



OPEN

# Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes

Jingwei Hao<sup>✉</sup>, Senlin Luo & Limin Pan

Due to concealed initial symptoms, many diabetic patients are not diagnosed in time, which delays treatment. Machine learning methods have been applied to increase the diagnosis rate, but most of them are black boxes lacking interpretability. Rule extraction is usually used to turn on the black box. As the number of diabetic patients is far less than that of healthy people, the rules obtained by the existing rule extraction methods tend to identify healthy people rather than diabetic patients. To address the problem, a method for extracting reduced rules based on biased random forest and fuzzy support vector machine is proposed. Biased random forest uses the k-nearest neighbor (k-NN) algorithm to identify critical samples and generates more trees that tend to diagnose diabetes based on critical samples to improve the tendency of the generated rules for diabetic patients. In addition, the conditions and rules are reduced based on the error rate and coverage rate to enhance interpretability. Experiments on the Diabetes Medical Examination Data collected by Beijing Hospital (DMED-BH) dataset demonstrate that the proposed approach has outstanding results (MCC = 0.8802) when the rules are similar in number. Moreover, experiments on the Pima Indian Diabetes (PID) and China Health and Nutrition Survey (CHNS) datasets prove the generalization of the proposed method.

The incidence of diabetes is increasing in the world and has become a major global public health problem in the twenty-first century. Diabetes is a common chronic disease characterized by chronic hyperglycemia with carbohydrate, fat and protein metabolic disorders, which is caused by insulin secretion deficiency or insulin action deficiency. Diabetes is usually divided into type I diabetes mellitus (T1DM), type II diabetes mellitus (T2DM) and gestational diabetes mellitus. Among them, the cause of T2DM patients is mainly insulin resistance, accompanied by insufficient insulin secretion, accounting for 90–95% of diabetic patients<sup>1</sup>. According to the latest statistics from the International Diabetes Federation (IDF), approximately 8.8% of adults aged 20–79 have diabetes worldwide, of which 46.5% are undiagnosed, and approximately 5 million die each year from diabetes<sup>2</sup>. By 2040, the number of people suffering from diabetes is expected to reach 642 million. The World Health Organization (WHO) estimates that diabetes consumes 12% of global medical expenses (approximately \$673 billion). At present, the number of people with diabetes in China is approximately 110 million, and it has become the country with the largest number of patients with diabetes<sup>3</sup>. Additionally, the growth rate of direct diabetes medical expenses in China reached 19.90%, exceeding the growth rate of GDP and national total health expenditure during the same period. A study published by JAMA in 2017 showed that the average life expectancy of Chinese adult diabetic patients will be shortened by 9 years compared with those without diabetes<sup>4</sup>. Diabetes also poses a heavy economic burden on families and society while endangering human health. The initial symptoms of diabetes are relatively concealed, and it is difficult to confirm the diagnosis in the early stage of the disease. Patients are often diagnosed when the condition is serious or when there are one or more complications. According to the American Diabetes Association, early screening and detection in people with impaired blood glucose can delay the onset of diabetes through diet and exercise intervention<sup>5</sup>. In addition, early identification of high-risk groups of diabetes mellitus is also conducive to avoiding or controlling the development of complications.

In recent years, using data mining and machine learning methods to analyze collected medical data has attracted the interest of many researchers. At the same time, the accurate prediction of a disease outcome is one of the most challenging tasks for physicians. Some outstanding research has been successfully applied in breast

Information System and Security and Countermeasures Experiments Center, Beijing Institute of Technology, Beijing 100081, People's Republic of China. ✉email: 3120195386@bit.edu.cn

cancer<sup>6</sup>, hepatocellular carcinoma<sup>7</sup>, lung cancer<sup>8</sup>, and other cancer types automatic recognition and survival prediction tasks<sup>6–8</sup>. Similar to cancer research, many machine learning methods have been proposed for the effective diagnosis, prognosis, management, and treatment of diabetes<sup>9–11</sup> in the past few years. For the diagnosis and treatment of diseases, the methods used by researchers are divided into rule-based methods<sup>12–15</sup>, such as C4.5, CART, and random forest (RF), and nonrule-based methods<sup>16–19</sup>, such as support vector machine (SVM), Bayes, k-nearest neighbor (k-NN), and neural network (NN). When using a machine learning model to screen or diagnose individuals with diabetes mellitus, the model should not only have good discriminant performance but also ensure the transparency and interpretability of the discriminant process. Discriminant rules with causal logic can provide decision support for professionals, construct a model that can produce transparent discriminant rules, help people understand the internal mechanism of disease occurrence, and promote the development of disease research. Additionally, extracting and displaying the intrinsic discriminant rules of the model can help the expert question the discriminating mechanism. In addition, the transparent discrimination rule can also be checked twice by the expert to avoid errors and improve the reliability of the discrimination system<sup>18</sup>. Therefore, the interpretability of the model is of great value to ensure the usability of the diagnostic assistant system. The model used for diabetes diagnosis should have good interpretability and produce clear discriminant logic and intuitive results that can be easily understood by medical workers and patients.

To ensure the interpretability of the model, rule-based methods are usually used in existing research. The rule-based model compares its information gain based on the split value of a single feature. Although it is interpretable, it does not consider the correlation between features, which limits the discriminant performance of the model to a certain extent. The SVM is a classifier that maximizes the interval. The goal is to find a classification hyperplane that can separate samples of different categories. Due to the strong classification ability of SVM for medical data, SVM has been widely used in the diagnosis of various diseases, and its performance superiority has been confirmed<sup>21–23</sup>. However, SVM implicitly maps the input to the high-dimensional feature space in the process of calculating the hyperplane used for classification, which destroys the actual physical meaning of the input feature, resulting in the lack of interpretability, and it cannot clearly show the discriminant criteria and process. Therefore, SVM is generally used as a “black box”, which reduces its practical value, especially in the field of disease diagnosis. Considering the poor interpretability of the SVM model, the rule extraction method can be generally used to transform the model decision process into a rule set to improve the interpretability of the model.

This paper proposes a method for extracting the reduced rule from the fuzzy SVM and the biased random forest (BRF). First, we build the fuzzy SVM model with acceptable accuracy and extract the support vectors (SVs) from the SVM. Then, the fuzzy SVM is used to predict the SV labels. The SVs and predicted labels make up the artificial dataset. The artificial dataset is provided to the BRF to generate rules. Finally, the rule reduction module is introduced to remove redundant conditions and rules and improve the interpretability of the obtained rules.

The experimental results show that the proposed method generates more succinct and accurate rules than other methods, which is helpful for a broader assessment of diabetic patients. In addition, the results of the study indicate that the method can be used as a tool to detect diabetes and its associated risk factors. In summary, this work has the following major contributions:

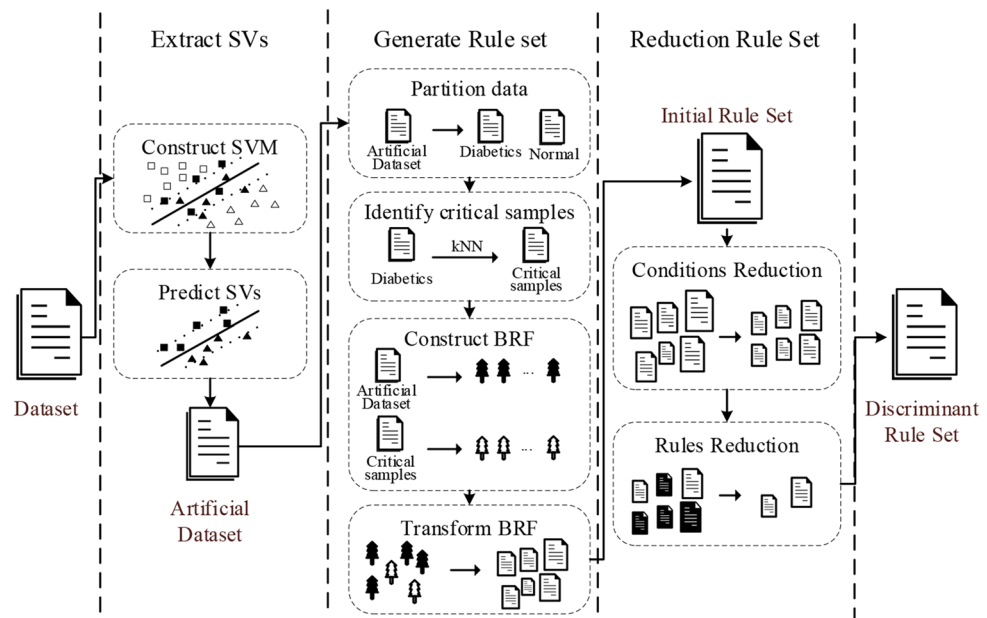
- Developing a hybrid framework based on reduced rules extracted by BRF.
- It is proposed to utilize BRF to deal with the problem of data imbalance caused by diabetic patients far less than normal people.
- A reduction method based on the error rate and coverage rate is developed to remove the problems of similar, repetitive, and inefficient conditions and rules caused by the independent learning of each tree in the ensemble method.

The rest of this paper is organized as follows. The second section discusses the related work of SVM rule extraction. In third section, first, the framework of the algorithm is introduced, and then the algorithm is introduced in detail. The fourth section introduces the dataset and the experimental process. In fifth section, the experimental results are discussed. Finally, the last section is the conclusion.

## Related work

To achieve early detection and early intervention of diabetic patients, many methods have been proposed in recent years. Nilashi et al. used the EM method to cluster data, applied the PCA method to reduce the data dimensionality, filtered out the potential noise, and applied CART to find the decision rules from diabetes data<sup>11</sup>. Patil et al. proposed the HPM method, using C4.5 to classify the data denoised by the k-means clustering algorithm<sup>24</sup>. Due to the tree structure of CART, C4.5, and other decision tree models, the classification process is transparent, but they are weak classifiers. To improve the classification effect, the model with stronger learning ability is used. SVM has attracted attention for the diagnosis of diabetes due to its excellent classification ability. Shen et al. proposed an SVM parameter adjustment method using a fruit fly optimization algorithm and applied it to diabetes diagnosis<sup>25</sup>. It was verified that the method can obtain more suitable model parameters and greatly reduce the calculation time compared with other SVM parameter adjustment methods. Santhanam et al. used k-means to remove noise data, used a genetic algorithm to find the best feature set, and used SVM as a classifier to classify the diabetes data<sup>26</sup>. Uzer et al. proposed using an artificial bee colony algorithm for feature selection to eliminate the influence of unimportant features on SVM classification results<sup>27</sup>. Choubey et al. compared the effects of SVM using different kernel functions in the diagnosis of diabetes and used genetic algorithms to eliminate redundant features to reduce calculation costs and improve classification accuracy<sup>28</sup>.

SVM has a rigorous statistical learning theoretical basis, which can better solve the problems of overfitting, local minima, and dimension disasters. However, its classification process is not transparent, and it is used as a



**Figure 1.** Schematic diagram of the proposed method.

black box, which reduces its reliability. Rule extraction is an effective technique to solve this problem. At present, rule extraction for SVM can be divided into three categories: decomposition methods, pedagogical methods and eclectic methods<sup>29</sup>. The basic idea of the decomposition method is to decompose the SVM into several sets in units of SVs, search and extract rules for each SV, and finally combine these rules, such as SVM + prototype<sup>30</sup> and HRE<sup>31</sup>. The pedagogical method does not consider the type and structure information of the SVM, ignores the knowledge provided by the SVs or decision boundary of the SVM, only pays attention to the mapping result of the SVM input–output, and uses the SVM as a “black box” to extract rules from the SVM prediction labels by the rule generation method. Other machine learning algorithms are used to extract rules, such as the GEX and G-REX algorithms, which generate rule sets using algorithms such as C4.5, CART, and Bayesian trees<sup>32</sup>. The advantage of this algorithm is that it is highly versatile. It is different from the decomposition method, which is usually applied to the linear SVM model. The pedagogical method is not limited by the type and structure of the SVM. However, the rule set is too large due to the use of all data generation rules. The eclectic method combines the advantages of the pedagogical method and the decomposition method, makes full use of the SV information in the SVM, and can also use a rule generation model to extract rules. To some extent, the SVM decision function information is considered, and the number of generated rules is also reduced. Han et al. proposed the SVM + RF algorithm, which uses random forests to generate rules from artificial datasets constructed from SVs<sup>33</sup>. The rules extracted by this method have good accuracy. However, the rules generated by the ensemble method are similar or even repeated, which harms the interpretability. Liu et al.<sup>34</sup> and Khanam et al.<sup>35</sup> used CART to extract rules from the SVM. Deshmukh et al.<sup>36</sup> developed a hybrid fuzzy deep learning approach for the detection of diabetes. Firstly, the data was fuzzified. After that, a  $5 \times 5$  fuzzy matrix was constructed. Lastly, the fuzzy matrix was fed into the convolution neural network (CNN). The results demonstrated that the fuzzified CNN approach outperformed the traditional NN approach. Azad et al.<sup>37</sup> proposed a model PMSGD to classify diabetes. Synthetic minority over-sampling technique (SMOTE), genetic algorithm (GA), and DT were used in the proposed model. Wang et al.<sup>38</sup> deleted the repeated rules and the repeated conditions in the rules to obtain a more concise rule set. Hayashi et al.<sup>39</sup> proposed to combine rule extraction algorithm and sampling selection technique to achieve interpretable and accurate classification rules for PID data set. Similarly, Chakraborty et al.<sup>40</sup> proposed the eclectic rule extraction from neural network recursively (ERENNR) algorithm, which generated rules from dataset with mixed attributes in the guise of attribute data ranges.

Overall, Han et al. noted that the eclectic method can reduce the degree of imbalance in the dataset<sup>33</sup>, but the effect is limited. Most of the existing rule extraction methods do not consider how to deal with the imbalance problem that is prevalent in medical datasets. In addition, the rules extracted by ensemble learning methods are redundant, which improves the risk of model overfitting. Using a decision tree to extract rules, because the model is generated by heuristic learning, there is a problem that cannot effectively minimize the global training error. To solve the above problems, a method for extracting reduced rules from SVM based on biased random forest is proposed.

### Proposed method

In this section, the proposed rule extraction method is introduced. Figure 1 shows the algorithmic principle of the method for extracting reduced rules from SVM based on biased random forest. First, the SVM model is constructed by using the data preprocessed training set, and the hyperparameters are tuned to make the model

have acceptable classification performance. Extracting the SVs, the richest information points containing partitioning patterns from SVM. The SVs are predicted by the trained SVM to obtain the labels. The SVs and their labels make up the artificial data to eliminate label noise. Then, the potential distribution of the artificial dataset is inferred through BRF, and each tree is traversed from the root node to the leaf node to generate “if-then” rules. Finally, the rule set generated by the BRF is reduced to obtain the discriminant rule set.

**Extract SVs.** The purpose of the fuzzy SVM is to find the optimal hyperplane that can separate samples of different classes, while the hyperplane meets the constraints of maximizing samples and hyperplane spacing. In essence, fuzzy logic is used to classify the level of risks from data, SVM is used to design the fuzzy rules, and the dataset is used to train the SVM using Linear Parameter and test the fuzzy system. Finding the classification hyperplane can be transformed into a convex optimization problem:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \quad \text{s.t. } y_i ((\mathbf{w})^T \varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i, i = 1, 2, \dots, N \quad (1)$$

$\xi_i$  is a relaxation variable, which converts hard interval maximization into soft interval maximization.  $C$  is the penalty factor to represent the penalty size of the misclassified samples.  $\varphi(\cdot)$  indicates that the kernel technique is used to map the input space into the high-dimensional space, which can transform the linear indivisible problem into a linearly separable solution problem in high-dimensional space. Spatial mapping is usually implemented by the radial basis function (RBF):

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

$\|\mathbf{x} - \mathbf{x}'\|^2$  represents the Euclidean distance between two vectors.  $\sigma$  is a tunable parameter; the smaller  $\sigma$  is, the more SVs there are, and the easier the model is overfitted.

To simplify the solution, the Lagrange multiplier  $\alpha_i$  is introduced. By using the Lagrange dual property, the solution of Formula (1) is transformed into its dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0, i = 1, 2, \dots, N \\ & \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned} \quad (3)$$

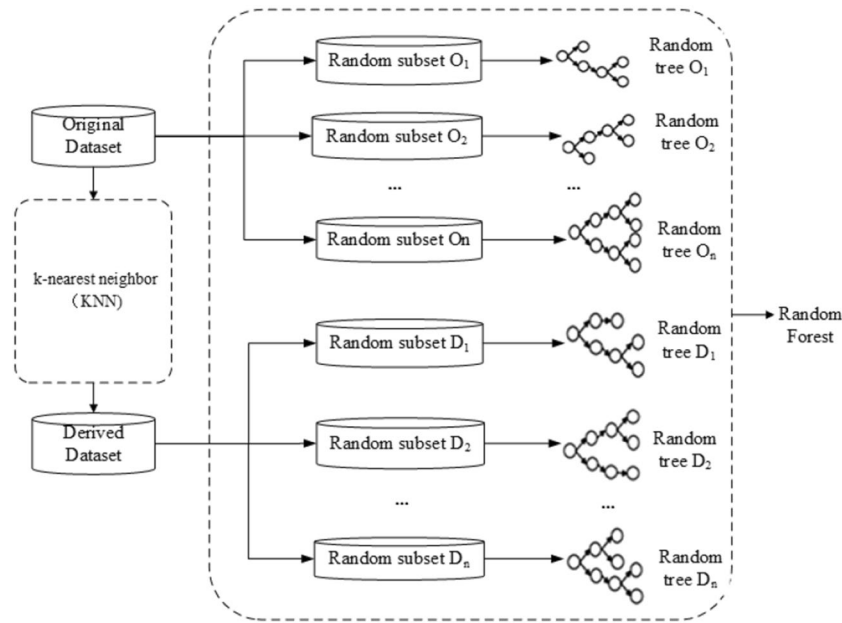
The gradient descent method is used to solve  $\alpha_i$ . Then, the SVM classification decision function can be written as:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{SV} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (4)$$

**SVs** An SV is a sample of training data corresponding to a Lagrange multiplier greater than 0. Formula (4) shows that the discriminant result of the SVM discriminant model for new samples is entirely determined by SVs, and discriminant rule set extraction using SVs can retain the discriminant effect of the SVM model to a large extent. Through Formula (4), the researchers can prove that the rules in SVM are implied in SVs or decision boundaries. Therefore, rule extraction from SVM can be transformed into rule extraction from SVs. The complexity of computation depends on the number of SVs, not the dimension of the sample space, which avoids the “dimension disaster” in a sense and reduces the number of rules generated by rule extraction. It is worth noting that to strengthen the output accuracy, fuzzy SVM is used to optimize the traditional SVM classifier. Fuzzy SVM is able to emphasize the support vector node to avoid any redundant training since the crisp sets will be converted to fuzzy sets.

**Generation rule set.** Figure 2 shows the schematic diagram of BRF. It is an ensemble method to alleviate the data imbalance by increasing the number of classifiers representing the minority class<sup>41</sup>. Compared with RF, BRF defines the minority samples and their k-nearest neighbors as critical samples. For this part of the samples, more tree models are generated for classification. Move the sampling operation from the data level to the model level to obtain better results in imbalanced data classification. In the diagnosis of diabetes, the number of diabetic patients is far less than that of healthy people, which leads to an imbalance of the collected dataset. Although in the previous step, the imbalance problem of the artificial dataset constructed by SVs is slightly alleviated compared with that of the training dataset, the problem still exists and cannot be ignored. Taking advantage of BRF to generate rule sets is better than other ensemble learning methods due to its adaptability to imbalanced data.

Specifically, the dataset is first divided into a majority class set (normal) and a minority class set (diabetics). Then, the k-NN algorithm is used to find the k-nearest neighbors in the majority class set for each sample in the minority class set. If one sample in the majority class set appears repeatedly, only one is retained. The minority class set and the k-nearest neighbors in the majority class set form a new dataset. In addition to using the undivided dataset to build a random forest, the new dataset is also used to build a random forest. These forests are combined to obtain the final BRF. BRF can be seen as a method to learn from the original dataset and the



**Figure 2.** Schematic diagram of BRF.

undersampling subdataset generated from the original dataset. This kind of bias to the minority class compensates for its low presence in the dataset to overcome the data imbalance problem.

Rule generation is divided into two steps. First, the BRF model is induced based on an artificial dataset. Then, according to the BRF model, each tree is searched from the root node to the leaf node to extract the rules. The rules extracted from all trees are combined to form the initial rule set.

**Reduction rule set.** The rules contained in the initial rule set have the problem of redundancy. The problem increases the risk of the rule set overfitting and affects the practicability of the rule set. Therefore, it is necessary to simplify the rule set. The reduction includes two steps: the first step is to remove the redundant conditions, and the second step is to reduce the redundant rules.

First, let the initial rule set be  $\mathfrak{R}_{initial} = \{R_i \rightarrow L_i, i = 1, 2, \dots, K\}$ , where  $K$  is the number of rules,  $R_i$  is the  $i$  th discrimination rule, and  $L_i$  is the label corresponding to the  $i$  th rule. Rules consist of multiple conditions, such as  $f_1 = v_1 \& f_2 \leq v_2 \& f_3 \geq v_3 \cdots \& f_j \geq v_j \rightarrow 1$ , where  $f_j$  represents the  $j$  th attribute in the rule, and  $v_j$  represents the value of  $f_j$ . The pruning rule  $R_i$ , according to the removal of a certain condition, calculates the change in the error rate of rule  $R_i$  to the sample to determine whether the condition should be removed, and the specific calculation formula is as follows:

$$D_j = \frac{err_{-j} - err_0}{\max(err_0, s)} \tag{5}$$

In the formula,  $err_0$  and  $err_{-j}$  indicate the discrimination error rate of rule  $R_i$  before and after, respectively, the  $j$  th condition is removed. It should be noted that the discriminant error rate of the rule is the proportion of the misjudged samples in the samples satisfying the rule.  $s$  is a normal number to constrain the size of  $D_j$ . Set a threshold value (0.05 here). If  $D_j$  is less than the threshold value, it denotes that the  $j$  th condition has little impact on the discrimination. It should be removed from  $R_i$  and updated with  $err_0$ . Otherwise, the condition is kept, and the next condition is evaluated. After all the rules in the initial rule set are processed, the conditions reduced rule set  $\mathfrak{R} = \{R'_i \rightarrow L_i, i = 1, 2, \dots, K\}$  is obtained, where  $R'_i$  is the reduced rule  $R_i$ .

The next step is to reduce the redundant rules. First, an empty set  $\mathfrak{R}'_{final} = \{\}$  is constructed to store the filtered rule set. Then, the rule set  $\mathfrak{R}$  is roughly screened by rule coverage, which is expressed as:

$$freq = \frac{N_{R'_i}}{N} \tag{6}$$

where  $N_{R'_i}$  represents the number of training samples that meet rule  $R'_i$ , and  $N$  represents the total number of training samples. Set the threshold  $g$ , and remove the rules whose coverage is less than  $g$  from  $\mathfrak{R}$ . At the same time, a default rule  $R_{def} = \{\} \rightarrow L^*$  is built, where  $L^*$  represents the label with the largest number of samples in the training set. Remove the rules with low coverage in rule set  $\mathfrak{R}$ , and add  $R_{def}$  to form rule set  $\mathfrak{R}'$ . Then, the training dataset and rule set  $\mathfrak{R}'$  are used to filter the rules iteratively, in which rule  $R_{best}$  with the minimum discrimination error rate is selected into  $\mathfrak{R}'_{final}$  for each iteration, the samples satisfying rule  $R_{best}$  are removed from the training dataset,  $R_{best}$  is removed from  $\mathfrak{R}'$ , and the output label  $L^*$  of the default rule  $R_{def}$  is updated according

---

Input :  $\mathfrak{R} = \{R'_i \rightarrow L_i, i = 1, 2, \dots, K\}$ ,  $R_{def} \leftarrow \{\}$   $\rightarrow L^*$ ,  $\mathfrak{R}_{final} \leftarrow \{\}$ , training dataset

Output :  $\mathfrak{R}_{final}$

---

1. Calculate the coverage of each rule  $freq_i, i = 1, 2, \dots, K$
  2. Eliminate the rules  $freq_i < g$  in  $\mathfrak{R}$ , and add  $R_{def}$  to  $\mathfrak{R}$  to construct  $\mathfrak{R}'$
  3. repeat
  4. Calculate the error rate of each rule in  $\mathfrak{R}'$  and extract the rule  $R_{best}$  with the lowest error rate
  5. if  $R_{best} = R_{def}$  then
  6. Add  $R_{def}$  to  $\mathfrak{R}_{final}$
  7. end if
  8. If training dataset  $= \phi$  then
  9. Update the label of  $R_{def}$  to  $L^*$ , and add  $R_{def}$  sequentially to  $\mathfrak{R}_{final}$
  10. end if
  11. Add  $R_{best}$  sequentially to  $\mathfrak{R}_{final}$
  12. Update  $R_{def}$ ,  $\mathfrak{R}'_{initial}$ , training dataset
  13. until 1
  14. return  $\mathfrak{R}_{final}$
- 

**Table 1.** Pseudo-code of reduce the redundant rules.

to the updated training dataset. Finally, when the rule  $R_{best}$  selected is  $R_{def}$  or the training dataset is empty, the iterative process of rule filtering is stopped. If  $R_{best}$  is the default rule, add the default rule to  $\mathfrak{R}_{final}$ . If the training dataset is empty, update the output label of the default rule to the initial value, and add the rule to  $\mathfrak{R}_{final}$ .  $\mathfrak{R}_{final}$  is the final set of reduced discriminant rules. The pseudocode for reducing the redundant rules is shown in Table 1.

## Experiments

In this study, a new interpretability approach for rule extraction from the fuzzy SVM is proposed. This technology integrates the information provided by the SVs of the SVM model into the BRF method to extract rules from the black box SVM model and reduces the conditions and rules to improve interpretability. First, to verify the rule extraction motivation from the SVM, the SVM is compared with the RF, C4.5, ID3, CART, and RIPPER methods. Then, SVM + BRF (not reduced) and fuzzySVM + BRF (not reduced) are compared with SVM + RF<sup>33,35</sup>. Finally, the proposed method is compared with Re-RX + J48graft(2016)<sup>36</sup>, Fuzzy + CNN(2019)<sup>33</sup>, ERENNR(2019)<sup>37</sup>, SVM + XGBoost(2019)<sup>17</sup>, RF + XGBoost(2021)<sup>16</sup>, and PMSGD(2019)<sup>34</sup> methods. Finally, all methods are tested on the test set.

**Dataset.** The experimental data are the Diabetes Medical Examination Data collected by Beijing Hospital (DMED-BH). The DMED-BH consists of 17 features, which can be divided into noninvasive and invasive features, including routine physical examination indicators, blood test indicators and questionnaire survey indicators. According to WHO's definition of diabetes, the fasting blood glucose  $\geq 7.0$  mmol/L and/or the postprandial blood glucose  $\geq 11.1$  mmol/L, 6503 people are marked as diabetes positive and 36,853 people are marked as diabetes negative.

**Experiment environment.** Table 2 shows the experiment environment. Each model is implemented using Tensorflow v2.2.0 and trained by a personal computer with Intel i7-8750 CPU, 8 GB RAM, Windows 10 operating system.

**Evaluation metrics.** The primary goal of rule extraction is to improve the interpretability of the model while ensuring the discriminant ability of the model. In particular, in the diagnosis of diabetes, we cannot blindly pursue interpretability without considering the accuracy of the discriminant results. For all methods, accuracy, precision, recall, F1- measure and Mathews correlation coefficient (MCC) are used as evaluation metrics in the experiment, and the number of rules is also used to measure the interpretability of rules. Formulas for these metrics are shown in Eqs. (7)–(11).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Environment	Description
Computer	Intel(R) Core(TM) i7-8750H CPU
Application platform	Windows10
Software	R 4.0.3 Tensorflow v2.2.0

**Table 2.** Experiment environment.

Features	Chi square (p-value)	IG	RF
AGE	140.5 (0.0000)	0.071	12.399
WEIGHT	559.45 (0.0000)	0.259	10.529
HEIGHT	392.51 (0.0000)	0.192	10.224
CHOL	406.19 (0.0021)	0.196	9.844
TG	415.31 (0.0000)	0.201	11.041
HDL	221.72 (0.0000)	0.118	13.174
LDL	391.37 (0.0001)	0.190	10.617
SBP	175.24 (0.0000)	0.092	10.855
DBP	108.89 (0.0000)	0.056	9.998

**Table 3.** Feature selection results.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (11)$$

where  $TP$  indicates the true positive frequency,  $FP$  indicates the false positive frequency,  $TN$  indicates the true negative frequency, and  $FN$  indicates the false negative frequency.  $F1$  is the weighted harmonic average of *precision* and *recall* and gives them the same weight.  $MCC$  is considered to be a relatively balanced metric, which can be applied even when the data are imbalanced.

**Feature selection.** Many machine learning methods may lead to worse performance because of a large number of redundant features. Feature selection has important practical significance<sup>42</sup>. It not only reduces overfitting, reduces the number of features, and improves the generalization ability of the model but also accelerates the training speed of the model. Generally, feature selection can improve the model performance. Therefore, the filtering method and embedding method are used for feature selection. Among them, the filtering method uses the chi square test and information gain, and the embedding method is realized by RF.

The chi square test is one of the commonly used methods for feature selection to determine whether the two variables are independent by observing the deviation between the actual value and the theoretical value<sup>43</sup>. In addition to the chi square test, information gain is also a very effective feature selection method. Unlike the chi square test, which uses correlation between features and labels to quantify the importance of features, information gain is based on the amount of feature information<sup>44</sup>. Random forest is a typical ensemble learning method that is often used for feature selection<sup>45</sup>. The idea is to compare the contribution of each feature in random forest; the greater the contribution is, the more important the feature. Generally, the Gini index is used to measure the contribution of features<sup>46</sup>.

Considering the effect and efficiency of diabetes diagnosis, the features evaluated by the chi square test, information gain and RF are ranked, and the average rank is calculated. The top 9 features with the highest average rank and statistical significance (p value < 0.05) were selected to build the models. They are AGE, WEIGHT, HEIGHT, CHOL (cholesterol), TG (triglyceride), HDL (high-density lipoprotein), LDL (low-density lipoprotein), SBP (systolic blood pressure) and DBP (diastolic blood pressure). The result of feature selection is shown in Table 3.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC
SVM	<b>98.84</b>	<b>98.67</b>	<b>61.77</b>	<b>0.7496</b>	<b>0.7472</b>
RF	92.70	92.36	60.54	0.7351	0.7210
C4.5	50.60	41.52	40.77	0.4061	0.3088
ID3	46.81	45.76	37.69	0.4049	0.3285
CART	44.86	44.09	28.46	0.3361	0.2624
RIPPER	49.99	49.21	22.31	0.2705	0.2391

**Table 4.** Average results of fivefold CV for positive class. Significant values are in bold.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC
SVM + RF	91.96	89.87	83.77	0.8671	0.8258
SVM + BRF(ours)	<b>96.84</b>	<b>94.78</b>	<b>90.08</b>	<b>0.9237</b>	<b>0.8428</b>
FuzzySVM + BRF(ours)	<b>98.73</b>	<b>96.83</b>	<b>93.01</b>	<b>0.9488</b>	<b>0.8524</b>

**Table 5.** Average results of fivefold CV for ensemble methods. Significant values are in bold.

**Rule extraction performance.** To obtain reliable and stable models, fivefold cross validation (fivefold CV) is used to determine the model parameters and test models. The dataset is randomly and evenly divided into 5 parts, one of which is used as the test set, one of which is used as the validation set, and the remaining three parts are used as the training set. The training set is used to train the SVM, the validation set is used to evaluate the performance of the model under different hyperparameters, and the test set is used to evaluate the performance of the SVM using the hyperparameters that perform best on the validation set. First, through grid search, the optimal hyperparameters (gamma and cost) of the SVM are 1.5 and 4. It is worth noting that the SVM uses the radial basis function (RBF) as the kernel function and normalizes the data to [0,1] during training. Then, the SVM is trained on the new training set consisting of the training set and the validation set, and the test results are obtained on the test set. This process is also carried out in fivefold CV.

In addition, to prove the motivation for extracting rules from SVM, RF, C4.5, ID3, CART and RIPPER are used as the comparison methods. As with SVM, these methods are adjusted by fivefold CV to obtain test results. The performance of these models is evaluated by accuracy, precision, recall, F1-measure and MCC. The results are shown in Table 4.

In the process of fivefold CV, SVs are extracted from the trained SVM model. The average number of SVs is 653.5 (standard deviation is 4.394), and the average ratio of positive and negative diabetes in SVs is 1:5.45 (standard deviation is 0.406), which is slightly lower than the ratio of 1:5.7 in the original dataset, but the imbalance problem still exists. This is the motivation for using the BRF method, which can effectively deal with the imbalance problem to extract rules. The SVs and prediction results via the SVM are combined into an artificial dataset. The new dataset is used to extract rules from the SVM, training rule-based learners to obtain rules that can express the connotation of the SVM. RF, which is an ensemble method similar to BRF, is used as a comparison method. The results are shown in Table 5.

The rules obtained from BRF are reduced by the method in Sect. 3.3. The reduced rule sets (FuzzySVM + BRF + reduced, SVM + BRF + reduced) are compared with the rule sets reduced by the Re-RX + J48graft(2016)<sup>36</sup>, Fuzzy + CNN(2019)<sup>33</sup>, ERENNR(2019)<sup>37</sup>, SVM + XGBoost(2019)<sup>17</sup>, RF + XGBoost(2021)<sup>16</sup>, and PMSGD(2019)<sup>34</sup> methods. We tested these comparison methods on DMED-BH dataset. In addition to using accuracy, precision, recall, F1-measure and MCC to evaluate the rule set performance in the diagnosis of diabetes, the number of rules is also used to represent the interpretability of rules. The results are shown in Table 6.

**Generality analysis.** To verify the generality of the proposed method, two open datasets related to diabetes were selected and tested. The selected datasets are described as follows:

Pima Indian Diabetes (PID)<sup>47</sup>. A PID dataset was used to test the effectiveness of various diagnostic methods for diabetes. There are 768 samples in the dataset (268 cases 1 and 500 cases 0), and the ratio of positive samples to negative samples is 1:1.87. Each sample is represented by 8 features: pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.

China Health and Nutrition Survey (CHNS)<sup>45,46</sup>. The CHNS dataset was collected by the Chinese Center for Disease Control and Prevention and the Carolina Population Center at the University of North Carolina at Chapel Hill. The dataset we selected was collected in 2009, covering nine provinces in China. After data preprocessing, there were 7,913 samples. The samples with fasting blood glucose  $\geq 7.0$  mmol/L were labeled positive samples, and the ratio of positive samples to negative samples was 1:11.2. After feature selection, a total of 9 features were used in the experiment: WEIGHT, AGE, WAIST, DBP, CHOL, TG, HBA1C, UA, and HDL.



Methods	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC	Rules
Re-RX + J48graft(2016) <sup>36</sup>	83.96	83.25	85.38	0.8430	<b>0.8726</b>	<b>8.2 ± 1.0</b>
Fuzzy + CNN(2019) <sup>33</sup>	<b>94.74</b>	<b>94.94</b>	<b>93.02</b>	<b>0.9397</b>	0.8626	29.3 ± 1.1
ERENNR(2019) <sup>37</sup>	83.71	81.25	83.96	0.8258	0.8617	72.4 ± 6.0
SVM + XGBoost(2019) <sup>17</sup>	90.89	89.53	82.26	0.8574	0.8534	13.4 ± 7.6
RF + XGBoost(2021) <sup>16</sup>	90.93	89.22	85.47	0.8730	0.8722	18.5 ± 2.9
PMSGD(2021) <sup>34</sup>	83.86	82.16	85.47	0.8378	<b>0.8759</b>	21.5 ± 3.7
SVM + BRF + reduced (ours)	<b>95.75</b>	<b>93.50</b>	<b>91.49</b>	<b>0.9248</b>	0.8653	<b>12.4 ± 2.5</b>
FuzzySVM + BRF + reduced (ours)	<b>96.92</b>	<b>94.81</b>	<b>93.11</b>	<b>0.9395</b>	<b>0.8802</b>	<b>9.2 ± 1.6</b>

**Table 6.** Average results of fivefold CV for extracted rule sets on DMED-BH dataset. Significant values are in bold.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC	Rules
Re-RX + J48graft(2016) <sup>36</sup>	84.93	83.83	78.64	0.8115	<b>0.8796</b>	<b>8.2 ± 2.1</b>
Fuzzy + CNN(2019) <sup>33</sup>	<b>95.74</b>	<b>95.00</b>	<b>92.01</b>	<b>0.9348</b>	0.8355	28.9 ± 9.5
ERENNR(2019) <sup>37</sup>	84.71	83.12	81.56	0.8233	0.8518	79.1 ± 6.9
SVM + XGBoost(2019) <sup>17</sup>	76.77	75.32	73.62	0.7446	0.7747	23.2 ± 2.7
RF + XGBoost(2021) <sup>16</sup>	89.60	88.32	86.55	0.8742	0.8578	19.4 ± 0.8
PMSGD(2021) <sup>34</sup>	83.64	82.13	80.09	0.8109	0.84026	24.7 ± 3.1
SVM + BRF + reduced (ours)	<b>95.92</b>	<b>94.78</b>	<b>92.95</b>	<b>0.9385</b>	<b>0.8746</b>	<b>18.2 ± 1.5</b>
FuzzySVM + BRF + reduced (ours)	<b>96.84</b>	<b>95.23</b>	<b>93.64</b>	<b>0.9442</b>	<b>0.8752</b>	<b>8.9 ± 1.3</b>

**Table 7.** Average results of fivefold CV for extracted rule sets on PID dataset. Significant values are in bold.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC	Rules
Re-RX + J48graft(2016) <sup>36</sup>	83.59	80.87	79.56	0.8020	0.7794	<b>7.6 ± 2.0</b>
Fuzzy + CNN(2019) <sup>33</sup>	<b>93.46</b>	<b>92.83</b>	<b>92.22</b>	<b>0.9252</b>	<b>0.8857</b>	23.5 ± 8.0
ERENNR(2019) <sup>37</sup>	84.76	83.21	82.30	0.8275	0.8227	76.4 ± 3.2
SVM + XGBoost(2019) <sup>17</sup>	81.91	79.37	71.54	0.7525	0.7790	24.5 ± 2.8
RF + XGBoost(2021) <sup>16</sup>	88.64	87.36	87.55	0.8745	0.8401	19.4 ± 0.8
PMSGD(2021) <sup>34</sup>	90.85	89.15	83.94	0.8646	<b>0.8775</b>	17.8 ± 2.1
SVM + BRF + reduced (ours)	<b>93.62</b>	<b>92.78</b>	<b>91.44</b>	<b>0.9210</b>	0.8122	<b>13.3 ± 0.7</b>
FuzzySVM + BRF + reduced (ours)	<b>94.92</b>	<b>92.87</b>	<b>92.98</b>	<b>0.9292</b>	<b>0.8956</b>	<b>7.2 ± 1.6</b>

**Table 8.** Average results of fivefold CV for extracted rule sets on CHNS dataset. Significant values are in bold.

Five-fold cross validation was carried out according to the process in “Feature selection”, and some experimental results were extracted from their original paper. The summarized experimental results are shown in Tables 7 and 8.

## Discussion

The main purpose of this study was to achieve a diabetes diagnosis. The models and rule sets are evaluated by accuracy, precision, recall, F1-measure and MCC. Among them, in the disease diagnosis field, false negatives need to be minimized, and the dataset has the characteristics of class imbalance, so recall and MCC should be given priority<sup>48,49</sup>.

In Table 4, compared with rule-based classifiers such as RF, C4.5, ID3, CART, and RIPPER, SVM has the highest accuracy, precision, recall rate, F1-measure, and MCC, which proves that SVM has better performance than the rule-based models. The results also demonstrate the rationality of our motivation to choose SVM as the basic classifier for diabetes detection. In Table 5, the rule sets extracted by BRF are superior to the rule sets extracted by RF in all indicators. After fuzzy logic is combined, our method achieves a better separation effect. Since fuzzy SVM can highlight support vector nodes to minimize duplicate training and meet the goal of improving output accuracy. In Table 6, compared with the six rule extraction models, except the fuzzy + CNN method, our method has obvious advantages in accuracy, precision, recall rate, F1-measure, and the number of MCC and reduction rules. Furthermore, while the fuzzy + CNN method has high accuracy, precision, and recall rate, the

classifier tends to select the majority classes due to the naturally imbalanced character of the diabetes dataset. As a result, these indicators cannot accurately reflect the classifier's performance. Because MCC has little to do with the distribution of positive and negative samples, we focus more on MCC value comparison. In this way, fuzzy SVM + BRF outperforms fuzzy + CNN. It is worth mentioning that although the PMSGD method does not have high accuracy and rule reduction effect, it also has good classification performance on imbalanced data sets. The rule reduction number of the Re-Rx + J48graft method is also ideal, but the classification effect is not as good as our method in the diabetes prediction task. Tables 7 and 8 provide similar experimental results to Table 6, indicating that the proposed method also performs well on different data sets, proving the method's generality.

In summary, the proposed method can adapt to imbalanced data and extract rules that tend to diagnose patients with diabetes and further enhance interpretability by reducing rules. It is an effective method to extract rules from SVM for diabetes diagnosis.

Needless to say, the diagnosis of diabetes remains a complex problem; therefore, the fuzzySVM + BRF method should be tested on more recent and complete diabetes datasets in future studies to ensure that the most highly accurate rules can be extracted for diagnosis.

## Conclusion

Diabetes mellitus is a common chronic disease that seriously endangers human health. In recent years, machine learning methods have been widely used in diabetes diagnosis. Fuzzy SVM can emphasize support vector nodes, avoid redundant training, and simplify classification without sacrificing classification accuracy. Although fuzzy SVM has achieved great discrimination effects, the lack of interpretability due to mapping features to high-dimensional spaces during the classification process limits its application in the field of disease diagnosis. Therefore, it is necessary to extract rules for SVM. Considering the poor adaptability of the existing rule extraction methods to imbalanced data, the extracted rules tend to identify healthy people, and the BRF with a reduction module was proposed for rule extraction to solve the problem. First, the support vectors are extracted from the SVM model with acceptable classification ability, and the SVM is used to predict the support vectors. The support vectors and prediction results constitute an artificial dataset. Then, the critical samples are defined by the k-NN algorithm. Based on the critical samples, more trees are generated to be a part of the BRF. BRF is used to infer the potential distribution of the artificial dataset and obtain the initial rule set. Finally, the rule set is reduced to obtain the final rule set. The extracted rule set provides a basis for early intervention measures for diabetic patients and control of diabetes.

The experimental results show that the proposed model performs well in the four metrics of accuracy, recall, F1-measure, and MCC when the sizes of the rule sets are almost the same. This shows that the model is promising in diabetes diagnosis. A possible extension of this work is to consider how to generate the rule set to improve the accuracy, while maintaining recall.

Received: 4 November 2021; Accepted: 15 April 2022

Published online: 14 June 2022

## References

- Zhao, M., Wang, X. & Zhu, X. Understanding diabetes from the diagnosis of diabetes mellitus [J]. *J. Diagn. Concepts Pract.* **2**, 226–228 (2014).
- Cho, N., Whiting, D., & Forouhi, N. IDF Diabetes Atlas [R]. Brussels, Belgium: International Diabetes Federation (2016).
- Chinese Diabetes Society. Guidelines for the prevention and control of type 2 diabetes in China [J]. *Chin. J. Pract. Internal Med.* **38**(4), 292–344 (2018).
- Bragg, F., Holmes, M. V. & Iona, A. Association between diabetes and cause-specific mortality in rural and urban areas of China [J]. *J. Am. Med. Assoc.* **317**(3), 280–289 (2017).
- Patil, R. *et al.* A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus[J]. *Int. J. Electr. Comput. Eng.* **12**(1), 524 (2022).
- Montazeri, M. *et al.* Machine learning models in breast cancer survival prediction[J]. *Technol. Health Care* **24**(1), 31–42 (2016).
- Książek, W., Gandor, M. & Pławiak, P. Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma[J]. *Comput. Biol. Med.* **134**, 104431 (2021).
- Doppalapudi, S., Qiu, R. G. & Badr, Y. Lung cancer survival period prediction and understanding: Deep learning approaches[J]. *Int. J. Med. Informatics* **148**, 104371 (2021).
- Faura, G. *et al.* Colorimetric and electrochemical screening for early detection of diabetes mellitus and diabetic retinopathy—application of sensor arrays and machine learning[J]. *Sensors* **22**(3), 718 (2022).
- Choubey, D. K. *et al.* Classification of diabetes by kernel based SVM with PSO[J]. *Recent Adv. Comput. Sci. Commun.* **14**(4), 1242–1255 (2021).
- Dremin, V. *et al.* Skin complications of diabetes mellitus revealed by polarized hyperspectral imaging and machine learning[J]. *IEEE Trans. Med. Imaging* **40**(4), 1207–1216 (2021).
- Latchoumi, T. P., Dayanika, J. & Archana, G. A comparative study of machine learning algorithms using quick-witted diabetic prevention [J]. *Ann. Roman. Soc. Cell Biol.* **1**, 4249–4259 (2021).
- Tiddi, I. & Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey[J]. *Artif. Intell.* **302**, 103627 (2022).
- Du, Y. *et al.* An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus [J]. *Sci. Rep.* **12**(1), 1–14 (2022).
- Tama, B. A. & Rhee, K. H. Tree-based classifier ensembles for early detection method of diabetes: An exploratory study [J]. *Artif. Intell. Rev.* **51**(3), 355–370 (2019).
- Kumari, S., Kumar, D. & Mittal, M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier[J]. *Int. J. Cognit. Comput. Eng.* **2**, 40–46 (2021).
- Farran, B., AlWotayan, R. & Alkandari, H. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: A retrospective cohort study of health data from Kuwait [J]. *Front. Endocrinol.* **10**, 1 (2019).
- Hasan, M. K. *et al.* Diabetes prediction using ensembling of different machine learning classifiers[J]. *IEEE Access* **8**, 76516–76531 (2020).

19. Singh, A. *et al.* eDiaPredict: An Ensemble-based framework for diabetes prediction[J]. *ACM Trans. Multimed. Comput. Commun. Appl.* **17**(2s), 1–26 (2021).
20. Singh, N., Singh, P. & Bhagat, D. A rule extraction approach from support vector machines for diagnosing hypertension among diabetics [J]. *Expert Syst. Appl.* **130**, 188–205 (2019).
21. Thairaynayaki, K. Classification of diabetes using deep learning and svm techniques[J]. *Int. J. Curr. Res. Rev.* **13**(01), 146 (2021).
22. Jaiswal, V., Negi, A. & Pal, T. A review on current advances in machine learning based diabetes prediction[J]. *Prim. Care Diabetes* **15**(3), 435–443 (2021).
23. Almansour, N. A., Syed, H. F. & Khayat, N. R. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study [J]. *Comput. Biol. Med.* **109**, 101–111 (2019).
24. Patil, B. M., Joshi, R. C. & Toshniwal, D. Hybrid prediction model for Type-2 diabetic patients [J]. *Expert Syst. Appl.* **37**(12), 8102–8108 (2010).
25. Shen, L., Chen, H. & Yu, Z. Evolving support vector machines using fruit fly optimization for medical data classification [J]. *Knowl.-Based Syst.* **96**, 61–75 (2016).
26. Santhanam, T. & Padmavathi, M. S. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis [C]. In *proceedings of Graph algorithms, High performance implementations and its applications, India* **47**, 76–83 (2015).
27. Uzer, M. S., Yilmaz, N., & Inan, O. Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification [J]. *Sci. World J.* (2013).
28. Choubey, D. K., & Paul, S. GA\_SVM: A classification system for diagnosis of diabetes [M]. *Handbook of research on soft computing and nature-inspired algorithms*, 2017, 359–397.
29. Barakar, N. & Bradley, A. P. Rule extraction from support vector machines: A review [J]. *Neurocomputing* **74**(1–3), 178–190 (2010).
30. Núñez, H., Angulo, C., & Català, A. Rule extraction from support vector machines [C]. In *proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2002, 107–112.D
31. Zhang, Y. *et al.* Rule extraction from trained support vector machines [J]. *Adv. Tech. Knowl. Discov. Data Min* **1**, 92–95 (2005).
32. Martens, D. *et al.* Comprehensive credit scoring models using rule extraction from support vector machines [J]. *Eur. J. Oper. Res.* **183**(3), 1466–1476 (2007).
33. Han, L. *et al.* Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes [J]. *IEEE J. Biomed. Health Inform.* **19**(2), 728–734 (2015).
34. Liu, C. *et al.* A predictive model for acute allograft rejection of liver transplantation [J]. *Expert Syst. Appl.* **94**, 228–236 (2018).
35. Khanam, J. J. & Foo, S. Y. A comparison of machine learning algorithms for diabetes prediction[J]. *ICT Express* **7**(4), 432–439 (2021).
36. Deshmukh, T., Fadewar, H. S., & Shukla, A. The detection of Prameha (diabetes) in Ayurvedic way with the help of fuzzy deep learning. In *International Conference on Intelligent Computing and Communication Technologies*, pp. 152–158 (Springer, Singapore, 2019).
37. Azad, C. *et al.* Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus[J]. *Multimed. Syst.* **1**, 1–19 (2021).
38. Wang, Y. *et al.* Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection [J]. *Appl. Soft Comput.* **77**, 188–204 (2019).
39. Hayashi, Y. & Yukita, S. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset[J]. *Inf. Med. Unlocked* **2**, 92–104 (2016).
40. Chakraborty, M., Biswas, S. K. & Purkayastha, B. Rule extraction from neural network using input data ranges recursively[J]. *N. Gener. Comput.* **37**(1), 67–96 (2019).
41. Bader-El-Den, M., Teitei, E. & Perry, T. Biased random forest for dealing with the class imbalance problem [J]. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(7), 2163–2172 (2019).
42. Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics [J]. *Bioinformatics* **23**(19), 2507–2517 (2007).
43. Jin, X., Xu, A., & Bie, R. Machine learning techniques and Chi-square feature selection for cancer classification using SAGE gene expression profiles [C]. In *proceedings of the 1st workshop on Data Mining for Biomedical Applications, Singapore*, 106–115.
44. Brown, G. A new perspective for information theoretic feature selection [C]. In *Proceedings of the twelfth international conference on artificial intelligence and statistics (AISTATS)*, 49–56.
45. Qi, Y. Random Forest for Bioinformatics [J]. *Ensemble Mach. Learn.* **1**, 307–323 (2012).
46. Menze, B. H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data [J]. *BMC Bioinf.* **10**(1), 1 (2009).
47. Cheruku, R., Edla, D. & Kuppili, V. An optimized and efficient radial basis neural network using cluster validity index for diabetes classification [J]. *Int. Arab J. Inf. Technol.* **16**(5), 816–826 (2019).
48. Zhang, X. T. *et al.* A parallel ensemble fuzzy classifier for diabetes diagnosis [J]. *J. Med. Imaging Health Inf.* **10**(3), 544–551 (2020).
49. Hu, Y. X. *et al.* Deep supervised learning with mixture of neural networks [J]. *Artif. Intell. Med.* **102**, 1 (2020).

## Acknowledgements

This work was supported in part by the 13th 5-year National Science and Technology Supporting Project, PR China under Grant 2018YFC2000302 and the Information Security Software Project (2020) of the Ministry of Industry and Information Technology, PR China under Grant CEIEC-2020-ZM02-0134.

## Author contributions

J.H.: conceptualization—ideas; application of statistical-mathematical; visualization; writing—original draft. S.L.: supervision; project administration; writing—review and editing. L.P.: project administration; funding acquisition. All authors reviewed the manuscript.

## Funding

This article was funded by the 13th 5-year National Science and Technology Supporting Project, PR China (2018YFC2000302), the Information Security Software Project (2020) of the Ministry of Industry and Information Technology, PR China (CEIEC-2020-ZM02-0134).

## Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to J.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022