetes prediction model (SADPM) [12] where a logistic regression model was constructed using a subject's physiological parameters such as systolic blood pressure and cholesterol level. The underlying causes of type 2 diabetes

H. Abbas is with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874 Qatar; hasan.abbas@qatar.tamu.edu

L. Alic was with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874 Qatar; lejlaresearch@gmail.com

S. Islam and M. Qaraqe are with the College of Science and Engineering, Hamad Bin Khalifa University Doha, Qatar

M. Erraguntla is with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843

M. Abdul-Ghani is with the UT Health, San Antonio, TX 78229

Q. Abbasi is with the School of Engineering, University of Glasgow, UK

Manuscript received April 19, 2005; revised August 26, 2015.

in the form insulin resistance and insulin secretion were studied to develop a prediction model in [5]. In another study, multivariate logistic models using the plasma glucose values measured in the OGTT were used to predict the future risk of developing type 2 diabetes [6], [13].

In this paper, we propose to develop a future diabetes prediction model by first identifying the variables computed from the OGTT that strongly correlate to future diabetes and then develop a support vector machine prediction model using these feature variables. For this purpose, we use the OGTT data generated from the population-based, epidemiological study, the San Antonio Heart Study (SAHS) [14], [15].

II. Materials and Methods

A. San Antonio Heart Study

We developed the diabetes prediction models using the data extracted from a population-based epidemiological study, the San Antonio Heart Study (SAHS). The aim of the SAHS was to assess the risk factors of diabetes and cardiovascular diseases [14], [15], for which 5,158 men and non-pregnant women of Mexican-American and non-Hispanic white residents of San Antonio, Texas were recruited. The age of the subjects 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 an oral glucose tolerance test (OGTT), which measures the subject's body response to a standard 75 g dose of glucose after fasting overnight. The OGTT was performed both at the baseline and follow-up phases of the study, which had an average span of 7.5 years. The SAHS subjects were recruited in 2 cohorts, the first during the period 1979 to 1982, and the second from 1984 to 1988 [16]. 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. For the future T2DM prediction problem, we construct the machine learning model using the data from the second cohort, in which the plasma glucose and insulin levels of 1,496 of healthy subjects were recorded during the OGTT at times 0, 30, 60 and 120 minutes in the baseline evaluation. During the course of the study, a total of 171 subjects developed T2DM within which 10 subjects also reported of at least one cardiovascular event such as a heart attack, stroke or angina.

At the follow-up assessment, the participants were classified as having type 2 diabetes (T2D), cardiovascular disease (CVD) or normal. For T2D diagnosis, the WHO criteria, defining fasting glucose level $\geq$126 mg/dL or 2-hour glucose level $\geq$200 mg/dL was followed [17]. Any participant reportedly taking anti-diabetic medications was also classified as diabetic. For CVD classification, any cardiovascular event such as a heart attack, stroke or angina reported by the participant, was considered as an identifier. Table I outlines the distribution of patient classification used in this study. In order to construct a binary classifier for this study, the subjects categorized under the `DMI` and `DMI+CVD` were encoded as the positive class whereas the `Heahlty` labels were the negative class. We restricted the classification to only two classes and the samples with the label `CVD` were ignored.

TABLE I: 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 %

B. Preparation of the Data

The dataset included the glucose and insulin values recorded at the baseline, 30, 60 and 120 minute intervals, and a distribution of these values marked by the follow-up labels of 'healthy' and 'diabetic' is shown in Fig. 1. Moreover, 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 (ratio of insulin and glucose slopes between any two time intervals) and Matsuda indices, as defined in [8], [18]. 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 prepared 68 features 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. As can be observed from the Table I, the SAHS dataset is intrinsically imbalanced with the class distribution skewed toward the majority class with a ratio of 7.5:1. Considering the Therefore, we took two approaches for the training set, one in which the dataset was balanced by a randomized undersampling of the majority class, and second in which we persisted with natural class distribution. In both the cases, the validation set had the same class distribution representative of the original dataset. A total of 1360 samples were used for training that accounted for 91 % of the dataset, and 99 samples, consisting of 11 diabetic and 83 healthy subjects were reserved for the validation of the trained prediction models.

C. 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 works on the principle of *structural risk minimization* (SRM) in which the goal is to develop a model from the given training data such that it generalizes well to new datasets and minimizes the empirical risk associated with misclassification of samples in the training set [19], [20]. 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,

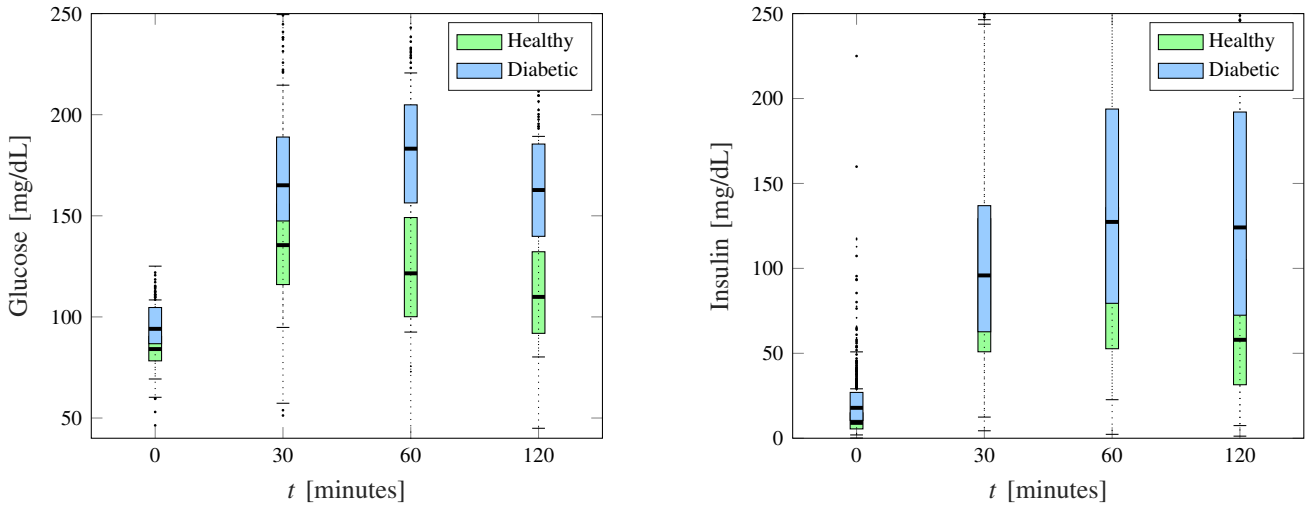

Fig. 1: Box plots of glucose and insulin measurements for healthy and diabetic subjects.

which generally gives better training performance [21]. Let us consider a training data (x_1, y_1),(x_2, y_3),...,(x_k, y_k) of k pairs containing $x_i \in \mathbb{R}^N$ features and the binary classes $y_i \in -1, 1$. The SVM approach transforms the input features using a nonlinear mapping $\Phi : x \mapsto \phi(x)$ into a higher dimension space $\mathbb{R}^P$, where in general $P \gg N$. Due to the transformation, the classes can then be separated using a linear decision boundary in the enlarged space. The non-linear SVM classifier $\mathcal{F}$ is expressed in terms of the higher dimensional hyperplane,

$$\mathcal{F} = \text{sign}\left(\phi(x)^T\beta + \beta_0\right). \tag{1}$$

When the classes may not be completely separable, introducing a slack variable ζ in the higher dimension space $\mathbb{R}^P$ 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 $\sum_i \zeta_i$ and in turn misclassification rate, we introduce the convex optimization problem,

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{P}\zeta_i, \tag{2}$$

with the nonlinear constraints $y_i(\phi(x_i)^T\beta+\beta_0) \geq 1-\zeta_i \quad \forall i$ and $\zeta_i \geq 0$, 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 [21],

$$\mathcal{L}_\text{p} = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{P}(1-\mu_i)\zeta_i - \sum_{i=1}^{P}\alpha_i\left[y_i\left(\phi(x_i)^T\beta+\beta_0\right) - (1-\zeta_i)\right]. \tag{3}$$

By minimizing $\mathcal{L}_\text{primal}$ with respect to β, β_0, and ζ_i, we get the corresponding dual form of the Lagrange function,

$$\mathcal{L}_\text{dual} = \sum_{i=1}^{P}\alpha_i - \frac{1}{2}\sum_{i=1}^{P}\sum_{i'=1}^{P}\alpha_i\alpha_{i'}y_iy_{i'}\langle\phi(x_i),\phi(x_{i'})\rangle \tag{4}$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$ and the constraints, $\zeta_i, \mu_i \geq 0 \forall i$. The nonzero coefficients α_i and β_0 are determined using (1). As the dimension of the input feature space goes up, the computation of the mapping Φ gets excessive in complexity. With the introduction of a kernel,

$$\mathcal{K}(x, x') = \langle\phi(x), \phi(x')\rangle, \tag{5}$$

we can compute the inner product on (4) without computing the mapping Φ [22], which becomes computationally expensive as the dimension of the input feature space increased. In this paper, we used the Gaussian radial basis function,

$$\mathcal{K}(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \tag{6}$$

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.

D. Feature Selection

Before constructing the SVM model to predict the future risk of diabetes, 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 in to 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 $\mathbb{R}^N$ 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, $\mathcal{I}$ defined as [23]:

$$\mathcal{I}(x,y) = -\int p_i \ln p_i \, \mathrm{d}x - \sum_j p_j \ln p_j + \sum_j \int p_{ij} \ln p_{ij} \, \mathrm{d}x, \tag{7}$$

where p_i, and p_j are the probabilities of the random variables x and y taking a particular value x_i and $y_j \in (-1,1) \forall j$ respectively. The term p_{ij} denotes the joint probability $P\{x = x_i, y = y_j\}$. The three terms in (7) 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 $\mathcal{I}$. 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 [24]. 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 $\mathbb{X}$. We follow the minimal-redundancy-maximal-relevance (mRMR) algorithm [25], that selects the features, that not only yield the maximal mutual information (7) with respect to the class label, but minimizes the mutual correlation among the features expressed in terms of redundancy $\mathcal{R}$ as:

$$\mathcal{R}(\mathbb{X}) = \sum_{x_i, x_j \in \mathbb{X}} \mathcal{I}(x_i, x_j). \tag{8}$$

where $\mathcal{I}$ follows its definition in (7). By minimizing $\mathcal{R}$, 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 $\mathcal{I}$ with respect to the diabetic class. The application of the mRMR algorithm produces the features that are listed in Table II that are ranked in order of their relevance. The prefixes *AuC* and *Sl* denote the area under the curve and slope respectively, and the OGTT time interval corresponding to the feature appears in the subscripts.

TABLE II: List of ten most relevant features ranked by the mRMR algorithm

Rank	Feature
1	AuC-$Glu_{0\text{-}120}$
2	Sl-$Glu_{120\text{-}0}$
3	Sl-$Glu_{120\text{-}60}$
4	Sl-$Glu_{60\text{-}0}$
5	Sl-$Glu_{30\text{-}0}$
6	AuC-$Glu_{60\text{-}120}$
7	PG_0
8	PG_{120}
9	PG_{60}
10	AuC-$Glu_{30\text{-}120}$

In the second phase, we further refined the number of variables to four by only selecting the ones which provided the best performance using the SVM classification scheme with the parameters C and γ preconfigured to a value of 1. For this purpose, we employed the accuracy achieved in the validation set as the evaluation criterion. Table III shows the mean validation accuracies of the variables which is obtained by performing 100 iterations of the SVM classifier supplied with only one variable at a time.

TABLE III: Average performance of the individual features gauged by the accuracy

Features	Mean Accuracy (Standard Deviation)
AuC-$Glu_{60\text{-}120}$	0.973 (0.013)
PG_{120}	0.971 (0.015)
PG_{60}	0.967 (0.022)
AuC-$Glu_{0\text{-}120}$	0.958 (0.019)
AuC-$Glu_{30\text{-}120}$	0.950 (0.018)
Sl-$Glu_{60\text{-}0}$	0.946 (0.025)
Sl-$Glu_{120\text{-}0}$	0.931 (0.026)
PG_0	0.816 (0.039)
Sl-$Glu_{120\text{-}60}$	0.763 (0.050)
Sl-$Glu_{30\text{-}0}$	0.745 (0.044)

III. SVM Model Generation

1. Balanced Training -> Unbalanced Validation 2. Unbalanced Training -> Balanced Validation Cross-validation 100 iterations to remove any randomized errors grid search of the best parameters C and gamma In this study, we employed the linear SVM kernel by utilizing the Matlab's `svmtrain` function. The training data was first scaled to have a unit standard deviation. The misclassification cost was configured by setting the value of the `boxconstraint` parameter to a high value of 100, which would cause a stricter partitioning of the data with respect to the class labels.

To predict the future risk of type 2 diabetes, we defined a positive class (occurrence of diabetes at the follow-up) and a negative class (healthy). As illustrated in Table I, the OGTT data used in this study is heavily unbalanced. With 171 positive class instances as compared to 1281 that of the negative class, the size of class labels is unbalanced with the ratio of positive-to-negative instances of 1:8. To avoid the problem of overfitting to the majority class during the learning phase of the technique, we under-sampled the majority class (healthy) to the size of the minority class (diabetic) by a randomly selecting equal number of samples. During the prediction model generation, we employed 10-fold cross-validation framework in which 90 % of the training data, consisting of 360 samples was used for training and the remaining 10 % was used to test the model. To validate the trained models, we used a holdout data set with the same unbalanced ratio of negative-to-positive classes in the original data, i.e., 11 samples of the positive class, and 88 samples of the negative class. We started our experiments using one feature at a time, and then more number of features were incrementally added. This exercise assists in discovering any feature dependencies. In total, we performed 1,023 classification experiments. Each of these experiments was trained as a 10-fold cross-validation (CV) and, to minimize the effect of random selection of samples from the majority class, 100 iterations were performed for each experiment. Owing to the small sample size of the holdout

dataset, this strategy ensures the unbiased reporting of the classifier performance. To maximize reliability of the model to predict diabetes events, we maximized the recall metric during the training phase, which is defined as,

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

where TP and FN are the true-positives and false-negatives respectively. During the validation phase, we tracked the confusion matrices for all the models yielding the maximum training recall for all the feature combinations.

IV. Conclusion

The conclusion goes here.

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] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes risk calculator," *Diabetes Care*, vol. 31, no. 5, pp. 1040–1045, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] S. M. Haffner, M. P. Stern, H. P. Haztjda, J. A. Pugh, and J. K. Patterson, "Hyperinsulinemia 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.

[19] 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.

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

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

[22] 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.

[23] 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.

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

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

[26] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on pattern analysis and machine intelligence*, no. 3, pp. 306–307, 1979.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1226–1238, 2005.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.