Screening for pre-diabetes using support vector machine model

Jai Won Chung, Won Jae Kim, Soo Beom Choi, Jee Soo Park, Deok Won Kim, *Life member, IEEE**

***Abstract*—The global prevalence of diabetes is rapidly increasing. Studies support screening and interventions for pre-diabetes, which results in serious complications and diabetes. This study aimed at developing an intelligence-based screening model for pre-diabetes that could assist with decreasing the prevalence of diabetes through early identification and subsequent interventions. Data from the Korean National Health and Nutrition Examination Survey (KNHANES) were used, excluding subjects with diabetes. The KNHANES 2010 data (n = 4,685) were used for training and internal validation, while data from KNHANES 2011 (n = 4,566) were used for external validation. We developed a model to screen for pre-diabetes using support vector machine (SVM), and performed a systematic evaluation of the SVM model using internal and external validation. We compared the performance of the SVM model with that of a screening score model based on logistic regression analysis for pre-diabetes that had been developed previously. Backward elimination logistic regression resulted in associations between pre-diabetes and age, sex, waist circumference, body mass index, alcohol intake, family history of diabetes, and hypertension. The areas under the curves (AUCs) for the SVM model in the internal and external datasets were 0.761 and 0.731, respectively, while the AUCs for the screening score model were 0.734 and 0.712, respectively. The SVM model developed in this study performed better than the screening score model that had been developed previously and may be more effective for pre-diabetes screening.**

I. Introduction

The prevalence of type 2 diabetes is dramatically increasing, resulting in a global public health issue [1]. The prevalence of diabetes was estimated at 285 million or 6.4% of adults in the world in 2010 [2], and this prevalence is expected to rise to 552 million by 2030 [3]. The increasing rates of obesity are expected to result in a faster increase in the prevalence of type 2 diabetes in the future. However, owing to the absence of symptoms and/or disease-related knowledge, especially regarding the risk factors, diabetes often goes undetected, and approximately one-third of people with diabetes are not aware of their status [4]. Therefore, a simple, accurate screening method is needed. Historically, the majority of the clinical screening methods consisted of surveys developed using logistic regression analyses to predict diabetes [5], [6].

Pre-diabetes was first recognized as an intermediate diagnosis and indication of a relatively high risk for the future development of diabetes by the Expert Committee on Diagnosis and Classification of Diabetes Mellitus in 1997 [7], and it has been reported that approximately 5–10% of patients with untreated pre-diabetes subsequently develop diabetes [8]. This is significant considering that pre-diabetes was estimated to affect 4.9 million people, accounting for 17.4% of Korean adults, in 2005 [4], with a further 35% of adults in the US with pre-diabetes in 2008 [9]. The definition of pre-diabetes includes an impaired fasting plasma glucose (FPG) level in the range of 100–125 mg/dL (5.6–6.9 mmol/L) or impaired glucose tolerance (oral glucose tolerance test [OGTT] 2-h measurement in the range of 140–199 mg/dL [7.8–11.0 mmol/L]) [7]. Similar to diabetes, the risk of microvascular complications is increased with pre-diabetes [10], and the risk for cardiovascular disease and total mortality is almost twice as high in individuals with pre-diabetes [11]. Early diagnosis and intervention for pre-diabetes could prevent these complications, delay or prevent the transition to diabetes [1]-[3], and be cost-effective.

In this study, we aimed to develop and validate a model to predict pre-diabetes using support vector machine (SVM) method, which could be effective as simple and accurate screening tool. The SVM model performance was compared to that of the screening score model for pre-diabetes based on the screening score for diabetes by Lee et al. [5], with respect to accuracy and area under the curve (AUC) of the receiver operating characteristic (ROC).

II. Materials and Methods

A. Data source and subjects

Data from the Korean National Health and Nutrition Examination Survey (KNHANES) 2010 and 2011were used to develop and validate, respectively, the SVM model for pre-diabetes. The KNHANES is a cross-sectional survey that includes approximately 800 questions; it is conducted by the Division of Chronic Disease Surveillance, Korea Centers for Disease Control and Prevention.

The KNHANES 2010 included 8,958 subjects, for which the following exclusion criteria applied: <20 years of age [n = 2,293]; missing data for waist circumference, smoking status, alcohol intake, body mass index (BMI), physical activity, or family history of diabetes [n = 1,321]; or undetermined diabetes or hypertension status [n = 256]. Of the remaining 5,088 subjects, 403 subjects with diabetes and undiagnosed diabetes were also excluded from the study, resulting in 4,685 subjects who were included. A similar process was used with

J. W. Chung is with the Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Korea (e-mail: chjw0915@yuhs.ac).

W. J. Kim is with Dept. of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: jayasd36@hotmail.com).

S. B. Choi is with Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, Korea (e-mail: plains7@yuhs.ac).

J. S. Park is with Dept. of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: sampark@yuhs.ac).

*D. W. Kim is a Professor with the Department of Medical Engineering, Yonsei University College of Medicine, Seoul, Korea (phone: 82-2-2228-1916; fax: 82-2-364-1572; e-mail: kdw@yuhs.ac).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

the data from KNHANES 2011, which resulted in 4,566 subjects. The subjects of the KNHANES 2010 and 2011 data sets were not overlapped.

Fig. 1 illustrates the study flow. The development dataset from KNHANES 2010 was randomly divided into training and internal validation sets using a 2:1 ratio. The training set (n = 3,134) was used to construct the SVM model. The internal validation set (n = 1,551) was used to assess the ability to predict pre-diabetes. Additionally, data from KNHANES 2011 were used as an external validation set (n = 4,566). All individuals in the surveys participated voluntarily, and informed consent was obtained from all participants. The survey protocol was approved by the Institutional Review Board of the Korean Centers for Disease Control and Prevention.

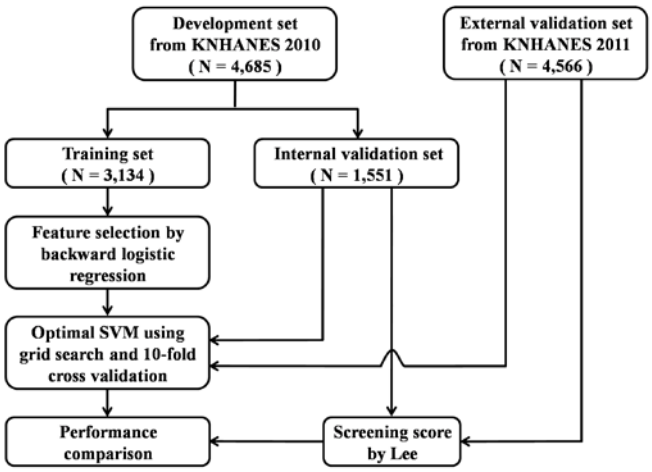

Figure 1. Chart depicting the flow of data from the KNHANES 2010 and 2011 to develop and validate, respectively, a pre-diabetes model.

B. Risk factors

We adopted the most frequently used 9 variables from previous studies regarding diabetes prediction models: age, gender, family history of diabetes, hypertension, alcohol intake, BMI, smoking status, waist circumference, and physical activity [5], [6]. FPG was determined using glucose levels that were collected following ≥8 hours of fasting . Pre-diabetes was defined as impaired FPG: 100 mg/dL ≤ FPG < 126 mg/dL. Diabetes was defined as FPG ≥126 mg/dL or non-fasting plasma glucose ≥200 mg/dL [7]. A family history of diabetes was limited to parents and siblings. Hypertension was defined as systolic blood pressure (SBP) >140 mmHg, diastolic blood pressure (DBP) >90 mmHg, or use of medication for blood pressure control [5]. Alcohol intake was calculated using 2 questions: (1) alcohol consumption frequency during the previous 12 months, and (2) average number of drinks on those days. The amount of alcohol was calculated based on the number of glasses, regardless of the kind of beverage, assuming that the amount of alcohol was approximately the same in each glass (approximately 8 g alcohol per glass). Smoking status was divided into "currently smoking regularly" and "others," with the latter group including subjects who had never smoked or had quit smoking. The subjects who answered more than "moderate" to the question "how intense is your everyday activity?" were considered as physically active.

C. Support vector machine model

SVM maps data to a higher dimensional space through a kernel function that linearly separates data patterns. The data are separated into 2 groups by the training data referred to as a support vector. SVM models are determined by choosing the maximum-margin hyper plane with the nearest support vector of the 2 groups [12]. SVM improves the accuracy of a model through the optimization of separating space using the kernel function, but one of the disadvantages of SVM is that it requires many trials to construct an optimal SVM model in comparison with other machine learning techniques [13].

The SVM was trained with 7 risk factors including age, gender, waist circumference, BMI, family history of diabetes, hypertension, and alcohol intake, which were selected using backward logistic regression. To obtain the optimal model, we adopted a grid search in which a range of parameter values (penalty parameter [C] of 0.01, 0.1, 1, 10, and 100 and scaling factor [σ] of 0.001, 0.01, 0.1, 1, 10, and 100) was tested using the 10-fold cross-validation strategy. The optimal parameter values with a C of 10 and σ of 10 for SVM using the Gaussian kernel function were obtained. The SVMs model were constructed using MATLAB Version 2012a (Mathworks Inc., Natick, MA).

D. Screening score of our models for pre-diabetes

The model constructed by SVM were compared with a previously developed screening survey to prove performance of our model and the possibility of their use in real situations. For this purpose, we used a screening score model for diabetes for the Korean population constructed by Lee et al. [5]; we felt this was appropriate because both studies constructed models for the Korean population. Lee et al. used data from KNHANES 2001 and 2005 for training and data from KNHANES 2007 and 2008 for external validation. In addition, the screening score model by Lee et al. used very similar risk factors to ours, with the exception of current smoking status. Those 6 variables independently associated with undiagnosed diabetes were chosen for their model: age, family history of diabetes, hypertension, waist circumference, smoking, and alcohol intake.

The risk score was assigned according to the odds ratio for each risk factor in the logistic regression model defined by Lee et al. [5]. Within the total score range of 0–11 points, a cut-off score of ≥5 points was selected to indicate an individual at high risk for undiagnosed diabetes; this cut-off resulted in the highest value for the Youden index. The 6 risk factors jointly yielded an AUC of 0.730 for both the internal and external validation sets [5]. To compare with our SVM model for pre-diabetes, we constructed a new screening score model for pre-diabetes by adjusting the cut-off point value based on our definition of pre-diabetes (100 mg/dL≤ FPG <126 mg/dL), given that the screening score for diabetes used by Lee et al. was based on FPG ≥126 mg/dL [5]. The screening score for pre-diabetes was designed with the same

risk score model of the 6 risk factors using our training set for pre-diabetes (KNHANES 2010) and the Youden index; as a result, a cut-off score of ≥5 points was identified to indicate an individual with pre-diabetes.

E. Statistical analyses

Characteristics of the data from the KNHANES 2010 in different pre-diabetes statuses are summarized by descriptive statistics. For comparison between normal glucose tolerance and impaired fasting glucose, the continuous and categorical characteristics were tested using t-test and chi-square test, respectively.

To obtain the optimal variables for the prediction model, backward logistic regression was performed with the training set. Each step of the backward regression excluded the variables without a statistically meaningful correlation with the outcome, pre-diabetes. Three steps of backward regression were executed, and the selected 7 variables were age, BMI, hypertension, gender, alcohol intake, waist circumference, and family history of diabetes. The Hosmer-Lemeshow test resulted in a p-value of 0.132, indicating that the chosen variables were well-fitted.

ROC curve analysis is the most commonly used method in clinical analysis to establish an optimal cut-off point [14]. Therefore, we generated ROC curves and the selected cut-off points that maximized the Youden index [15] to compare the performance of our optimal SVM model with that of the screening score model for pre-diabetes based on the screening score by Lee et al. [5], using our internal and external validation sets. Following the ROC analysis, the AUC, accuracy, sensitivity, and specificity of our SVM model and screening score model for pre-diabetes were calculated. We used SPSS 20.0 (IBM Corp, Armonk, NY) for statistical analysis and MedCalc 12.4 (MedCalc Inc., Mariakerke, Belgium) for ROC analysis. Statistical significance was set at $p < 0.05$.

III. Result

A. Characteristics of the development dataset

The characteristics of the KNHANES 2010 data are summarized in Table I. The variables that were significantly related to pre-diabetes were age, gender, family history of diabetes, alcohol intake, BMI, waist circumference, FPG, systolic and diastolic blood pressures, and hypertension.

B. Performance of the SVM model

Cross-validation of the optimal SVM parameters with the training set resulted in an AUC of 0.742 and accuracy of 69.9%. These results are not included in Table II. The similar performance observed between the training and validation sets indicates that the trained model was not over-fitting.

With both the internal and external validation sets, our SVM model showed better performance than the existing screening score model using logistic regression, especially in terms of AUC, which is known as a better predictor than accuracy in evaluating learning algorithms [16] (Table II).

TABLE I. Characteristics of the data from the KNHANES 2010

Variable*	Normal glucose tolerance (n=3,681)	Impaired fasting glucose (n=1,004)	p†
Age (years)	46.0 (15.2/0.3)	55.6 (13.3/0.4)	< 0.001
Gender (% of men)	39.3	54.0	< 0.001
Family history of diabetes (%)	18.4	21.2	0.048
Current smoker (%)	21.3	21.2	0.940
Alcohol intake (drinks/day)	0.7 (1.2/0.0)	0.9 (1.4/0.0)	< 0.001
Physically active (%)	50	51.9	0.284
Body mass index (kg/m^2)	23.1 (3.2/0.1)	24.9 (3.3/0.1)	< 0.001
Waist circumference (cm)	79.3 (15.1/0.2)	85.5 (9.1/0.3)	< 0.001
Fasting plasma glucose (mg/dL)	89.0 (6.0/0.1)	107.5 (6.7/0.2)	< 0.001
Systolic blood pressure (mmHg)	115.3 (16.8/0.3)	125.6 (17.0/0.5)	< 0.001
Diastolic blood pressure (mmHg)	73.8 (10.4/0.2)	77.9 (10.4/0.3)	< 0.001
Hypertension (%)	19.9	43.6	< 0.001

* Table values are given as mean (standard deviation/standard error) or %.
† p were obtained by t-test and chi-square test.

TABLE II. Performance of SVM and screening score by Lee for internal and external validation sets for predicting pre-diabetes

Dataset	Screening method	AUC	Acc. (%)	Sen. (%)	Spec. (%)
Internal validation set (n=1,551)	SVM	0.761	64.9	78.9	61.2
	Screening score	0.734	63.4	76.1	60.0
External validation set (n=4,566)	SVM	0.731	66.1	69.4	65.3
	Screening score	0.712	59.9	74.3	56.4

AUC : area under the curve, Acc. : Accuracy, Sen. : sensitivity, Spec. : specificity, SVM : support vector machine

IV. Discussion and Conclusion

The results of the present study indicate that the SVM model that we developed to predict pre-diabetes, defined as impaired FPG, performed better than the existing clinical screening score model, as indicated by the AUC and accuracy measures (Table III). The SVM model performed particularly well due to the ability of SVM to efficiently find a unique optimal solution, incorporate multiple types of data with a degree of flexibility, and model nonlinear patterns [17].

Although similar statistical analyses were conducted (i.e., backward regression models), there were slight differences in the variables included in the present study and those in the

study by Lee et al [5]. Lee et al. included current smoking status as a risk factor in the training set based on the data from KNHANES 2001 and 2005; however, current smoking status was not included in our training set using data from KNHANES 2010. This may have resulted from lifestyle changes in the Korean population between those years, including a decline in the overall smoking rate and stronger anti-smoking laws [18].

Although several screening score models have been developed and used clinically, our prediction model is unique in several ways. First, owing to the similarity between our SVM model and the existing screening score model, we were able to compare the performance of our SVM model with the existing model. Second, to the best of our knowledge, there are very few studies investigating pre-diabetes; instead, the majority of the other models have been developed to predict undiagnosed diabetes. However, pre-diabetes is increasingly becoming a significant public health issue. Using our model to screen patients for pre-diabetes would enable interventions at an earlier stage, which would be easier to implement and more successful than interventions implemented following diabetes screening.

The SVM model that we developed is limited in terms of convenience and potential widespread use. Although the screening score model is not the most effective one for disease prediction, it is simple and accessible. However, SVM model could also become more accessible through the use of calculator software, particularly with the widespread use of devices such as computers, smart phones, and tablet PCs. Future studies could develop a calculator in which the values are entered via a website or application and the results are immediately delivered to the end-user.

Our study constructed a reasonably good model to predict pre-diabetes in the Korean population. By applying similar methods in other countries, researchers could develop country-specific machine learning models for nationwide use. The creation of a user-friendly calculator program would enable access to screening by the general population, in addition to medical professionals. This widespread use could result in early diagnosis and treatment for people with pre-diabetes and diabetes, helping to relieve the public health diabetes burden and reducing the number of people who remain undiagnosed.

REFERENCES

[1] B.C. Zyriax, R. Salazar, W. Hoeppner, E. Vettorazzi, and C. Herder et al., "The Association of Genetic Markers for Type 2 Diabetes with Prediabetic Status-Cross-Sectional Data of a Diabetes Prevention Trial," *PLoS One*, vol. 8, no. 9, pp. e75807, Sep. 2013.

[2] J. Shaw, R. Sicree, and P. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes Res. Clin. Pract.*, vol. 87, no. 1, pp. 4-14, Jan. 2010.

[3] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes Res. Clin. Pract.*, vol. 94, no. 3, pp. 311-321, Dec. 2011.

[4] Y. J. Choi, H. C. Kim, H. M. Kim, S. W. Park, and J. Kim et al., "Prevalence and Management of Diabetes in Korean Adults Korea National Health and Nutrition Examination Surveys 1998–2005," *Diabetes Care*, vol. 32, no. 11, pp. 2016-2020, Nov. 2009.

[5] Y. H. Lee, H. Bang, H. C. Kim, H. M. Kim, and S. W. Park et al., "A Simple Screening Score for Diabetes for the Korean Population Development, validation, and comparison with other scores," *Diabetes Care*, vol. 35, no. 8, pp. 1723-1730, Aug. 2012.

[6] J. Lindström, and J. Tuomilehto, "The Diabetes Risk Score A practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725-731, Mar. 2003.

[7] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. Supplement 1, pp. S81-S90, Jan. 2014.

[8] Diabetes Prevention Program (DPP) Research Group, "The Diabetes Prevention Program (DPP) description of lifestyle intervention," *Diabetes Care*, vol. 25, no. 12, pp. 2165-2171, Dec. 2002.

[9] T. L. Sentell, G. He, E. W. Gregg, and D. Schillinger, "Racial/ethnic variation in prevalence estimates for United States prediabetes under alternative 2010 American Diabetes Association criteria: 1988-2008," *Ethn. Dis.*, vol. 22, no. 4, pp. 451, 2012.

[10] A. G. Tabak, C. Herder, W. Rathmann, E. J. Brunner, and M. Kivimäki, "Prediabetes: a high-risk state for diabetes development," *Lancet*, vol. 379, no. 9833, pp. 2279-2290, Jun. 2012.

[11] M. Coutinho, H. C. Gerstein, Y. Wang, and S. Yusuf, "The relationship between glucose and incident cardiovascular events. A metaregression analysis of published data from 20 studies of 95,783 individuals followed for 12.4 years," *Diabetes Care*, vol. 22, no. 2, pp. 233-240, Feb. 1999.

[12] C. Cortes, and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, Mar. 1995.

[13] K. A. Kim, J. Y. Choi, T. K. Yoo, S. K. Kim, and K. Chung et al., "Mortality prediction of rats in acute hemorrhagic shock using machine learning techniques," *Med. Biol. Eng. Comput.*, vol. 51, no. 9, pp. 1059-1067, Sep. 2013.

[14] T. K. Yoo, S. K. Kim, D. W. Kim, J. Y. Choi, and W. H. Lee, "Osteoporosis Risk Prediction for Bone Mineral Density Assessment of Postmenopausal Women Using Machine Learning," *Yonsei Med. J.*, vol. 54, no. 6, pp. 1321-1330, Nov. 2013.

[15] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its associated cutoff point," *Biom. J.*, vol. 47, no. 4, pp. 458-472, Aug. 2005.

[16] J. Huang, and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299-310, Mar. 2005.

[17] R. C. Thurston, K. A. Matthews, J. Hernandez, and F. De La Torre, "Improving the performance of physiologic hot flash measures with support vector machines," *Psychophysiology*, vol. 46, no. 2, pp. 285-292, Mar. 2009.

[18] Y. H. Khang, S. C. Yun, H. J. Cho, and K. Jung-Choi, "The impact of governmental antismoking policy on socioeconomic disparities in cigarette smoking in South Korea," *Nicotine Tob. Res.*, vol. 11, no. 3, pp. 262-269, Mar. 2009.