Paper Review

Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes - Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury

Name: Kavana Manvi Krishnamurthy
ID: 2158984

Part 1: Problem being addressed. What was the data?

The study addresses the need for improved methods to classify individuals with and without common diseases, specifically focusing on diabetes. The authors present an approach using support vector machine (SVM) techniques, which they argue may have advantages over traditional methods like logistic regression. The problem being addressed is the challenge of early detection and classification of diabetes and pre-diabetes in the population, which is crucial for implementing effective prevention strategies. With an estimated 23.6 million people affected by diabetes in the U.S., of whom about one-third are unaware of their condition, and an additional 57 million with pre-diabetes, there is a pressing need for accurate and efficient classification tools.

The study aims to develop and validate SVM models using data from the National Health and Nutrition Examination Survey (NHANES) data. It is an ongoing, cross-sectional, probability sample survey of the U.S. population. Data was collected by trained professionals by inviting participants for detailed physical, physiological and lab exams. The study focused on non-pregnant participants aged 20 or older.

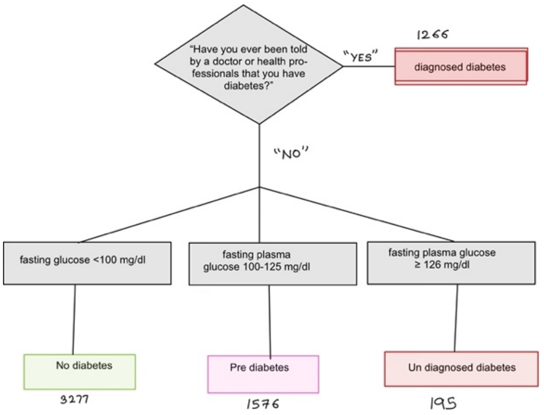
"Have you ever been told by a doctor or health pro-fessionals that you have diabetes?"
"YES"
1266
diagnosed diabetes
"NO"
fasting glucose <100 mg/dl
fasting plasma glucose 100-125 mg/dl
fasting plasma glucose ≥ 126 mg/dl
No diabetes
3277
Pre diabetes
1576
Un diagnosed diabetes
195

Fig 1. Flowchart of Participant classification

In the study, participants were classified based on their diabetes status through two main criteria. Those who responded affirmatively to the question "Have you ever been told by a doctor or health professional that you have diabetes?" were considered to have diagnosed diabetes. Participants who had not been diagnosed but had a fasting plasma glucose level of ≥ 126 mg/dl were categorized as having undiagnosed diabetes. Individuals with fasting glucose levels ranging from 100-125 mg/dl were classified as having pre-diabetes, while those with fasting glucose levels below 100 mg/dl were considered free from diabetes. This classification process is illustrated in Figure 1's flowchart.

Classification Scheme I - diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes
Classification Scheme II - undiagnosed diabetes or pre- diabetes vs. no diabetes

Total number of the cases for classification scheme I = 1461
Total number of the non-cases for classification scheme I = 4853
Total number of the cases for classification scheme II = 1709
Total number of the non-cases for classification scheme II = 3206

Part 2: What machine learning methods were used, and how?

Pre-processing: To prepare the dataset for training, an equal number of cases and non-cases were selected to ensure balance and prevent bias in the model. All feature values were normalized to fit within a range of -1 to +1 to standardize the data, making it easier for the model to process. Categorical variables, like race, were assigned arbitrary numerical values within this range, while continuous variables, such as age, were scaled by dividing by an appropriate number (e.g., age was divided by 100). Finally, the first column of the dataset was used to store the outcome labels, where 1 represented positive cases and -1 represented negative cases. These preprocessing steps helped make the data more consistent and suitable for training the model effectively.

Feature Extraction: Fourteen key variables commonly linked to diabetes risk were chosen, including family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, physical activity, smoking, alcohol use, education, and household income**.** The selection process followed an automated method developed by Chen et al., which identified relevant features based on predefined criteria. To further refine the selection, manual evaluation and parameter adjustments were conducted. The final set of variables was determined based on their ability to effectively distinguish between cases and non-cases, ensuring optimal performance for the model.

Training model: Support Vector Machine (SVM), a supervised machine learning algorithm, was used for classification by constructing an optimal hyperplane that maximizes the margin between two data clusters. Since real-world data is often not linearly separable, SVM uses kernel functions (linear, polynomial, sigmoid, and radial basis functions) to transform the input space into a higher-dimensional space where separation is possible. The LibSVM library was used to generate the models, and the grid.py utility helped fine-tune key parameters, C (controls misclassification tolerance) and gamma (controls model nonlinearity), using 5-fold cross-validation. Various kernel functions were tested, and the best-performing models were selected.
Key Aspects of SVM Implementation:

1. Misclassification Handling: A parameter C was introduced to balance minimizing errors and maximizing the margin.
2. Kernel Selection: Different kernel functions (linear, polynomial, sigmoid, and RBF) were tested to find the best fit for the data.

Additionally, multiple logistic regression (MLR) modeling was performed using the same selected variables for comparison, with probability estimations calculated for each case using SAS-callable SUDAAN version 9.

Part3: What did the authors find, what were the results? And, in your opinion, were those findings valid, why or why not?

Evaluate Performance: A 10-fold cross-validation was performed on the training dataset to assess model robustness. Performance metrics (AUC, sensitivity, specificity, PPV, and NPV) were averaged over all iterations. SVM and Logistic Regression showed comparable discriminative power. SVM is a flexible model-free method that performs well in multivariate settings. The Diabetes Classifier, a web-based tool, was implemented using the SVM model.

Classification Scheme	**Model**	**AUC (%)**	**Best Kernel (SVM)**	**P-Value**
Scheme I (Diagnosed/Undiagnosed Diabetes vs. No Diabetes/Pre-Diabetes)	SVM	83.47	RBF	0.3672
	Logistic Regression	83.19	N/A	
Scheme II (Undiagnosed Diabetes/Pre-Diabetes vs. No Diabetes)	SVM	73.18	Linear	0.6718
	Logistic Regression	73.35	N/A	

CRITIQUE:

Pre-processing: I loved that normalization was used here! SVM is very sensitive to unnormalized data, so standardizing all feature values to -1 to +1 ensures stability and improves model performance. Balancing cases and non-cases also prevents bias, making the dataset well-prepared for training.

Feature Extraction: The selection of 14 key diabetes risk variables is well-structured, following an automated approach by Chen et al., but more details on their method would have been helpful. Additionally, while manual refinement adds rigor, the specific criteria for adjustments and exclusions could have been better explained. Clarifying how each variable contributes to model performance would also strengthen the justification.

Training model: I loved the clear and simple explanation of SVM, especially the way it breaks down the role of C and gamma parameters in controlling overfitting and nonlinearity. The diagram illustrating feature transformation from a 2D input space to a 3D feature space was particularly effective in making the concept intuitive. However, a more detailed explanation of kernel functions would have been fantastic, as they are crucial to SVM's performance. While this explanation is great for a general audience, a data science-focused audience would benefit from more mathematical insights into hyperplane optimization and kernel mechanics.

Evaluate Performance: The evaluation seems overly complex, using multiple methods (test sets and cross-validation) without clear justification for both. I love how sensitivity and specificity were prioritized for diabetes prediction instead of just using accuracy, as they provide a more meaningful clinical perspective. The diagrams of AUC are great—they effectively illustrate both model performances. While SVM seems to offer a slight improvement over logistic regression, the added complexity may not be justified without a more significant performance boost. Moreover, the absence of test results comparing various kernel functions makes it difficult to assess the suitability of RBF and linear kernels for different classification tasks, highlighting the need for more comprehensive evaluations across a range of kernel options.Incorporating confidence intervals, hypothesis testing, and comparisons of multiple metrics (beyond just AUC) would provide a more comprehensive evaluation of model effectiveness.