Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

An improved support vector machine-based diabetic readmission prediction

Shaoze Cui[a], Dujuan Wang[b,*], Yanzhang Wang[a], Pay-Wen Yu[c], Yaochu Jin[a,d]

[a] *School of Management Science and Engineering, Dalian University of Technology, Dalian 116023, PR China*
[b] *Business School of Sichuan University, Chengdu 610064, China*
[c] *Department of Physical Education, Fu Jen Catholic University, New Taipei City 24205, Taiwan*
[d] *Department of Computer Science, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom*

ARTICLE INFO

Article history:
Received 30 May 2018
Revised 7 October 2018
Accepted 12 October 2018

Keywords:
Hospital readmission
Diabetes
Support vector machine
Synthetic minority over-sampling
Feature selection

ABSTRACT

Background and objective: In healthcare systems, the cost of unplanned readmission accounts for a large proportion of total hospital payment. Hospital-specific readmission rate becomes a critical issue around the world. Quantification and early identification of unplanned readmission risks will improve the quality of care during hospitalization and reduce the occurrence of readmission. In clinical practice, medical workers generally use LACE score method to evaluate patient readmission risks, but this method usually performs poorly. With this in mind, this study presents a novel method combining support vector machine and genetic algorithm to build the risk prediction model, which simultaneously involves feature selection and the processing of imbalanced data. This model aims to provide decision support for clinicians during the discharge management of patients with diabetes.
Method: The experiments were conducted from a set of 8756 medical records with 50 different features about diabetic readmission. After preprocessing the data, an effective SMOTE-based method was proposed to solve the imbalance data problem. Further, in order to improve prediction performance, a hybrid feature selection mechanism was devised to select the important features. Subsequently, an improved support vector machine-based (SVM-based) method was developed and the genetic algorithm was used to tune the sensitive parameter of the algorithm. Finally, the five-fold cross-validation method was applied to compare the performance of proposed method with other methods (LACE score, logistic regression, naïve bayes, decision tree and feed forward neural networks).
Results: Experimental results indicate that the proposed SVM-based method achieves an accuracy of 81.02%, a sensitivity of 82.89%, a specificity of 79.23%, and outperforms other popular algorithms in identifying diabetic patients who may be readmitted.
Conclusions: Our research can improve the performance of clinic decision support systems for diabetic readmission, by which the readmission possibility as well as the waste of medical resources can be reduced.

1. Introduction

With the development of economy and society, people pay more attention to healthcare which consumes large amount of resources at the same time. In medical research, hospital readmission has been recognized as "common, expensive and often preventable" [1], and it has become a major topic in healthcare system. In the United States, readmission is a very common problem, with 20% Medicare beneficiaries readmitted within 30-days after hospital discharge, readmission costs roughly $17 billion in annual spending [2]. To reduce the aforementioned costs, Centers for Medicare and Medicaid Services have launched a readmission payment reduction program aiming at reducing the readmission rates [3]. The program punishes those hospitals with high readmission rates and reduces financial allocations for these hospitals as punishment. Thus, it helps hospitals attach more importance to readmission problems, and prompts them use effective and efficient interventions to reduce readmission rate.

Nowadays, healthcare resources are costly and limited. Moreover, hospital readmission is now widely accepted as a quality of care barometer [4,5], and it has joined the ranks of mortality and complication rates in the world of "quality of care outcomes measures". By identifying patients at high risk of readmission, doctors

[*] Corresponding author.
E-mail address: wangdujuan@dlut.edu.cn (D. Wang).

https://doi.org/10.1016/j.cmpb.2018.10.012

can take targeted interventions to prevent readmission. Further, prevention of avoidable readmission can significantly improve the quality of patient life and improve the financial condition of the health care system.

In order to solve the readmission problem, researchers are assessing the effectiveness of various methods in improving the discharge process. Interventions for readmission are generally divided into three categories: pre-discharge interventions (patient education, discharge plans, drug integration and scheduled follow-up time), post-discharge interventions (telephone follow-up, communication with outpatient care staff and home visits), and transitional interventions (transitional guidance, patient-centered discharge guidance). It is shown that interventions for patients who may be readmission can effectively reduce the incidence of readmission [6]. To corroborate this, some researchers found that the effectiveness of post-discharge phone calls on 30-day preventable readmission rates with the pediatric hospital setting had been evaluated [7]. In this study, we focus on pre-discharge intervention, because many preventable hospital readmission derive from a low quality of care during hospitalization as well as poor discharge process arrangement [8]. Through the modeling of patient's pre-discharge indicators, we can predict whether the patient is likely to be readmitted in the short term.

Research has been done to determine which factors contribute to early re-hospitalization, which mainly focuses on a particular disease. In addition, most researches are about heart failure (HF) [4,9–12], and relatively little research focus specifically on diabetic readmission [13]. Diabetes is a major global health problem that affects hundreds of millions of people around the world, and it takes 11.6% of the total health care expenditure in the world in 2010 [14]. According to the World Health Organization's survey [15], 1.5 million people died of diabetes in 2012 around the world, and the number of diabetes will increase remarkably in the next decade [16]. Besides the diabetes can affect individual health, it brings high social costs [17]. With the rapid growth of hospitalized patients with diabetes, the burden of that is substantial, growing, and costly. Readmission issue makes this situation worse. Through the identification of potential diabetic readmission patients, we can pay more attention to them and try to avoid the occurrence of diabetic readmission.

In medical field, clinicians commonly use scoring methods to predict readmission risk of a patient. The scoring methods generally contain several attributes related to readmission. By scoring these patient's features, patients can be divided into several groups, such as high-risk patients and low-risk patients, and then doctor can speculate the possibility of readmission in the future. The most well-known scoring methods include LACE score and HOSPITAL score, and many studies use these two methods to predict readmission risk in a particular disease [18–21]. Although these scoring methods are more convenient for clinical practitioners, the accuracy of these methods is often unsatisfactory. And when applying to certain scenes, the score methods perform only a little better than random guess [19].

The reason for low accuracy of current approaches is that the factors affecting readmission are very complex, including patients' health condition, quality of inpatient care and social determinants etc. In addition, current readmission prediction methods rely strongly on the experience of clinicians. In order to solve the problem of low prediction accuracy, this study introduces machine learning methods into readmission prediction. Machine learning is a hot research field in recent years, which involves many subjects of knowledge, and is now widely used in various fields. Especially for prediction problems, machine learning methods usually perform better than other methods. Nowadays, machine learning techniques have found use in various health-care applications. For example, Baskaran et al. [22] use neural networks and a supervised learning method, to predict breast screening attendance. Zolbanin et al. [23] predict overall survivability in comorbidity of cancers using random forest algorithm.

Prediction methods have been proposed to address the readmission problem for diabetic patients. In this study, the readmission rate of diabetic patients in 130 United States hospitals is investigated. The majority of past research on hospital readmission use LACE score, HOSPITAL score, logistic regression, cohort study. Considering the imbalance of data, an effective SMOTE-based class imbalance processing method is proposed in this study. Moreover, for the problem of feature selection, we propose a new feature selection method with accuracy and efficiency under consideration. In addition, considering the prediction accuracy of support vector machines is easily affected by kernel function and parameter selection, we compare the performance of several different kernel functions and use genetic algorithm to tune the sensitive parameter. Beyond that, we compare the proposed diabetic readmission prediction method with other methods such as LACE score, Naïve Bayes, Decision Tree, Logistic Regression and Back Propagation Neural Network (BPNN). The experimental results show that the proposed method is superior to the above methods.

This paper is organized as follows: Section 2 explains data used in this study and its handling methods, and describes the proposed method. Section 3 presents the results of extensive numerical studies to evaluate the performance of the proposed method. Section 4 discusses related works and the results of this paper. We conclude the study and suggest topics for future research in the last section.

2. Materials and methods

In this section, we will describe the proposed method in detail. First, we describe the data together with data pre-processing and feature selection used in this study. Then, we will introduce the framework of the proposed algorithm. After that, we will introduce an improved support vector machine method (SVM) used in this study. In the last part of this section, a genetic algorithm (GA), which is a type of evolutionary algorithms, is used to optimize the parameters of the support vector machine.

2.1. Diabetic readmission dataset

In order to verify the effectiveness of the proposed method in predicting diabetic readmission, we need to obtain the corresponding data set. A data set derived from medical records is used to investigate the implicit regularities in hospital readmission of diabetes patients. The data set belongs to Health Facts database (Cerner Corporation, Kansas City, MO).

In this data set, there are 50 features as presented in Table A1. Various nominal feature values are indexed by the numeric values for predictive model preferences. The features consist of four major parts: the basic information of the patient, the patient's past medical record, medication and readmission.

In order to test the generalization performance of the proposed prediction method, we use a 5-fold cross-validation method to divide the data set into five parts, where the five parts will take turns as a test set. When a part serves as a test set, the other four parts will serve as a training set. Compared to the 10-fold cross-validation, the 5-fold cross-validation has a faster calculation speed, and the details are shown in Section 3.1. The training set is used to train the model and optimize the parameters, and the test set is used exclusive for evaluating the performance of the proposed method.

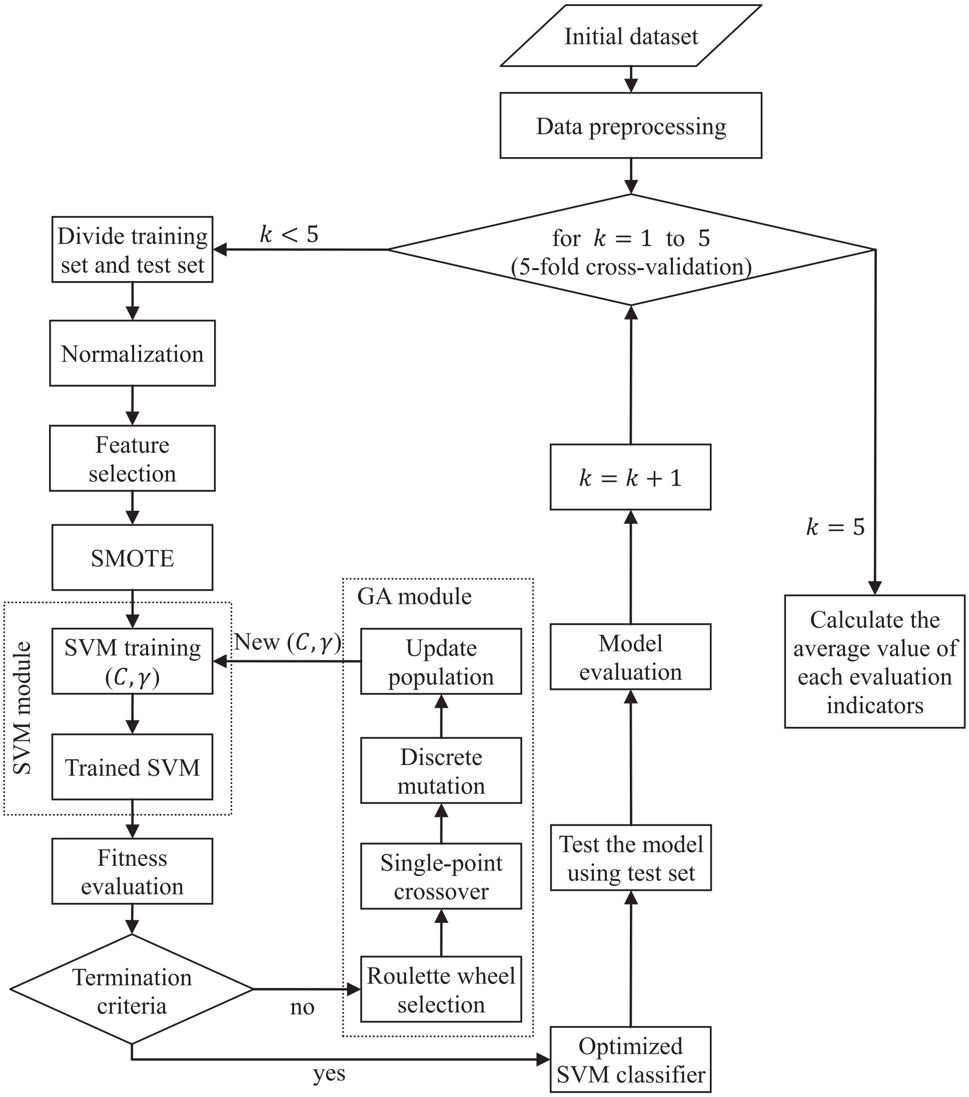

Fig. 1. A flowchart of the proposed method.

2.2. Modeling approach

This study develops a hybrid method to predict diabetic readmission risks, which uses GA to tune the parameters of SVM to improve its performance. The flowchart of the proposed method is shown in Fig. 1. First, we deal with problems with missing values and text content in the initial dataset, and get the data that can be used for modeling. Subsequently, the training set and the test set are divided by the 5-fold cross-validation method, and the training set data of each fold is then processed in the order of normalization, feature selection, and SMOTE [24–26]. After that, the parameters of the SVM classifier are tuned using GA. Finally, the trained model is evaluated using the test set.

Since the misclassification of non-readmission patients has less impact on patient health and fewer consequences than readmitted patient misclassification, we should pay more attention to readmitted patient. Thus, we introduce F-measure and G-mean to evaluate the ability of a model to correctly identify the readmitted patient. Compared with the commonly used indicators such as accuracy and specificity, F-measure as a comprehensive indicator considers both the precision and recall (sensitivity). G-mean is another indicator that can evaluate the model performance of imbalanced data. A detailed description of model evaluation indicators is described in Section 3.1.

2.3. Data pre-processing

2.3.1. Extraction of initial dataset

Information is extracted from the database for encounters that satisfy the following criteria.

(1) It is an inpatient encounter (a hospital admission).
(2) It is a “diabetic” encounter, that is, one during which any kind of diabetes is entered to the system as a diagnosis. We select the sample from database which the ICD9 codes value of “Diagnosis 1” equal 250.xx (250.00–250.99).
(3) The length of stay is at least 1 day and at most 14 days.
(4) Laboratory tests are performed during the encounter.
(5) Medications are administered during the encounter.

Criteria 3 and 4 are applied to remove admissions whose duration are less than 23 hours. In sum, 8,756 encounters are identified that fulfill all of the above five criteria and are used in the following process.

2.3.2. Data cleaning

2.3.2.1. Treatment of missing values. We all know it is common to have incomplete data in classification cases. In the UCI repository, one most commonly used dataset collection, only 55% of datasets are complete while the rest 45% have missing values. And it is more serious in clinical information databases [27]. Missing values usually have great influence on the results, and there are several methods that can solve this problem [28]. In general, if the number of missing values is small, we can remove the sample which has the missing values. In this study, the detailed information on the removed samples are as follows:

- 189 samples are removed because of lacking in “race”;
- 600 samples are removed because of lacking in “Diagnosis 2”;
- 382 samples are removed because of lacking in “Diagnosis 3”;

After the above processing, 7965 valid samples are kept.

2.3.2.1. Feature processing. For the given samples, it is evident that some features have little impact on the results. So, we should remove these irrelevant features. Under the guidance of a clinical doctor, we remove the following features: Encounter ID, Patient number, Payer code.

The missing values of “Weight” feature reach 97%, so the “Weight” feature cannot be used in this study. But “medical specialty” feature is maintained, adding the value “missing” in order to account for missing values. In addition, just a few samples of different values in some medication features, these features have little impact on the classification, so we can remove those. The features belong to this kind including chlorpropamide, acetohexamide, tolbutamide, miglitol, troglitazone, tolazamide, examide, citoglipton, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone. Because “Diagnosis 1″ is the primary diagnosis considered in this study, we just take the feature “Diagnosis 1″ into consideration and remove “Diagnosis 2″ and “Diagnosis 3″.

After the feature processing, 32 features are considered for classification. In these features, some are numeric data and others are nominal data. In order to process and analyze data, we convert the nominal features to numeric features. Take feature “A1c test result” for example, it has 4 distinct integer values and each integer value corresponds to a categorical value (0 = “ > 8″, 1 = “ > 7″, 2 = “normal”, 3 = “none”). In addition, we divide feature “Readmitted” into two categories. In the original data set, feature “Readmitted” has 3 distinct values. On such hospital quality benchmark, known as 30-day, all-cause risk-standardized readmission, is compiled for each hospital relative to a national average and reported publicly on the Internet (https://www.medicare.gov/hospitalcompare/search.html) [29]. Therefore, we pay more attention on 30-day readmission.

2.3.3. SMOTE-based imbalanced data processing

After the data preprocessing, the samples are labeled as the readmission or non-readmission class. The ratio between the two classes is 6.4:1, which is highly imbalanced. This imbalance prevents us from reporting the single classification error number, because one class would dominate the other [12].

Generally, there are three resampling strategies that are used to obtain balanced classes in data mining: over-sampling, under-sampling and hybrid methods [30]. In this study, we use an over-sampling method named synthetic minority over-sampling technique (SMOTE) algorithm to address data imbalance. Although there are many classification algorithms for imbalanced data sets in the existing literature, the SMOTE is perhaps one of the most used approaches to improve the performance of classifiers on skewed data sets [31]. Fig. 2 depicts the process of synthesizing new data.

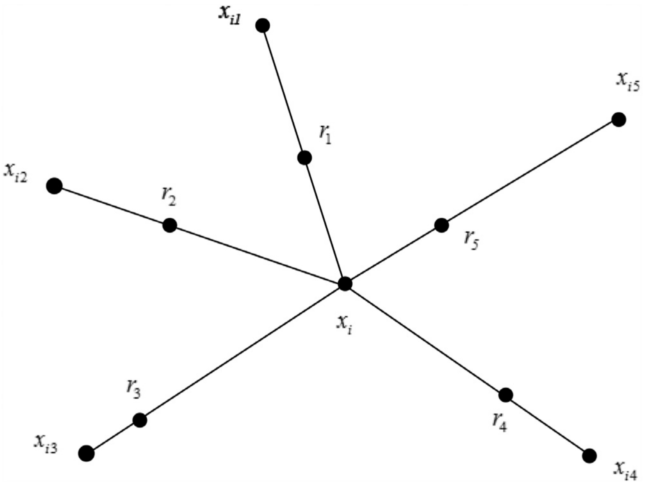

Fig. 2. Schematic diagram for generating new data with SMOTE.

Let x_i be a sample of minority class samples, $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$ the five nearest same class neighbors of x_i, and r_1, r_2, r_3, r_4, r_5 the five new synthetic data generated. This method generates the new minority class samples and improves the generalization ability of the model. A new sample is synthesized by the following formula: $r_j = x_i + rand(0, 1) \times (x_i - x_{ij})$. In addition, the SMOTE method is able to alleviate the over-fitting problem which simple copy of the sample may cause. And SMOTE is modified for discrete features, which is originally used for continuous features. In addition, we have designed a mechanism to automatically determine the amount of SMOTE $N\%$ to balance both classes to 50% [33]. The details of the SMOTE algorithm are shown in Table 1.

2.4. A hybrid feature selection mechanism

In machine learning and statistics, feature selection is the process of selecting a subset of relevant features which will be used in model construction. In general, the commonly used feature selection methods can be divided into three categories: filter, wrapper and embedded methods [34]. The filter method is a preprocessing step and use criteria not involving any learning machine and, by doing that, it does not consider the effects of a selected feature subset on the performance of the algorithm. Wrapper method evaluates a subset of features according to accuracy of a given predictor. Embedded method performs feature selection during the process of training, which is usually specific in given learning machines.

In this study, we proposed a hybrid feature selection mechanism that combines the filter method with the wrapper method. First, we use five widely used filter feature scoring methods (Logistic regression stepwise selection, Information gain, Information gain ratio, Chi-square test and Gini index) to evaluate the importance of each feature. After that, the features are ranked according to their importance scores. Finally, the features are added to the classifier one by one in order to observe the change in time and accuracy brought about by the increase in features. The detailed feature selection process in this study is shown in Table 2.

2.5. Improved support vector machine-based learning

The principle of support vector machine method is as follows. Consider a classification problem constituted with n sample-label pairs, $S = (x_i, y_i), (i = 1, 2, \ldots, n)$, $x_i \in R$ is a training set, and $y_i \in \{-1, +1\}$ is a class label. A hyperplane is constructed by the equation $\omega^T x + b = 0$ (where “ω” is the vector of hyperplane coefficients, “b” is a bias term) to maximize the margin between the

Table 1
The SMOTE algorithm.

```
Input: The initial dataset $D=(x_i,y_i), (i=1, 2, ..., n), x_i \in R^d$, where $d$ is the number of features; Number of nearest neighbors $k$; The amount of SMOTE $N$
Output: The synthetic minority class samples
Divide the initial dataset $D$ into a training set $S_{train}$ and a test set $S_{test}$
Calculate the number of minority class samples $Sample_{mi}$ marked as $Num_{mi}$ and the number of majority class samples $Sample_{ma}$ marked as $Num_{ma}$ in $S_{train}$
Calculate the number of new samples Synthetic that should be synthesized by each minority class sample $N$: $N = \frac{Num_{ma}}{Num_{mi}} - 1$ ("⌊⌋" indicates round down, "– 1" is to avoid the total number of minority class samples $Num_{mi}{}^{*}(N+1)$ after synthesis is more than the $Num_{ma}$)
for $j=1$ to $Num_{mi}$ do
    Compute $k$ nearest neighbors for $j$ in minority class, and save the indices in the narray
    for $l=1$ to $N$ do
        Randomly choose one of the $k$ nearest neighbors of $j$, call it nn
        for $feat=1$ to $d$ do
            Compute: $dif=Sample_{mi}[narray[nn]][feat]-Sample_{mi}[j][feat]$
          Compute: $gap=$ random number between 0 and 1
          $Synthetic[newindex][feat]=Sample_{mi}[j][feat]+gap^{*}dif$
            if feat is a discrete value
                Round $Synthetic[newindex][feat]$
            end if
          end for
      end for
end for
return the synthetic minority class samples Synthetic
```

Table 2
Process of the hybrid feature selection mechanism.

```
Input: The initial dataset $D=(x_i, y_i), (i=1, 2, ..., n), x_i \in R^d$, where $d$ is the number of features
Output: Optimal feature subset
Calculate the score of each feature using five filter feature scoring methods, and save the results to $Score_{LR}$, $Score_{IG}$, $Score_{IGR}$, $Score_{CS}$, $Score_{GI}$ ("LR" indicates Logistic regression stepwise selection, "IG" indicates Information gain, "IGR" indicates Information gain ratio, "CS" indicates Chi-square test, and "GI" indicates Gini index)
Sort the features in $Score_{LR}$, $Score_{IG}$, $Score_{IGR}$, $Score_{CS}$, $Score_{GI}$ from large to small according to the score
Create an array FSarray[] for loading features one by one
Perform the following operations on the sorting results of each filter feature scoring method, and use $RScore_{each}$ to represent the sorting result of any of the five methods.
for $j=1$ to $d$ do
    Add $RScore_{each}[j]$ to FSarray[]
    $NewD=D[1{:}\ n][FSarray[]]$
    Divide NewD into 5 parts for 5-fold cross-validation
    Record current system time as tic
    for $k=1$ to 5 do
        One part as the test set $NewD_{test}$, the other four parts as the training set $NewD_{train}$
        Train the classifier using $NewD_{train}$
        Test trained classifier and calculate the values of each evaluation indicator using $NewD_{test}$
    end for
    Record current system time as toc, and calculate single runtime $T = \frac{1}{5}(toc-tic)$
    Calculate the average of each evaluation indicator
end for
Calculate the ratio of the evaluation indicator to the runtime in each feature subset $Ratio_l$: $Ratio_l = the\ evaluation\ indicator\ value_l/T_l$, $(l=1, 2, 3, ..., d)$
return the feature set with the largest $Ratio_l$ is selected as the optimal feature subset within a certain evaluation indicator interval.
```

hyperplane and the support vectors (nearest data points). For finding a hyperplane that can separate the positive $(+1)$ samples from the negative (-1) samples, we use classifier training to do this.

To obtain the optimal separating hyperplane $\omega^T x + b = 0$, an optimization problem needs to be solved.

$$\begin{aligned} &\min_{\omega,b} \quad \tfrac{1}{2}\|\omega\|^2 \\ &s.t. \quad y_i(\langle \omega, x_i\rangle + b) \geq 1. \end{aligned} \tag{1}$$

However, it is difficult to find a hyperplane that can separate data points completely and correctly in some problems. Such complex classification hyperplane may lead to overfitting of the model, resulting in a reduction in generalizability of the prediction model. To circumvent this difficulty, a soft margin is used, and then the optimization problem (1) is reformulated as follows:

$$\begin{aligned} &\min_{\omega,b,s} \quad \tfrac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ &s.t. \quad y_i(\langle \omega, x_i\rangle + b) \geq 1-\varepsilon_i, \ \ i=1,\ 2,\ \ldots,\ n. \end{aligned} \tag{2}$$

where ε_i is called a slack variable, and C is the penalty coefficient.

In a real-world task, however, it is very common that data are only non-linearly separable. To solve the problem of nonlinearity, we should project the original data into the high-dimensional space through a nonlinear mapping Φ_x. In the high-dimensional space, the data may be linearly separable.

Both (1) and (2) can be solved in dual form using the Lagrange method. In a nonlinear case, the dual form is:

$$\begin{aligned} &\min_{\alpha} \quad \sum_i \alpha_i - \tfrac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \langle \Phi(x_i), \Phi(x_j)\rangle_F \\ &s.t. \quad 0 \leq \alpha_i \leq C, \ \sum_i \alpha_i y_i = 0. \end{aligned} \tag{3}$$

where α_i is the Lagrange multiplier. The final decision function is:

$$class(x) = sign\left(\sum_i \alpha_i y_i \langle \Phi(x_i), \Phi(x)\rangle_F + b\right). \tag{4}$$

The key to entire construction process of an SVM is the kernel function, which can be expressed as the inner product $K(x,$

Table 3
Four types of Kernel functions.

Kernel type	Function
Linear function	$K(x, y) = \langle x, y\rangle$
Polynomial function	$K(x, y) = (\langle x, y\rangle + p)^q, q \in N$
Radial basis function (RBF)	$K(x, y) = exp(-\gamma\|x-y\|^2), \gamma > 0$
Sigmoid function	$K(x, y) = tanh(\gamma\langle x, y\rangle + c),\ \gamma > 0,\ c < 0$

$y) = \langle\Phi(x),\ \Phi(y)\rangle_F$. The kernel defines the similarity or a distance measure between new data and the support vectors. The dot product is the similarity measure used for linear SVM or a linear kernel because the distance is a linear combination of the inputs. Other kernels can be used that transform the input space into higher dimensions such as a polynomial kernel and a radial basis function kernel. Frequently used kernel functions are listed in Table 3.

When using the support vector machine to solve classification problems, there are two factors that play a key role. One is the choice of support vector machine kernel function, and the other is the selection of support vector machine parameters. For selection of kernel functions to solve practical problems, there are two common approaches: One uses a priori knowledge of experts to select a kernel function; another is the cross-validation method. In this study, we take the four kernel functions in Table 3 into consideration and use each kernel function to build the prediction model. By comparing the performance of SVM models with different kernel functions, we try to find the most appropriate kernel function for predicting the diabetes readmission.

In addition to the kernel function selection we mentioned above, the parameters of the support vector machine have a great influence on the classification results [35]. Despite this fact, there are still some effective approaches to select the proper parameters for an SVM, among which grid search (GS) algorithm is the most straightforward [36]. However, GS algorithm suffers from a heavy computational burden because the SVM model has to be rebuilt for all combinations of parameters. Compared with exhaustive search algorithms such as the grid search algorithm, meta-heuristic method can find the approximate optimal solution at a faster speed, so it has been widely used in parameter optimization. In this study, a genetic algorithm (GA) is implemented for parameters search in order to build an optimal classifier. There are two parameters to be optimized, one is punishment factor C, and the other is kernel parameter γ.

The basic elements of the genetic algorithm include: chromosome encoding method, fitness function, genetic operations and operation parameters. The key of genetic algorithms is the design of the fitness function, which is problem-specific. In this study, we use a number of different fitness functions as shown in Table 9 in the experimental studies. The SVM optimized with a GA (GA-SVM for short) proposed in this study is summarized in Table 4.

3. Results

3.1. Performance metrics

The confusion matrix is a tool that puts the true condition and predicted condition in the same matrix, as shown in Fig. 3.

The possible outcomes of a classification task can be interpreted as one of the four categories:

(1) True positive (TP): correctly classified as positive.
(2) False positive (FP): incorrectly classified as positive.
(3) True negative (TN): correctly classified as negative.
(4) False negative (FN): incorrectly classified as negative.

A positive pattern refers to a readmitted patient, whereas a negative pattern refers to a non-readmission patient [8]. Accuracy is the rate of correct classification and it is defined as:

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)**
	condition negative	**False Positive (FP)**	**True Negative (TN)**

Fig. 3. The confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{5}$$

Sensitivity (also known as recall) is the ability of a test to correctly identify those with the readmission (true positive rate), which is defined as:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{6}$$

Specificity is the ability of the test to correctly identify those without the readmission (true negative rate), which is defined as:

$$Specificity = \tfrac{TN}{TN + FP}. \tag{7}$$

Precision is indicated by the number of correctly classified positive examples divided by the number of labeled by the prediction method as positive, which is given as:

$$Precision = \frac{TP}{TP + FP}. \tag{8}$$

Many algorithms fail to optimize precision and recall at the same time. In statistical analysis of binary classification, the F-measure considers both the recall and the precious of the test to compute the score [37]. F-measure is defined as:

$$F_\beta - measure = \frac{\left(1 + \beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{9}$$

From the F-measure formula, we can see that the weight of precision and recall can be changed by adjusting the β value. However, when the sample number of negative class is much larger than that of positive class, even if the sensitivity is improved, the change of F-measure will not be great.

G-mean is another indicator that can evaluate the model performance of imbalanced data, which represents the geometric mean of all class recall rates. The higher the G-mean value, the better the classification performance. It is defined as:

$$\begin{aligned} G - mean &= \sqrt{Sensitivity \times Specificity} \\ &= \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}. \end{aligned} \tag{10}$$

K-fold cross-validation is a method for evaluating the generalization ability of a model, which is widely used in machine learning and statistics. Fig. 4 shows a schematic representation of a K-fold cross-validation. The method divides the total data in K parts, trains some of them, and then tests the others using the trained model. In general, 5-fold cross-validation and 10-fold cross-validation are the most commonly used, and the results of the two are not significantly different [38]. Hence, considering the computational cost, 5-fold cross-validation is chosen in this study. In addition, in order to avoid the randomness brought by cross-validation, we repeat the 5-fold cross-validation 25 times and use the average value as the final result [39].

Table 4
The GA-SVM algorithm.

```
Input: The dataset $S=(x_i,y_i), (i=1, 2, ..., n), x_i \in R^d$, where $d$ is the number of features, $y_i \in \{0, 1\}$
Output: The optimal parameters $(C, \gamma)$
Generate from $S$ the training dataset $P$ and the test dataset $V$
Call the GA to randomly initialize a population with $I$ individuals
while stopping criterion is not reached do
    for $k=1$ to $I$ do
        Extract the combination of SVM parameters $(C, \gamma)$ from $kth$ chromosome
        for $l=1$ to 5 do
            Divide $P$ into 5 copies, 1 as a test set $Q$, 4 as a training set $T$
            Perform the SVM model with training set $T$
            Get the predicted value $y^{(k)} = [y_1^{(k)}, y_2^{(k)}, \ldots, y_m^{(k)}]^T$, establish confusion matrix
          Calculate the TP, TN, FP, FN value
          end for
          Calculate the fitness value of the $kth$ individual using the fitness function
          end for
          Select the parents
          Make the crossing of the selected parents
          Perform the mutation on the new individuals
          Update the population
end while
Retrieve the fittest individual to perform the test phase using $V$
return the optimal parameters $(C, \gamma)$
```

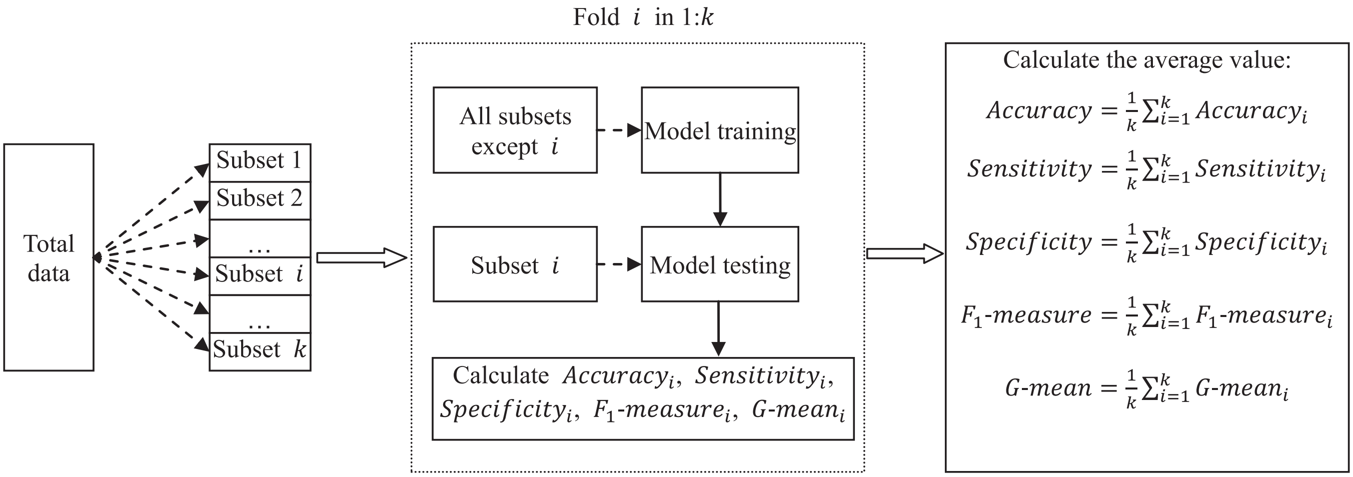
Fold i in 1:k
Total data
Subset 1
Subset 2
…
Subset i
…
Subset k
All subsets except i
Model training
Subset i
Model testing
Calculate $Accuracy_i$, $Sensitivity_i$, $Specificity_i$, F_1-$measure_i$, G-$mean_i$
Calculate the average value:
$Accuracy = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i$
$Sensitivity = \frac{1}{k}\sum_{i=1}^{k} Sensitivity_i$
$Specificity = \frac{1}{k}\sum_{i=1}^{k} Specificity_i$
$F_1\text{-}measure = \frac{1}{k}\sum_{i=1}^{k} F_1\text{-}measure_i$
$G\text{-}mean = \frac{1}{k}\sum_{i=1}^{k} G\text{-}mean_i$

Fig. 4. K-fold cross-validation.

3.2. Data normalization

In the field of data mining, data normalization is a fundamental step, which can eliminate the differences caused by different dimensions. In this study, the min-max normalization method is used to normalization the initial data before establish the model. The conversion function is defined as follows:

$$x_i'(k) = \frac{x_i(k) - \min\limits_i x_i(k)}{\max\limits_i x_i(k) - \min\limits_i x_i(k)}. \tag{11}$$

where $k=1, 2, \cdots, n,\ i=1, 2, \cdots, l,\ x_i(k)$ represents the k *th* feature of the i *th* sample.

The detailed experimental procedure of data normalization is as follows: Firstly, the training set is normalized according to Eq. (11), and the normalized parameters are recorded (mapminmax function in MATLAB); then the test set is normalized using the recorded normalization parameters. In this way, the relative independence between training set and test set can be maintained.

In machine learning, some methods (e.g. Logistic Regression, Decision Tree and Naïve Bayes) can be used without recoding of the features and data normalization, but for the comparison of subsequent experiments, we perform these operations on all classifiers. Specially, we will conduct a sensitivity analysis of the changes in the performance of the classifier before and after the operation. As shown in Table 5, we find that the recoding of the features and data normalization do not reduce the performance of the classifier. And Le et al. [40] also point out that data normalization can help the classification algorithm (including the Logistic Regression method) converge faster.

3.3. Feature selection results

In order to select features that have a significant impact on classification, we combine the filter method with the wrapper method to design a hybrid feature selection mechanism as shown in Table 2. In order to clearly demonstrate this feature selection process, we conducted an experiment. In this experiment, we use SVM with RBF as the classifier, and set the parameter values according to the default settings in e1071 package of the R language ($C=1,\ \gamma=1/(data\ dimension)$). When the hybrid feature selection mechanism is used with other classification methods, the classifier in Table 2 is set to the corresponding classification method. As for the five filter methods in the experiment, they are from the FSelector package (Information gain, Information gain, Chi-square test), the rpart package (Gini index), and the MASS package (Logistic regression stepwise selection) in the R language. What's more, in the selection of optimal feature subset, we need to set the inter-

Table 5
Sensitivity analysis of recoding of features and data normalization.

Training model	Processing type	Accuracy	Sensitivity	Specificity	F1-measure	G-mean
Decision tree	Basic	0.8650	0.0000	1.0000	0.0000	0.0000
	Recoding	0.8647	0.0000	1.0000	0.0000	0.0000
	Recoding + normalization	0.8650	0.0000	1.0000	0.0000	0.0000
Naïve Bayes	Basic	0.7981	0.2913	0.8821	0.2614	0.5069
	Recoding	0.7879	0.3172	0.8614	0.2861	0.5227
	Recoding + normalization	0.7863	0.3144	0.8599	0.2848	0.5200
Logistic regression	Basic	0.8634	0.0577	0.9890	0.1024	0.2377
	Recoding	0.8658	0.0530	0.9926	0.0960	0.2295
	Recoding + normalization	0.8662	0.0549	0.9927	0.0993	0.2334

Table 6
Experimental results of feature selection.

Filter method	Number of features	Accuracy	Sensitivity	Specificity	F_1-measure	G-mean	Time	Ratio
Information gain ratio	25	0.8159	0.8414	0.7904	0.8205	0.8154	129.46	0.0063
Information gain	20	0.7994	0.8222	0.7766	0.8038	0.7990	102.65	0.0078
Chi-square test	**19**	0.8050	0.8321	0.7778	0.8101	0.8055	**99.28**	**0.0081**
Gini index	20	0.8036	0.8284	0.7789	0.8084	0.8015	105.97	0.0076
Logistic regression stepwise selection	24	0.8044	0.8245	0.7843	0.8083	0.8041	125.87	0.0064

Table 7
Experimental performance of imbalanced data with SMOTE processing.

Training model	Processing type	Accuracy	Sensitivity	Specificity	F_1-measure	G-mean
SVM with RBF	without SMOTE	0.8652	0.0214	0.9968	0.1441	0.1460
	with SMOTE	0.8056	**0.8340**	0.7772	**0.8109**	**0.8051**
Decision tree	without SMOTE	0.8650	0.0000	1.0000	0.0000	0.0000
	with SMOTE	0.8150	**0.7048**	0.9251	**0.7820**	**0.8075**
Naïve Bayes	without SMOTE	0.7863	0.3144	0.8599	0.2848	0.5200
	with SMOTE	0.5950	**0.8352**	0.3549	**0.6735**	**0.5444**
Logistic regression	without SMOTE	0.8662	0.0549	0.9927	0.0993	0.2334
	with SMOTE	0.6676	**0.6226**	0.7126	**0.6519**	**0.6661**
BPNN	without SMOTE	0.8282	0.1265	0.9377	0.1655	0.3444
	with SMOTE	0.7429	**0.7899**	0.6960	**0.7544**	**0.7415**

Table 8
GA experimental parameters tuning results.

Parameter name	Symbolic representation	Value
Population size	*I*	50
Maximum number of iterations	*MAXGEN*	100
Individual length	*L*	2*20
Generation gap	*GGAP*	0.9
Crossover probability	*Cp*	0.7
Probability of mutation	*Pm*	0.02
The range of values for C	*cbound*	[0,100]
The range of values for γ	*gbound*	[0,1000]

val of an evaluation indicator so that the selected optimal feature subset can maintain a certain degree of precision. In this study, we set the interval to [$G\text{-}mean_{max} - 1\%$, $G\text{-}mean_{max}$].

From the experimental results of the feature selection in Table 6, the chi-square test filter method has the largest Ratio value, which means that this method takes the least amount of time while maintaining a certain G-mean value. At the same time, the number of features selected by the chi-square test filter method is the least. Therefore, we use the 19 features which are selected by chi-square test filter method as the optimal feature subset for subsequent experiments.

3.4. Imbalanced data processing results

In order to verify the effectiveness of the SMOTE method in dealing with the imbalanced data in this study, we select different classifiers (Decision Tree, Logistic Regression, Naïve Bayes, SVM with RBF, BPNN) to conduct experiments, and used the 5-fold cross-validation method to calculate the average value of the evaluation indicators. In the SMOTE experiment [32], the input parameters that need to be determined are: number of nearest neighbors k, and amount of SMOTE $N\%$. Analyzing the existing literature on SMOTE, one can notice that $k = 5$ neighbors is usually chosen [33] (we also try several different k values in the experiment and finally decide to use $k = 5$). To balance the number of majority class samples with minority class samples, the amount of SMOTE $N\%$ is determined by the ratio of majority class to minority class and it is dynamically determined and does not need to be given in advance in this study. This experiment is compiled and run in the R language environment, and the default values of functions are used for all parameter values that are not explicitly stated. The details of the SMOTE algorithm are shown in the Table 1, and we conduct the experiment according to the following steps.

(1) The initial data D is divided into a training set S_{train} and a test set S_{test} according to the 5-fold cross-validation method.
(2) Use SMOTE for the training set S_{train} to get a new balanced training set NS_{train}.
(3) Train the classifier with the new training set NS_{train}.
(4) Test the performance of classifier using the test set S_{test}, and calculate the value of each evaluation indicator.
(5) Calculate the average of evaluation indicators after completing the 5-fold cross-validation.

From the experiment results in Table 7 we can see that although the accuracy and specificity of the model have declined, the sensitivity, F_1*-measure* and *G-mean* have been greatly improved. Here the improvement of sensitivity means that the model's ability to predict diabetes readmission has been improved. In the clinical diagnosis process, the identification of patients who are truly readmission is particularly critical. At the same time, this

Table 9
Experimental results of SVM classifiers with various kernel functions.

Training model	Accuracy	Sensitivity	Specificity	F_1-measure	G-mean
Linear function	0.6547	0.5355	0.7740	0.6080	0.6438
Polynomial function	0.7452	0.8329	0.6575	0.7657	0.7400
Radial basis function	**0.7973**	**0.8205**	**0.7741**	**0.8019**	**0.7969**
Sigmoid function	0.5829	0.5816	0.5843	0.5823	0.5829

Table 10
Four types of fitness functions.

Fitness function name	Function
Accuracy	$\mathrm{f} = \frac{1}{5}\left[\sum_{l=1}^{5}\left(\frac{TP+TN}{TP\ +TN+\ FP\ +\ FN}\right)_l\right]$
F_1-measure	$\mathrm{f} = \frac{1}{5}\left[\sum_{l=1}^{5}\left(\frac{2\times TP}{2\times TP\ +\ FP\ +\ FN}\right)_l\right]$
F_2-measure	$\mathrm{f} = \frac{1}{5}\left[\sum_{l=1}^{5}\left(\frac{5\times TP}{5\times TP\ +\ FP\ +\ 4\times FN}\right)_l\right]$
G-mean	$\mathrm{f} = \frac{1}{5}\left[\sum_{l=1}^{5}\sqrt{\frac{TP}{TP\ +\ FN}\times\frac{TN}{TN\ +\ FP}}_l\right]$

experiment also proves the effectiveness of SMOTE in dealing with imbalanced data.

3.5. Classification results

SVMs with four kernel functions (linear, polynomial, radial basis functions and sigmoid) are built and trained in program language R. Experiments are carried out using the 5-fold cross-validation method. Since there are not enough possible parameter values for the linear, polynomial and sigmoid kernels and the low prediction accuracy, the exhaustive search, instead of the GA, is implemented for parameter tuning of these three kernel functions. In addition, the GA is applied to search for the optimal parameters C and γ for the RBF-based SVMs. The parameters of the GA are determined using pilot studies as listed in Table 8. And binary encoding is used in this study, the encoding length of each parameter is 20 bits.

The experimental results summarized in Table 9 show that the SVM with the radial basis function kernel has achieved the best performance on all indicators. Therefore, we will use the radial basis function as the kernel function of SVM.

In order to explore the influence of the fitness function on the performance of the SVM with RBF model, we have tried four different functions as shown in Table 10, each of them represents the importance of different indicators.

From Table 11, we can see that performances of the four fitness functions are relatively similar. But we can still notice that when we give more weight to the true positive (TP), the sensitivity of prediction method will be improved while the performance of specificity will decline. Therefore, in clinical application, we can choose different fitness functions according to our preferences.

Finally, we compare the proposed method with other popular machine learning methods. In order to verify the generalization ability of the method, all machine learning methods are tested using the 5-fold cross-validation, and the whole process is repeated 25 times. Moreover, all machine learning methods use SMOTE for unbalanced data processing and apply the hybrid feature selection mechanism for feature selection. In this way, we can obtain the reliable and unbiased evaluation of each model, and can compare the advantages and disadvantages among them under the same standard.

The results are depicted in Table 12 , which indicate that the proposed methods tend to have better generalization results when trained with the data after SMOTE. Moreover, we should notice that although the accuracy of GA-SVM with RBF is similar to decision tree, it is dominant in sensitivity. Since the misclassification of non-readmission patients has less impact on patient health and fewer consequences than readmitted patient misclassification, we should pay more attention to the increase in sensitivity.

4. Discussion of related works

Up to present, there are some related models aiming at predict readmission in general population settings, in which LACE score method is one of the popular models [41]. The LACE score identifies patients that are at risk of readmission or death within 30-day of discharge, which incorporates four parameters: “L” stands for the length of stay of the index admission; “A” denotes the acuity of the admission (Specifically, if the patient was admitted through the Emergency Department vs. an elective admission); “C” represents co-morbidities, incorporating the Charlson Co-Morbidity Index; and “E” measures the number of Emergency Department visited within the last 6 months. Through LACE score method, the patients’ readmission risk can be divided into three levels, low risk (0~4 points), moderate risk (5~9 points) and high risk (≥ 10 points).

However, with the further research in the field of readmission, the LACE score method has gradually shown its limitations. Donzé et al. [18] points out that the LACE score method has disadvantages of poor applicability and reliability, and this method also needs to calculate another score (the Charlson comorbidity index). Low et al. [20] compare performance of LACE score with a regression model among general medicine patients in Singapore, and they observe that the regression model performs better than LACE score in predicting 30-day readmission. What’s more, they expect that additional factors predicting risk and machine learning techniques should be considered to improve model performance. Cotter et al. [19] use LACE score in an older UK population for predicting readmission, but the result indicated that LACE score got poor performance for predicting 30-day readmission in older UK inpatients. In this study, we compare the LACE score method with some machine learning methods. And from Table 12, we can see that machine learning method is more accurate than current clinical LACE score method, which indicates that machine learning method performs well for diabetic readmission prediction.

In view of the above analysis, most existing studies on readmission prediction focus mainly on heart failure (HF) diseases (please refer to Table A2 for details), and few researchers study readmission with diabetes [42,43]. Realizing the importance of readmission with diabetes, this study attempts to predict readmission using a machine learning method together with a meta-heuristic algorithm to improve the performance. In addition, compared with Duggal et al. [43] and Tutun et al. [42], which use a single predictive model to predict hospital readmission, our study uses a variety of methods for predicting diabetes readmission.

Since clinically readmission cases are less than those without readmission, it raises the issue of data imbalance that generally exists in medical and financial fields [32,44,45,51,60]. Table 7 shows that the proposed SMOTE-based data imbalance processing method improves the performance of diabetic readmission prediction.

Table 11
Experimental performance of four fitness functions.

Training model	Accuracy	Sensitivity	Specificity	F_1-measure	G-mean	C	γ
GA-SVM-accuracy	**0.8270**	0.7752	**0.8783**	0.8173	0.8249	11.5657	13.5088
GA-SVM-F_1	0.8189	0.7606	0.8775	0.8073	0.8168	94.3553	13.3400
GA-SVM-F_2	0.8189	**0.8308**	0.8068	**0.8211**	0.8183	67.5723	6.5031
GA-SVM-G	0.8260	0.7871	0.8654	0.8185	**0.8252**	27.7191	12.0783

Table 12
The performances of proposed method and existing methods.

Training model	Accuracy	Sensitivity	Specificity	F1-measure	G-mean
LACE score	0.5323	0.4893	0.5994	0.5605	0.5416
Naïve Bayes	0.6551	0.5981	0.6421	0.6340	0.6523
Logistic regression	0.6702	0.6281	0.7146	0.6557	0.6689
BPNN	0.7366	0.7895	0.6837	0.7496	0.7328
SVM with RBF	0.7024	0.7213	0.6717	0.7046	0.6945
Decision tree	**0.8186**	0.7054	**0.9219**	0.7950	0.8002
GA-SVM with RBF	0.8102	**0.8289**	0.7923	**0.8198**	**0.8105**

Data dimension often influences the speed and accuracy of prediction method [43,46,47], and key features identification plays an important role in prediction process. Table 6 indicates that chi-square test method performs better than other popular feature selection methods with efficiency and accuracy under consideration. And we identify that numbers of visits in the year preceding the encounter, diagnostic information, discharge disposition, numbers of procedures during the encounter, duration in hospital and so on are key features that impacts most on diabetic readmission prediction.

Table 9 shows that radial basis function is the more suitable kernel to our problem other than linear, polynomial and sigmoid function. Besides, Genetic algorithm plays an important role in SVM parameters optimization and it improves the performance of our proposed prediction method.

5. Conclusion

In this study, an improved support vector machine-based method is proposed to predict diabetic readmission. SMOTE-based data preprocessing is introduced to address the imbalanced data. In addition, we compare five popular feature selection methods and find chi-square test method is more suitable for the problem under consideration.

Comparisons have been done between the proposed prediction method and LACE score, logistic regression, decision tree, BPNN and Naïve Bayes. The experimental results indicate that the decision tree and the proposed method outperform other popular methods for readmission predictions. In addition, the proposed method provides high sensitivity other than accuracy.

In the future, we will try to establish a more accurate readmission prediction method for other diseases, so as to provide more reliable support for clinical decision-making. In addition, we will try to propose more interpretable models and extract important features and rules that improve the effect of diagnosis and treatment. Future research may also use other popular methods, such as Random Forest, to predict diabetic readmission risks.

Conflicts of interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgements

This paper is our original work and has not been published or has it been submitted simultaneously elsewhere. All authors have agreed to the submission and declared that they have no conflict of interest. This paper was supported in part by the National Natural Science Foundation of China (Nos. 71533001, 71501024, 71871148, 71532007); by China Postdoctoral Science Foundation (Nos. 2018T110631, 2017M612099), and by Sichuan University (No. 2018hhs-47).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2018.10.012.

Appendix

Table A1
List of features and their description in the initial dataset.

Feature name	Type	Description and values	%missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0,10), [10,20), …, [90,100)	0%
Weight	Numeric	Weight in pounds	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharge to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9);848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9);923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9);954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: " > 200", " > 300", "normal", and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: " > 8," if the result was greater than 8%, " > 7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
23 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values : " < 30" if the patient was readmitted in less than 30 days, " > 30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission	0%

Table A2
Research map to match between search areas and approaches in the context of patient readmission predictions.

	Condition	Sample size	Features	Main methodology	Readmission length
C. Van Walraven et al. [41]	All	4,812	17	LACE	30 days
J.-Y. Yeh et al. [5]	Hemodialysis	6,284	26	C4.5	–
P.E. Cotter et al. [19]	Older UK population	507	–	LACE and Logistic regression	30 days
A.S. Fialho et al. [47]	Intensive care unit	26,655	23	Fuzzy logic	24 –72 hours
C. Ou-Yang et al. [48]	Stevens-Johnson syndrome	554	20	Rule-based classification	60 days
C. Walsh and G. Hripcsak [49]	All	263,859	8	Lasso regression and SVM	30 days
L.L. Low et al. [20]	All	5862	14	LACE and Logistic regression	30 days
J.D. Donzé et al. [18]	All	117,065	7	HOSPITAL score	30 days
M. Jovanovic et al. [50]	Pediatric disease	66,000	15	Tree-Lasso logistic regression	30 days
E. Shadmi et al. [51]	Psychiatric disease	2842	6	Logistic regression	6/12 months
Q.L. Huynh et al. [52]	Heart failure	565	–	Logistic regression	30 days
T. Cooksley et al. [21]	All	19,277	–	LACE and HOSPITAL score	30 days
P. Yazdan-Ashoori et al. [53]	Heart failure	378	29	LACE	30 days
N. McCabe et al. [54]	Heart failure	71	14	Logistic regression	30 days
O. Ben-Assuli et al. [55]	All	5103	4	Logistic regression	7 days
M.S. Hendryx et al. [56]	All	1384	–	Logistic regression	1 year
H. Uthoff et al. [57]	Heart failure	100	34	Cox regression	0–20 months
D. K.Moser et al. [58]	Heart failure	71	11	Cox regression	30 days
S. Yamada et al. [59]	Heart failure	215	15	Cox regression	30/90 days
K.J. Ottenbacher et al. [60]	Stroke	9584	6	Logistic regression and neural networks	3/6 months
R.E. Hodgson et al. [61]	Psychiatric disease	3404	15	Cox regression	21 days

References

[1] K.E. Joynt, A.K. Jha, Thirty-day readmissions — Truthtruth and consequences, N. Engl. J. Med. 366 (2012) 1366–1369.

[2] M.J. Swain, H. Kharrazi, Feasibility of 30-day hospital readmission prediction modeling based on health information exchange data, Int. J. Med. Inform. 84 (2015) 1048–1056.

[3] Centers for Medicare and Medicaid Services, Readmissions reduction program, 2012. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html.

[4] A. Pandey, H. Golwala, H. Xu, A.D. DeVore, R. Matsouaka, M. Pencina, D.J. Kumbhani, A.F. Hernandez, D.L. Bhatt, P.A. Heidenreich, C.W. Yancy, J.A. de Lemos, G.C. Fonarow, Association of 30-day readmission metric for heart failure under the hospital readmissions reduction program with quality of care and outcomes, JACC Hear. Fail. 4 (2016) 935–946.

[5] J.-Y. Yeh, T.-H. Wu, C.-W. Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decis. Support Syst. 50 (2010) 439–448.

[6] D.J. Rubin, K. Donnell-Jackson, R. Jhingan, S.H. Golden, A. Paranjape, Early readmission among patients with diabetes: A qualitative assessment of contributing factors, J. Diabetes Complications. 28 (2014) 869–873.

[7] R. Flippo, E. NeSmith, N. Stark, T. Joshua, M. Hoehn, Reduction of 30-day preventable pediatric readmission rates with postdischarge phone calls utilizing a patient- and family-centered care approach, J. Pediatr. Heal. Care. 29 (2015) 492–500.

[8] B. Zheng, J. Zhang, S.W. Yoon, S.S. Lam, M. Khasawneh, S. Poranki, Predictive modeling of hospital readmissions using metaheuristics and data mining, Expert Syst. Appl. 42 (2015) 7110–7120.

[9] L. Turgeman, J.H. May, A mixed-ensemble model for hospital readmission, Artif. Intell. Med. 72 (2016) 72–82.

[10] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm, Comput. Methods Programs Biomed. 141 (2017) 19–26.

[11] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, B. Krishnapuram, Predicting readmission risk with institution-specific prediction models, Artif. Intell. Med. 65 (2015) 89–96.

[12] W. Dai, T.S. Brisimi, W.G. Adams, T. Mela, V. Saligrama, I.C. Paschalidis, Prediction of hospitalization due to heart diseases by supervised learning methods, Int. J. Med. Inform. 84 (2015) 189–197.

[13] D.J. Rubin, Hospital readmission of patients with diabetes, Curr. Diab. Rep. (2015) 15.

[14] J. Zhu, Q. Xie, K. Zheng, An improved early detection method of type-2 diabetes mellitus using multiple classifier system, Inf. Sci. (Ny). 292 (2015) 1–14.

[15] World Health Organization 2016, Global report on diabetes, 2016.

[16] J.E. Shaw, R.A. Sicree, P.Z. Zimmet, Global estimates of the prevalence of diabetes for 2010 and 2030, Diabetes Res. Clin. Pract. 87 (2010) 4–14.

[17] International Diabetes Federation, IDF Diabetes Atlas 7th Edition (2015), 2015.

[18] J.D. Donzé, M.V. Williams, E.J. Robinson, E. Zimlichman, D. Aujesky, E.E. Vasilevskis, S. Kripalani, J.P. Metlay, T. Wallington, G.S. Fletcher, A.D. Auerbach, J.L. Schnipper, International validity of the HOSPITAL Score to predict 30-day potentially avoidable hospital readmissions, JAMA Intern. Med. 176 (2016) 496–502.

[19] P.E. Cotter, V.K. Bhalla, S.J. Wallis, R.W.S. Biram, Predicting readmissions: Poor performance of the LACE index in an older UK population, Age Ageing 41 (2012) 784–789.

[20] L.L. Low, K.H. Lee, M.E. Hock Ong, S. Wang, S.Y. Tan, J. Thumboo, N. Liu, Predicting 30-day readmissions: performance of the LACE index compared with a regression model among general medicine patients in Singapore, Biomed. Res. Int. 2015 (2015).

[21] T. Cooksley, P.W.B. Nanayakkara, C.H. Nickel, C.P. Subbe, J. Kellett, R. Kidney, H. Merten, L. Van Galen, D.P. Henriksen, A.T. Lassen, M. Brabrand, Readmissions of medical patients: An external validation of two existing prediction scores, QJM 109 (2016) 245–248.

[22] V. Baskaran, A. Guergachi, R.K. Bali, R.N. Naguib, Predicting breast screening attendance using machine learning techniques, IEEE Trans. Inf. Technol. Biomed. 15 (2011) 251–259.

[23] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst 74 (2015) 150–161.

[24] R. Blagus, L. Lusa, Joint use of over-and under-sampling techniques and cross–validation for the development and assessment of prediction models, BMC Bioinformatics 16 (2015) 1–10.

[25] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Natl. Cancer Inst. 95 (2003) 14–18.

[26] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. 99 (2002) 6562–6566.

[27] J. Xia, S. Zhang, G. Cai, L. Li, Q. Pan, J. Yan, G. Ning, Adjusted weight voting algorithm for random forests in handling missing values, Pattern Recognit 69 (2017) 52–60.

[28] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Comput. Biol. Med. 59 (2015) 125–133.

[29] S.M. Kurtz, E. Lau, K. Ong, E. Adler, F. Kolisek, M. Manley, Hospital, patient, and clinical factors influence 30- and 90-day readmission following primary total hip replacement, J. Arthroplasty. 31 (2016) 1–9.

[30] M. Bach, A. Werner, J. Żywiec, W. Pluskiewicz, The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis, Inf. Sci. (Ny). 384 (2016) 174–190.

[31] J. Cervantes, F. Garcia-Lamont, A. López, L. Rodriguez, J.S.R. Castilla, A. Trueba, PSO-based method for SVM classification on skewed data sets, Neurocomputing 9227 (2015) 187–197.

[32] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[33] J.A. Sáez, J. Luengo, J. Stefanowski, F. Herrera, SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Inf. Sci. (Ny). 291 (2015) 184–203.

[34] V.F. Rodriguez-Galiano, J.A. Luque-Espinar, M. Chica-Olmo, M.P. Mendes, Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods, Sci. Total Environ 624 (2018) 661–672.

[35] S. Yin, J. Yin, Tuning kernel parameters for SVM based on expected square distance ratio, Inf. Sci. (Ny). 370–371 (2016) 92–102.

[36] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, D. Liu, Evolving support vector machines using fruit fly optimization for medical data classification, Knowledge-Based Syst 96 (2016) 61–75.

[37] I. Pillai, G. Fumera, F. Roli, Designing multi-label classifiers that maximize F measures: State of the art, Pattern Recognit 61 (2017) 394–404.

[38] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2011) 137–146.

[39] L. Vandewater, V. Brusic, W. Wilson, L. Macaulay, P. Zhang, An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of Alzheimer's disease progression, BMC Bioinformatics 16 (2015) S1.

[40] T.H.M. Le, T.T. Tran, L.K. Huynh, Identification of hindered internal rotational mode for complex chemical species: A data mining approach with multivariate logistic regression model, Chemom. Intell. Lab. Syst. 172 (2018) 10–16.

[41] C. Van Walraven, I.A. Dhalla, C. Bell, E. Etchells, I.G. Stiell, K. Zarnke, P.C. Austin, A.J. Forster, Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community, Can. Med. Assoc. J. 182 (2010) 551–557.

[42] S. Tutun, S. Khanmohammadi, Lu He, C.-A. Chou, A meta-heuristic LASSO model for diabetic readmission prediction, Proc. 2016 Ind. Syst. Eng. Res. Conf. (2016).

[43] R. Duggal, S. Shukla, S. Chandra, B. Shukla, S.K. Khatri, Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India, Int. J. Diabetes Dev. Ctries. 36 (2016) 469–476.

[44] B. Krawczyk, M. Galar, Ł. Jeleń, F. Herrera, Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. 38 (2016) 714–726.

[45] B. Zhu, B. Baesens, S.K.L.M. vanden Broucke, An empirical comparison of techniques for the class imbalance problem in churn prediction, Inf. Sci. (Ny). 408 (2017) 84–99.

[46] C. Li, S. Rana, D. Phung, S. Venkatesh, Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records, Knowledge-Based Syst 99 (2016) 168–182.

[47] A.S. Fialho, F. Cismondi, S.M. Vieira, S.R. Reti, J.M.C. Sousa, S.N. Finkelstein, Data mining using clinical physiology at discharge to predict ICU readmissions, Expert Syst. Appl. 39 (2012) 13158–13165.

[48] C. Ou-Yang, S. Agustianty, H.C. Wang, Developing a data mining approach to investigate association between physician prescription and patient outcome - A study on re-hospitalization in Stevens-Johnson Syndrome, Comput. Methods Programs Biomed. 112 (2013) 84–91.

[49] C. Walsh, G. Hripcsak, The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions, J. Biomed. Inform. 52 (2014) 418–426.

[50] M. Jovanovic, S. Radovanovic, M. Vukicevic, S. Van Poucke, B. Delibasic, Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression, Artif. Intell. Med. 72 (2016) 12–21.

[51] E. Shadmi, M. Gelkopf, P. Garber-Epstein, V. Baloush-Kleinman, R. Doudai, D. Roe, Routine patient reported outcomes as predictors of psychiatric rehospitalization, Schizophr. Res. (2017).

[52] Q.L. Huynh, K. Negishi, L. Blizzard, M. Saito, C.G. De Pasquale, J.L. Hare, D. Leung, T. Stanton, K. Sanderson, A.J. Venn, T.H. Marwick, Mild cognitive impairment predicts death and readmission within 30 days of discharge for heart failure, Int. J. Cardiol. 221 (2016) 212–217.

[53] P. Yazdan-Ashoori, S.F. Lee, Q. Ibrahim, H.G.C. Van Spall, Utility of the LACE index at the bedside in predicting 30-day readmission or death in patients hospitalized with heart failure, Am. Heart J. 179 (2016) 51–58.

[54] N. McCabe, J. Butler, S.B. Dunbar, M. Higgins, C. Reilly, Six-minute walk distance predicts 30-day readmission after acute heart failure hospitalization, Hear. Lung J. Acute Crit. Care. 46 (2017) 287–292.

[55] O. Ben-Assuli, R. Padman, M. Leshno, I. Shabtai, Analyzing hospital readmissions using creatinine results for patients with many visits, Procedia Comput. Sci. 58 (2016) 357–361.

[56] M.S. Hendryx, J.E. Russo, B. Stegner, D.G. Dyck, R.K. Ries, P. Roy-Byrne, Predicting rehospitalization and outpatient services from administration and clinical databases, J. Behav. Health Serv. Res. 30 (2003) 342–351.

[57] H. Uthoff, C. Thalhammer, M. Potocki, T. Reichlin, M. Noveanu, M. Aschwanden, D. Staub, N. Arenja, T. Socrates, R. Twerenbold, S. Mutschmann-Sanchez, C. Heinisch, K.A. Jaeger, A. Mebazaa, C. Mueller, Central venous pressure at emergency room presentation predicts cardiac rehospitalization in patients with decompensated heart failure, Eur. J. Heart Fail. 12 (2010) 469–476.

[58] I. De Domenico, D.M.V. Ward, E. Nemeth, T. Ganz, E. Corradini, F. Ferrara, G. Musci, A. Pietrangelo, J. Kaplan, Molecular and clinical correlates in iron overload associated with mutations in ferroportin, Haematologica 91 (2006) 1092–1095.

[59] S. Yamada, Y. Shimizu, M. Suzuki, T. Izumi, Functional limitations predict the risk of rehospitalization among patients with chronic heart failure, Circ. J. 76 (2012) 1654–1661.

[60] K.J. Ottenbacher, P.M. Smith, S.B. Illig, R.T. Linn, R.C. Fiedler, C.V. Granger, Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke, J. Clin. Epidemiol. 54 (2001) 1159–1165.

[61] R.E. Hodgson, M. Lewis, A.P. Boardman, Prediction of readmission to acute psychiatric units, Soc. Psychiatry Psychiatr. Epidemiol. 36 (2001) 304–309.