BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

Accurate and rapid screening model for potential diabetes mellitus

Dongmei Pei[1], Yang Gong[2], Hong Kang[2], Chengpu Zhang[1] and Qiyong Guo[3*]

Abstract

Background: Prediction or early diagnosis of diabetes is crucial for populations with high risk of diabetes.

Methods: In this study, we assessed the ability of five popular classifiers (J48, AdaboostM1, SMO, Bayes Net, and Naïve Bayes) to identify individuals with diabetes based on nine non-invasive and easily obtained clinical features, including age, gender, body mass index (BMI), hypertension, history of cardiovascular disease or stroke, family history of diabetes, physical activity, work stress, and salty food preference. A total of 4205 data entries were obtained from annual physical examination reports for adults in the Shengjing Hospital of China Medical University during January–April 2017. Weka data mining software was used to identify the best algorithm for diabetes classification.

Results: The results indicate that decision tree classifier J48 has the best performance (accuracy = 0.9503, precision = 0.950, recall = 0.950, *F*-measure = 0.948, and AUC = 0.964). The decision tree structure shows that age is the most significant feature, followed by family history of diabetes, work stress, BMI, salty food preference, physical activity, hypertension, gender, and history of cardiovascular disease or stroke.

Conclusions: Our study shows that decision tree analyses can be applied to screen individuals for early diabetes risk without the need for invasive tests. This procedure will be particularly useful in developing regions with high epidemiological risk and poor socioeconomic status, and enable clinical practitioners to rapidly screen patients for increased risk of diabetes. The key features in the tree structure could further facilitate diabetes prevention through targeted community interventions, which can potentially improve early diabetes diagnosis and reduce burdens on the healthcare system.

Keywords: Data mining, Diabetes, Screening

Background

The worldwide incidence of diabetes rose from 108 million in 1980 to 422 million in 2014, and could potentially be the seventh-leading cause of death in 2030 [1]. However, half of the patients with diabetes are unaware of their disease. The incidence of diabetes (100 million adult patients) in China was the highest worldwide in 2015, whereas 52.7% of these patients (50 million) are undiagnosed [2, 3]. Hence, early detection and prevention of diabetes is a severe challenge in China.

The American Diabetes Association recommends annual screening for diabetes in patients older than 45 years and in younger patients with major risk factors [4]. China's National Plan for Non-Communicable Diseases Prevention and Treatment (2012–2015) identified diabetes as one of the priority diseases in China, and proposed several recommendations to predict diabetes based on blood glucose tests and routine physical examinations [5].

The main challenge in screening for diabetes is economic, including expensive blood work and additional human labor, which is even more challenging in developing countries [6]. The World Health Organization recommends that simple strategies should be developed to identify patients with risk for diabetes and then implement early lifestyle interventions [7]. To achieve these recommendations, it is crucial to develop a simple and accurate diabetes screening method.

* Correspondence: guoqy1111@hotmail.com
[3]Department of radiology, Shengjing Hospital, China Medical University, Shenyang, Liaoning, China
Full list of author information is available at the end of the article

Developing appropriate disease prediction algorithms can be technically challenging. In a Brazilian investigation, Lélis et al. [6] applied seven classification techniques to make a diagnosis of meningococcal meningitis and demonstrated this model is accurate and affordable. Choi et al. [8] developed two models to screen for prediabetes of 9251 individuals using an artificial neural network (ANN) and support vector machine (SVM) and performed a systematic evaluation of the models using internal and external validation, and concluded that the SVM model is superior to the ANN model in the screening for prediabetes. In another Brazilian study, Olivera et al. [9] utilized and compared machine-learning algorithms to develop predictive models using data from ELSA-Brasil and found that most of these predictive models yielded similar results and demonstrated the feasibility of identifying individuals with highest risk of having undiagnosed diabetes through easily-obtained clinical data. Data mining and machine learning are analytical methods that leverage artificial intelligence to identify patterns in large data sets, make decisions with minimal human intervention, and build models. There is considerable interest in determining how different classification techniques from machine learning can be utilized as disease prediction tools [10–19]. These tools have been used to diagnose diabetes [3, 8–10, 20–22], meningitis [6], glaucoma [11], asthma [12], coronary artery disease [13], cancer [14–17, 23], tuberculosis [18, 24], hypertension [25], and heart arrhythmia [26].

The objective of this study is to use easily obtained and directly observable clinical data to construct a predictive model to identify patients with increased risk for diabetes. Specifically, we utilize data mining and machine learning to develop an accurate diabetes classifier that can rapidly screen clinical data. Our approach will be particularly useful in locations with high epidemiological risk and poor socioeconomic status, where patients cannot afford medical laboratory costs [6]. Rapid identification of patients with high diabetes risk can help to avoid disease progression and prevent the incidence of disease complications.

Methods

Study population

A total of 8452 annual physical examination reports between January 2017 and April 2017 were collected from the electronic health records database in Shengjing Hospital of China Medical University, located in the center of Liaoning Province in China. We adopted the nine most frequently used features from previous studies of diabetes prediction models [8, 20, 27–30]. These features are either directly observable or easily obtained without expensive and invasive tests. Approval for this study was obtained from the Shengjing Hospital (reference number 2017PS42K).

The nine features included age, gender, body mass index (BMI), hypertension, history of cardiovascular disease or stroke, family history of diabetes, physical activity, work stress, and salty food preference (eating the salty meat or fish 4–7 times a week). Among 8452 records, a total of 3956 records were excluded due to missing data for BMI, blood pressure, family history of diabetes, history of cardiovascular disease or stroke, physical activity, work stress, or salty food preference. Records with past history of diabetes (291 records) also were excluded because we focused on predicting prediabetes and diabetes. Finally, a total of 4205 records were included in this study as shown in Fig. 1.

Data collection and transformation

The nine features were characterized for data analysis. Age and gender were demographic characteristics. Family history of diabetes was defined as any family member previously diagnosed by a physician as diabetic or prediabetic (Yes = 1, No = 0). BMI was calculated as body weight divided by the square of height in meters and BMI ≥ 25 was defined as overweight. History of cardiovascular disease or stroke was defined as the patient previously diagnosed with coronary heart disease or stroke by a physician (Yes = 1, No = 0). Physical activity indicated if the patient engaged in more than 30 min of exercise 3 days a week (More = 1, Less = 0). Work stress was grouped into three levels according to the patients' subjective impression (High = 2, Moderate = 1, Low = 0). Salty food preference (salty meat or fish) indicated if the person preferred salty food for 4–7 days a week (Yes = 1, No = 0).

BMI and hypertension were defined and measured as below. BMI was calculated as weight in kilograms divided by the square of height in meters (kg/m2); BMI ≥ 25 was defined as overweight. Hypertension was defined as systolic blood pressure ≥ 140 mmHg, or diastolic blood pressure ≥ 90 mmHg, and/or use of medication for blood pressure control.

Each report included a diagnosis (diabetes or normal) based on fasting plasma glucose. Diabetes diagnoses included prediabetes and type 2 diabetes, and was defined as fasting plasma glucose ≥5.6 mmol/L [8, 20].

Variable characteristics

After data preparation and transformation, the final database consisted of 4205 records and 10 variables. These 10 variables included 9 input variables and one target variable. The target variable consisted of two classes: one class was the diagnosis of diabetes, the other class was normal. The characteristics of participants and chi-square test results between two groups are presented

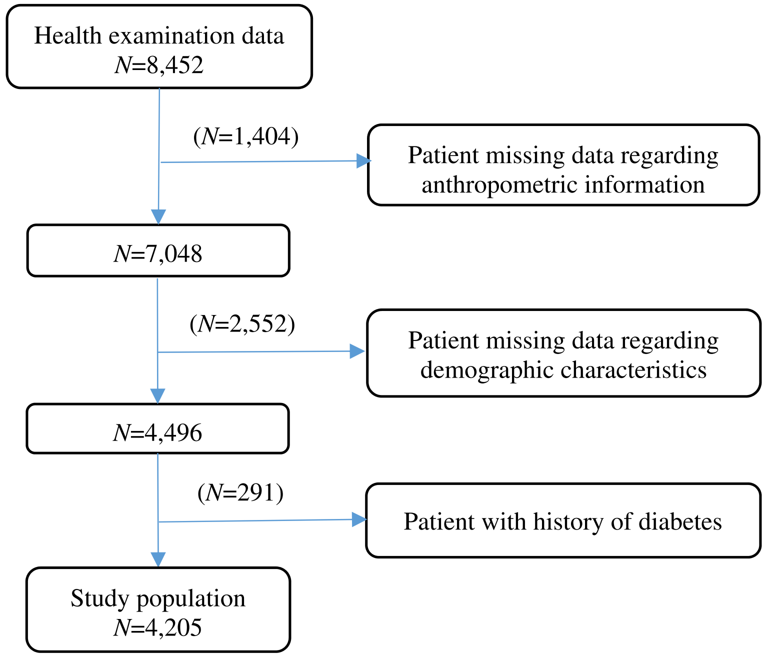
Health examination data
N=8,452
(N=1,404)
Patient missing data regarding anthropometric information
N=7,048
(N=2,552)
Patient missing data regarding demographic characteristics
N=4,496
(N=291)
Patient with history of diabetes
Study population
N=4,205

Fig. 1 Flow chart of records that were excluded from the physical examination database of Shengjing Hospital of China Medical University (January–April, 2017)

in Table 1. There were statistically significant differences in the nine features between the two groups, at a significance level of 0.05.

Classifier comparison

We applied five popular classifiers to train the dataset, including J48 (class for generating a pruned or unpruned), AdaboostM1 (method for boosting a nominal class classifier), SMO (implements John Platt's sequential minimal optimization algorithm for training a support vector classifier), Bayes Net (Bayes network learning method that implements a hill climbing algorithm restricted by an order on the variables), and Naïve Bayes (class for a naïve Bayes classifier using estimator classes). Weka software (version 3.8; University of Waikato, Hamilton, NZ) [6, 16] was used to assess the classifiers and identify the best algorithm for diabetes classification. To avoid over-fitting and unnecessary complexity, the decision tree created by the J48 algorithm was pruned by removing nonessential terminal branches. This pruning method was based on defined algorithms and did not affect the classification accuracy [6, 21, 25].

Classifier accuracy and performance evaluation

The entire dataset was randomly divided into two parts: the training set consisted of 70% of the data for model development, and the test set consisted of the remaining data (30%) for model validation [21, 31]. The algorithms were compared based on accuracy, precision, recall, *F*-measure, and the area under the receiver operating characteristic (ROC) curve (AUC), and the best-performing algorithm was selected [21, 32]. Eqs. 1–2 were used to calculate the accuracy, precision, recall, and *F*-measure, respectively.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{3}$$

$$Precision = TP/(TP + FN) \tag{4}$$

$$Recall = TP/(TP + FP) \tag{5}$$

$$F{-}measure = 2/(1/Precision) + (1/Recall) \tag{6}$$

The AUC summarizes ROC curves by indicating whether the classifier is more likely to distribute the score as positive rather than the randomly selected negative sample. Better models have larger AUC values. The relative accuracy of the classification test is graded according to the following scale [18]: Excellent = 0.90–1; Good = 0.80–0.90; Fair = 0.70–0.80; Poor = 0.60–0.70; Fail = 0.50–0.60.

Results

A total of 4205 records (2734 females and 1471 males) were selected for this analysis, which included 709 (16.86%) diabetes diagnoses and 3496 (83.14%) normal patients. Table 2 presents the performance of all classifiers, and shows that J48 exhibits better results than others (accuracy = 0.9503, precision = 0.950, recall = 0.950,

Table 1 Characteristics of variables in diabetes and normal groups

Variable	Possible values	Diabetes *N* = 709	Normal *N* = 3496	*p*-value	χ^2 test
Age	20–34 years old	36 (5.1%)	1718 (49.1%)	< 0.001	269.33
	35–49 years old	207 (29.2%)	1246 (35.7%)		
	50–65 years old	466 (65.7%)	532 (15.2%)		
Gender	Male	348 (49.1%)	1123 (32.1%)	< 0.001	16.25
	Female	361 (50.9%)	2373 (67.9%)		
Body mass index	< 25	250 (35.3%)	2806 (80.3%)	< 0.001	18.87
	≥25	459 (64.7%)	690 (19.7%)		
Hypertension	Yes	221 (31.2%)	755 (21.6%)	< 0.001	15.22
	Non-hypertension	488 (68.8%)	2741 (78.4%)		
Salty food preference	No	384 (54.2%)	2598 (74.3%)	< 0.001	9.33
	Yes	325 (45.8%)	898 (25.7%)		
History of cardiovascular disease or stroke	No	627 (88.4%)	3190 (91.2%)	0.018	122.25
	Yes	82 (11.6%)	306 (8.8%)		
Family history of diabetes	No	335 (47.2%)	3133 (89.6%)	< 0.001	154.21
	Yes	374 (52.8%)	363 (10.4%)		
Physical activity	Less	542 (76.4%)	2043 (58.4%)	< 0.001	33.68
	More	167 (23.6%)	1453 (41.6%)		
Work stress	Low	129 (18.2%)	1054 (30.2%)	< 0.001	81.54
	Moderate	353 (49.8%)	1993 (57.0%)		
	High	227 (32.0%)	449 (12.8%)		

F-measure = 0.948, and AUC = 0.964). Figure 2 presents the ROC curves of all classifiers.

The final tree contains 18 nodes and 19 leaves, as shown in Fig. 3.

The decision tree shows that age was assigned by as the first and most informative node, followed by family history of diabetes, work stress, BMI, salty food preference, physical activity, hypertension, gender, and history of cardiovascular disease or stroke. Most leaves in the left half of the decision tree (≤49 years old) were classified as normal, whereas most leaves in the right half of the decision tree (> 49 years old) were classified as diabetes.

The decision tree can be converted into a set of if-then rules by tracing the path from the root node to each terminal (leaf) node. The if-then rules created by the model are presented in Table 3.

Discussion

In this study, we employed data mining and machine learning to examine the performance of five classifiers (J48, AdaboostM1, SMO, Bayes Net, and Naïve Bayes) and nine non-invasive and easily obtained clinical features (age, gender, BMI, hypertension, history of cardiovascular disease or stroke, family history of diabetes, physical activity, work stress, and salty food preference) for the rapid and accurate identification of individuals with diabetes. The best classifier was trained with the decision tree generated by the J48 algorithm, which had accuracy = 0.9503, precision = 0.950, recall = 0.950, *F*-measure = 0.948, and AUC = 0.964. The results indicate that this strategy successfully achieves accurate and rapid diabetes screening. This approach can be applied for non-invasive prediction of prediabetes and diabetes without the need for expensive lab tests. Thus, this test could be particularly useful in regions with high epidemiological risk and low socioeconomic status.

Decision trees are powerful classification algorithms used in parallel with data mining methods [20, 21, 24, 33, 34]. The first variable (root) in the tree is the most important factor, whereas consecutively distant variables further from the root are ranked in order as less

Table 2 The results of classification algorithms

Model	Accuracy	Precision	Recall	*F*-Measure	AUC
AdboostM1	0.9127	0.908	0.913	0.906	0.933
J48	0.9503	0.950	0.950	0.948	0.964
SMO	0.9078	0.903	0.908	0.900	0.763
Naïve Bayes	0.8934	0.886	0.893	0.888	0.922
Bayes Net	0.8878	0.881	0.888	0.883	0.924

AUC the area under the receiver operating characteristic (ROC) curve

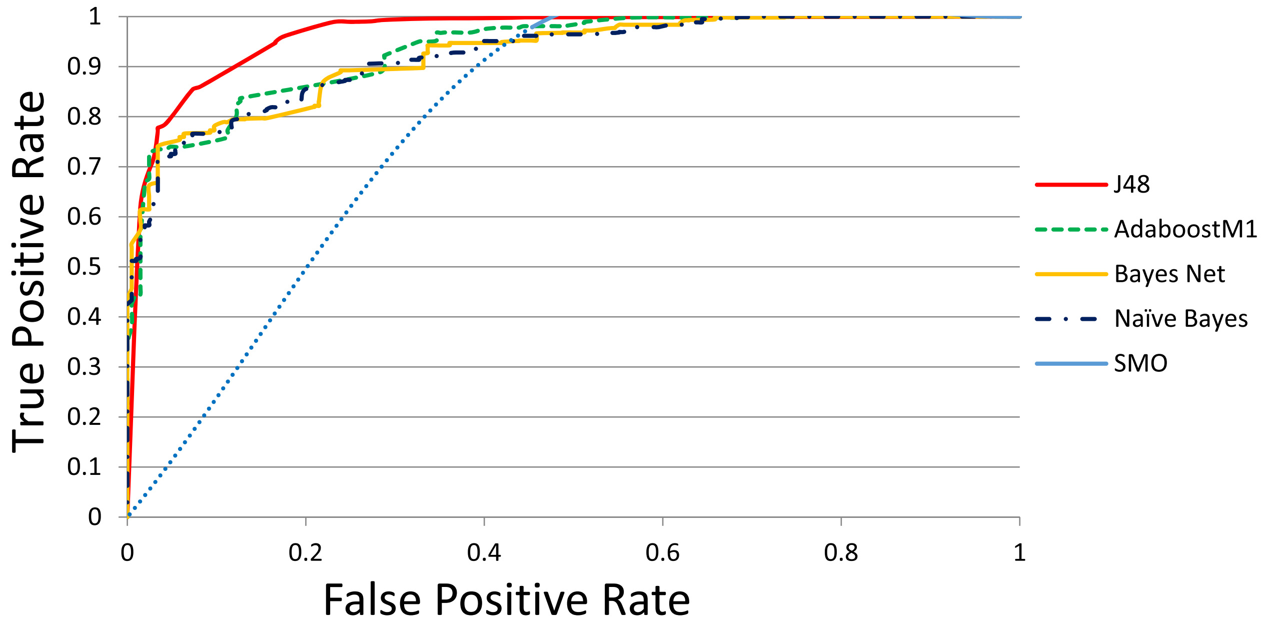

Fig. 2 ROC curve of all algorithms

important factors for data classification [21]. This study shows that age is the most important attribute discriminating between those with and without diabetes. Age is followed by family history of diabetes, work stress, BMI, salty food preference, physical activity, hypertension, gender, and history of cardiovascular disease or stroke. These results are consistent with those reported in previous studies [20, 35, 36].

The decision tree shows that family history of diabetes, work stress and BMI are the following important factors after age. The tree identified a subgroup of individuals [1457 patients (99%)] with age ≤ 49, without a family

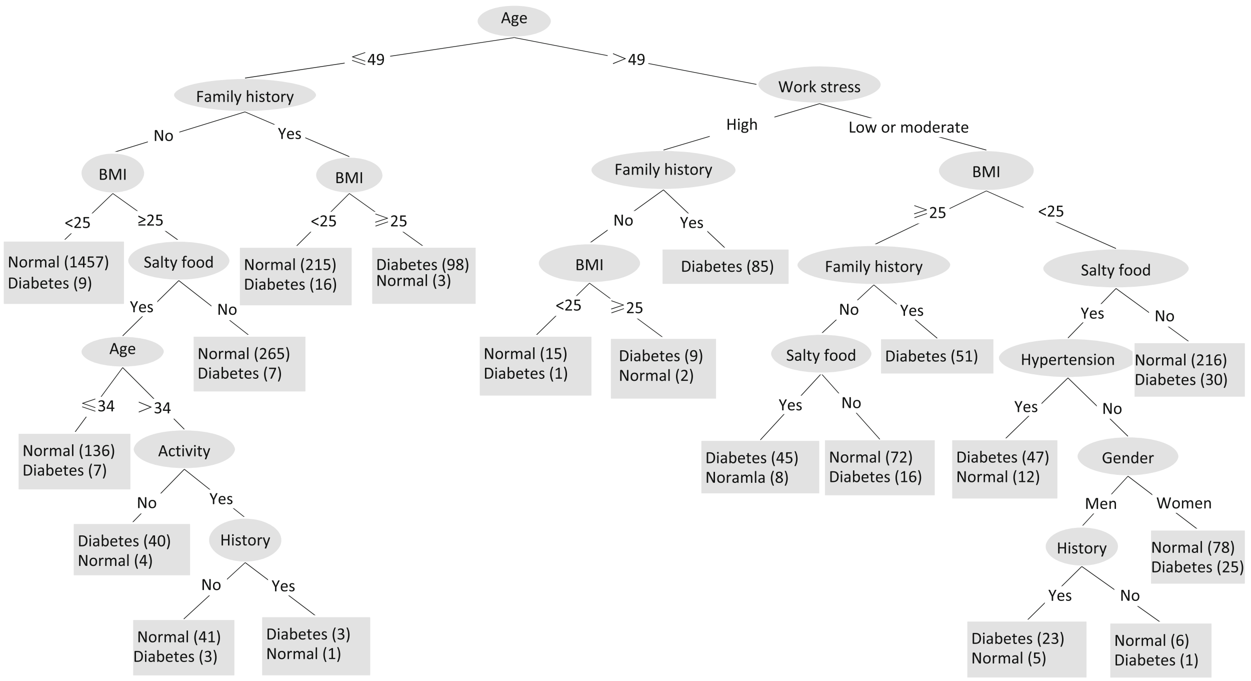

Fig. 3 Decision tree of diabetes classifiers. The sample size is given as the number in parentheses at each node

Table 3 Nineteen if-then rules extracted from the decision tree in Fig. 3

Rule 1: IF age ≤ 49, without a family history of diabetes, BMI ≤ 25, THEN patient is normal (1457/1466 or 99%)

Rule 2: IF age ≤ 34, without a family history of diabetes, BMI > 25, prefers salty food, THEN patient is normal (136/143 or 95%)

Rule 3: IF 35 < age ≤ 49, without a family history of diabetes, BMI > 25, prefers salty food, without physical activity, THEN patient is diabetic (40/44 or 91%)

Rule 4: IF 35 < age ≤ 49, without a family history of diabetes, BMI > 25, prefers salty food, with physical activity, without history of cardiovascular disease or stroke, THEN patient is normal (41/44 or 93%)

Rule 5: IF 35 < age ≤ 49, without a family history of diabetes, BMI > 25, prefers salty food, with physical activity, with history of cardiovascular disease or stroke, THEN patient is diabetic (3/4 or 75%)

Rule 6: IF age ≤ 49, without a family history of diabetes, BMI > 25, without preference for salty food, THEN patient is normal (265/272 or 97%)

Rule 7: IF age ≤ 49, with a family history of diabetes, BMI ≤ 25, THEN patient is normal (215/231 or 93%)

Rule 8: IF age ≤ 49, with a family history of diabetes, BMI > 25, THEN patient is diabetic (98/101 or 97%)

Rule 9: IF age > 49, with work stress high, without a family history of diabetes, BMI ≤ 25, THEN patient is normal (15/16 or 94%)

Rule 10: IF age > 49, with work stress high, without a family history of diabetes, BMI > 25, THEN patient is diabetic (9/11 or 82%)

Rule 11: IF age > 49, with work stress high, with a family history of diabetes, THEN patient is diabetic (85 or 100%)

Rule 12: IF age > 49, with work stress low or moderate, BMI > 25, without a family history of diabetes, prefers salty food, THEN patient is diabetic (45/53 or 85%)

Rule 13: IF age > 49, with work stress low or moderate, BMI > 25, without a family history of diabetes, without preference for salty food, THEN patient is normal (72/88 or 82%)

Rule 14: IF age > 49, without work stress high, BMI > 25, with a family history of diabetes, THEN patient is diabetic (51 or 100%)

Rule 15: IF age > 49, with work stress low or moderate, BMI ≤ 25, prefers salty food, with hypertension, with work stress, THEN patient is diabetic (47/59 or 80%)

Rule 16: IF age > 49, with work stress low or moderate, BMI ≤ 25, prefers salty food, without hypertension, gender male, with history of cardiovascular disease or stroke, THEN patient is diabetic (23/28 or 82%)

Rule 17: IF age > 49, with work stress low or moderate, BMI ≤ 25, prefers salty food, without hypertension, gender male, without history of cardiovascular disease or stroke, THEN patient is normal (6/7 or 86%)

Rule 18: IF age > 49, with work stress low or moderate, BMI ≤ 25, prefers salty food, without hypertension, gender female, THEN patient is normal (78/103 or 76%)

Rule 19: IF age > 49, with work stress low or moderate, BMI ≤ 25, without preference for salty food, THEN patient is normal (216/246 or 88%)

history of diabetes and BMI ≤ 25 that were normal cases. Another subgroup of individuals [98 patients (97%)] with age ≤ 49, with a family history of diabetes, BMI > 25 that were identified as diabetes cases. A subgroup of individuals [85 patients (100%)] with age > 49 and work stress high, with a family history of diabetes were identified as diabetes cases (Table 3). These key features could facilitate diabetes prevention through community interventions. Several large-scale trials have demonstrated the benefits of preventing diabetes with targeted lifestyle interventions [20, 37–39]. By reducing these risk factors, would be rewarded as Therefore, patients who are at a high risk of developing diabetes could be targeted to reduce established risk factors and provide educational programs, which will reduce the public health burden and the number of undiagnosed individuals [8, 40].

A major strength of this study is that we used a real medical dataset of annual physical examinations from Shengjing Hospital of China Medical University. All subjects were subjected to laboratory glucose tests to diagnose prediabetes or diabetes, so the results were more reliable than if the individuals were diagnosed by self-reporting. In 2014, Shengjing Hospital received the Stage Seven award from the Healthcare Information and Management Systems Society for successful implementation of electronic health records and rapid sharing of clinical information via standardized electronic transactions, data warehousing, and data continuity with the emergency department and other ambulatory care departments. Shengjing Hospital routinely collects and stores a large amount of data in the electronic hospital records. We used data mining, machine learning, and knowledge discovery capabilities to identify potential data patterns and specific features containing enough information to increase the accuracy of diabetes predictions [10, 41]. A large-scale study conducted in Iran compared different classification algorithms in the diagnosis of type 2 diabetes and demonstrated that it is therefore highly recommended that the choice and selection of features for data mining applications in disease diagnosis, be done by the help and advice of experts to obtain the best possible results. Artificial neural network is the most accurate method of classification with an accuracy of 97.18% [42].

In the future, we will test the model and develop prediction models with more sensitivity and specificity. We will focus on applying similar methods in different populations using more data. When the amount of data increases, the results will be more robust [17]. Our approach can be extended to larger databases that store more variables and risk factors related to diabetes [22]. The results of these studies could provide novel evidence-based prevention and treatment strategies. Clinical researchers can help to establish new priorities for further analyses by diabetes researchers.

Limitations

Our study has two limitations. Data were collected from only one large hospital in China. Further studies with additional data from this hospital and other centers need to be performed. This was a cross-sectional design study. The results should be confirmed in a prospective study.

Conclusion

We utilized data mining classifiers and machine learning to generate a decision tree that identified potential prediabetes and diabetes in clinical data extracted from annual health examination reports in a large Chinese hospital. We assessed the classifiers using nine clinical features that were easily obtained and non-invasive. The J48 classifier had the best performance, and indicates that decision tree analyses can be successfully applied to rapidly and accurately screen for diabetes in clinical practice. This type of work is essential in regions with high epidemiological risk and low socioeconomic status. The tree structure identifies the most important risk factors, and suggests that diabetes prevention programs could be applied through targeted community interventions. This would help improve early diabetes diagnosis and reduce burdens on the healthcare system.

Abbreviations

AUC: The area under the receiver operating characteristic curve; BMI: Body mass index; SMO: Sequential minimal optimization

Acknowledgements

Not applicable.

Funding

China Medical Board (Grant number 15–219). The funding bodies play a role in the design of the study and collection, analysis of data and in writing the manuscript.

Availability of data and materials

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

QG and YG had the initial conception of the idea and background for this study. DP and HK made contributions in the acquisition and analysis of the data. CZ preformed the literature search. All authors contributed to the writing, reviewing and final approval of the manuscript.

Ethics approval and consent to participate

Ethics approval was obtained by Shengjing Hospital of China Medical University Ethics Committee (ref. Ethics 2017PS42K) without requirement of consent for participation.

Consent for publication

Not applicable.

Competing interests

All authors declare that they have no conflicts of interests.

Publisher's Note

Author details

[1]Department of Family Medicine, Shengjing Hospital, China Medical University, Shenyang, Liaoning, China. [2]University of Texas Health Science Center at Houston, Houston, Texas, USA. [3]Department of radiology, Shengjing Hospital, China Medical University, Shenyang, Liaoning, China.

Received: 28 February 2018 Accepted: 3 March 2019
Published online: 12 March 2019

References

1. World Health Organization, Global report on diabetes 2016. http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf.
2. Hadaegh F, et al. High prevalence of undiagnosed diabetes and abnormal glucose tolerance in the Iranian urban population: Tehran lipid and glucose study. BMC Public Health. 2008;8:176.
3. Jahani M, Mahdavi M. Comparison of predictive models for the early diagnosis of diabetes. Healthc Inform Res. 2016;22(2):95–100.
4. Pippitt K, Li M, Gurgle HE. Diabetes mellitus: screening and diagnosis. Am Fam Physician. 2016;93(2):103–9.
5. The National Guideline for Basic Public Health Services. Available from: http://www.nhc.gov.cn/ewebeditor/uploadfile/2017/04/20170417104506514.pdf. (In Chinese).
6. Lelis VM, Guzman E, Belmonte MV. A statistical classifier to support diagnose meningitis in less developed areas of Brazil. J Med Syst. 2017;41(9):145.
7. World Health Organization, 2008–2013 action plan for the global strategy for the prevention and control of non-communicable disease.
8. Choi SB, et al. Screening for prediabetes using machine learning models. Comput Math Methods Med. 2014;2014:618976.
9. Olivera AR, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. Sao Paulo Med J. 2017;135(3):234–46.
10. Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob J Health Sci. 2015;7(5):304–10.
11. Huang ML, Chen HY. Glaucoma classification model based on GDx VCC measured parameters by decision tree. J Med Syst. 2010;34(6):1141–7.
12. Farion K, et al. A tree-based decision model to support prediction of the severity of asthma exacerbations in children. J Med Syst. 2010;34(4):551–62.
13. Gregori D, et al. Non-invasive risk stratification of coronary artery disease: an evaluation of some commonly used statistical classifiers in terms of predictive accuracy and clinical usefulness. J Eval Clin Pract. 2009;15(5):777–81.
14. Chao CM, et al. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst. 2014;38(10):106.
15. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. Int J Med Inform. 2017;97:304–11.
16. Takada M, et al. Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. BMC Med Inform Decis Mak. 2012;12:54.
17. Cakir A, Demirel B. A software tool for determination of breast cancer treatment methods using data mining approach. J Med Syst. 2011;35(6):1503–11.
18. Saybani MR, et al. Diagnosing tuberculosis with a novel support vector machine-based artificial immune recognition system. Iran Red Crescent Med J. 2015;17(4):e24557.
19. Zmiri D, Shahar Y, Taieb-Maimon M. Classification of patients by severity grades during triage in the emergency department using data mining methods. J Eval Clin Pract. 2012;18(2):378–88.
20. Meng X-H, et al. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93–9.
21. Ramezankhani A, et al. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran lipid and glucose study. Diabetes Res Clin Pract. 2014;105(3):391–8.
22. Yu W, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10:16.
23. Boyce S, et al. Evaluation of prediction models for the staging of prostate cancer. BMC Med Inform Decis Mak. 2013;13:126.
24. Kammerer JS, et al. Tuberculosis transmission in nontraditional settings: a decision-tree approach. Am J Prev Med. 2005;28(2):201–7.
25. Tayefi M, et al. The application of a decision tree to establish the parameters associated with hypertension. Comput Methods Prog Biomed. 2017;139:83–91.

26. Alickovic E, Subasi A. Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. J Med Syst. 2016;40(4):108.
27. Hu FB, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med. 2001;345(11):790–7.
28. Anderson JW, et al. Carbohydrate and fiber recommendations for individuals with diabetes: a quantitative assessment and meta-analysis of the evidence. J Am Coll Nutr. 2004;23(1):5–17.
29. Colditz GA, et al. Weight gain as a risk factor for clinical diabetes mellitus in women. Ann Intern Med. 1995;122(7):481–6.
30. Koppes LL, et al. Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. Diabetes Care. 2005;28(3):719–25.
31. Li CP, et al. Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chin Med J. 2012;125(5):851–7.
32. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med. 1999;16(1):3–23.
33. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81–97.
34. Marinov M, et al. Data-mining technologies for diabetes: a systematic review. J Diabetes Sci Technol. 2011;5(6):1549–56.
35. Ravikumar P, et al. Prevalence and risk factors of diabetes in a community-based study in North India: the Chandigarh urban diabetes study (CUDS). Diabetes Metab. 2011;37(3):216–21.
36. Reis JP, et al. Lifestyle factors and risk for new-onset diabetes: a population-based cohort study. Ann Intern Med. 2011;155(5):292–9.
37. Li G, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing diabetes prevention study: a 20-year follow-up study. Lancet. 2008;371(9626):1783–9.
38. Saaristo T, et al. Lifestyle intervention for prevention of type 2 diabetes in primary health care: one-year follow-up of the Finnish National Diabetes Prevention Program (FIN-D2D). Diabetes Care. 2010;33(10):2146–51.
39. Tuomilehto J, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med. 2001;344(18):1343–50.
40. Anderson AE, et al. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. J Biomed Inform. 2016;60:162–8.
41. Devoe JE, et al. Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. Ann Fam Med. 2011;9(4):351–8.
42. Heydari, M, et al. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev C. 2016;36(2):167–73.