Informatics in Medicine Unlocked 5 (2016) 26–38

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease

Yoichi Hayashi*, Kazuhiro Fukunaga

Department of Computer Science, Meiji University, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan

ARTICLE INFO

Keywords:
Rule extraction
Re-RX algorithm
Sampling selection technique
BUPA liver disorder dataset
Hepatitis dataset
Child-Pugh score

ABSTRACT

Although liver cancer is the second most common cause of death from cancer worldwide, because of the limited accuracy and interpretability of extracted classification rules, the diagnosis of liver disease remains difficult. In addition, hepatitis, which is inflammation of the liver, can progress to fibrosis, cirrhosis, or even liver cancer. Numerous methods for diagnosing liver disease have been applied, but most current diagnostic methods are black box models that cannot adequately reveal information hidden in the data. In the medical setting, extracted rules must be not only highly accurate, but also highly interpretable. The Recursive-Rule eXtraction (Re-RX) algorithm is a white box model that generates highly accurate and interpretable classification rules on the basis of both discrete and continuous attributes; however, it tends to generate more rules than other rule extraction algorithms. The objectives of this study were to use a new rule extraction algorithm, Continuous Re-RX combined with sampling selection techniques (Sampling-Continuous Re-RX), to achieve highly accurate and interpretable diagnostic rules for the BUPA and Hepatitis datasets and to quantify the associations between the presence and severity of ascites and serum biomarkers with the risk of developing hepatitis in consideration of Child-Pugh scores. The performance of Sampling-Continuous Re-RX was compared with existing techniques, and as a result, it was found to extract more accurate, concise, and interpretable rules for the BUPA and Hepatitis datasets compared with previous extraction algorithms. In addition, the rules extracted using the proposed method were close to the trade-off curve, which indicated that they were more accurate and interpretable, and therefore more suitable in the medical setting.

1. Introduction

Liver cancer is the second most common cause of death from cancer worldwide, accounting for 6% of global cancer incidence and 9% of mortality. In 2012, 746,000 deaths were directly attributable to liver cancer. It is the fifth most common type of cancer among men (554,000 new cases, 8% of all cases) and the ninth most common among women (228,000 cases, 3% of all cases) [1].

Although substantial progress has been made regarding the knowledge and management of liver disease over the past several decades, approximately 29 million patients in the EU are suspected to have a chronic liver condition. Unfortunately, the evaluation of liver disease in the EU is limited due to difficulties in accessing data from individual countries [2].

In order to grasp a clear understanding of the actual burden of liver disease, the prevalence of cirrhosis and primary liver cancer, which represent the end stage of liver pathology and are therefore indicative of the associated mortality, need to be accurately assessed; however, such details have rarely been reported. Available data show that about 0.1% of the EU population has cirrhosis, which corresponds to between 14 and 26 new cases per 100,000 inhabitants and 170,000 deaths per year [3].

One of the most serious outcomes of cirrhosis is hepatocellular carcinoma, which is the fifth most common cause of cancer in the EU and represents about 70–90% of primary liver cancer cases. According to the World Health Organization (WHO), liver cancer is estimated to be responsible for about 47,000 deaths per year in the EU [2].

The leading causes of cirrhosis and primary liver cancer in the EU are harmful alcohol consumption, viral hepatitis B and C, and metabolic syndromes related to overweight and obesity. The second major cause of both cirrhosis and liver cancer is chronic viral hepatitis B [2].

Hepatitis is inflammation of the liver. It is a condition that can be self-limiting and may progress to fibrosis, cirrhosis, or liver cancer. Although hepatitis viruses are the most common cause of hepatitis in

* Corresponding author.
E-mail addresses: hayashiy@cs.meiji.ac.jp (Y. Hayashi), kazukazu3536@gmail.com (K. Fukunaga).

http://dx.doi.org/10.1016/j.imu.2016.10.001
Received 2 August 2016; Received in revised form 6 October 2016; Accepted 6 October 2016
Available online 10 October 2016

the world, hepatitis can also be caused by other infections, toxic substances such as alcohol and drugs, and autoimmune diseases.

According to the WHO, viral hepatitis infection affects about 400 million people worldwide, which is more than 10 times the number of people infected with HIV. Nevertheless, over 90% of those infected with hepatitis C can be completely cured within 3–6 months. Worldwide, about 1.4 million people annually die from hepatitis, while 6–10 million are newly infected.

The liver is the largest glandular organ in the body and is absolutely crucial to life, as it performs a number of vital functions, including the synthesis of proteins, fats, and fatty acids, metabolism and the storage of carbohydrates, and bile production and excretion. The liver maintains blood volume and quality by filtering out potentially harmful biochemical products such as bilirubin, which forms during the breakdown of old blood cells, and ammonia, which forms during the breakdown of proteins. Both bilirubin and ammonia are produced constantly. The liver also filters out harmful substances from external sources, including drugs, alcohol, and environmental toxins. The failure of any of these detoxifying functions leads to poor health [3].

Liver disease can result from infection, injury, drug reactions, toxins, autoimmune processes, or genetic defects that cause a buildup of iron or copper. It can lead to inflammation, scarring, fibrosis, various obstructions, clotting abnormalities, and even liver failure [3].

For example, alcoholic liver disease (ALD) is a type of liver disease that is gaining increasing recognition around the world [4]. About 493,300 deaths worldwide were attributed to ALD in 2010, which accounted for 0.9% of all deaths that year [5]. Except for alcohol disorders and fetal alcohol syndrome, liver diseases have the highest alcohol-attributable fractions of any disease, and alcohol consumption contributes to about half of the disease burden of liver cirrhosis [5].

The diagnosis of liver disease can be formulated as a two-class classification problem. Although numerous methods for diagnosing liver disease have been successfully applied to the classification of different tissues, most current diagnostic methods [6–27,32] are black box models. A drawback of black box models is that they cannot adequately reveal information that may be hidden in the data.

For example, even in cases for which high-performance classifiers [6–8,10,28] allow the accurate assignment of instances to groups, black box models cannot explain the reasons underlying that assignment to physicians; therefore, algorithms that can provide insight into these underlying reasons are needed. Rules are one of the most popular symbolic representations of knowledge discovered from data, and are more comprehensible than other representations [33].

Rule extraction can provide detailed explanations underlying assignments, and is it therefore becoming increasingly popular; however, in the medical setting, extracted rules must be not only highly accurate, but also simple and easy to understand.

Recently developed by Setiono et al. as a rule extraction tool [34], the Recursive-Rule eXtraction (Re-RX) algorithm provides a hierarchical, recursive consideration of discrete variables prior to analysis of continuous data, and can generate classification rules from neural networks (NNs) that have been trained on the basis of both discrete and continuous attributes.

In contrast to black box models, the Re-RX algorithm [34] is a "white box" model that provides highly accurate classification. It is easy to both explain and interpret in accordance with the concise extracted rules associated with IF-THEN forms. Due to its ease of understanding, the Re-RX algorithm is typically preferred by both physicians and clinicians.

However, due to its recursive nature, the Re-RX algorithm tends to generate more rules than other rule extraction algorithms. Therefore, one of the major drawbacks of the Re-RX algorithm is that it typically generates expansive extraction rules for middle-sized or larger datasets.

It is important to consider both accuracy and interpretability for extracted classification rules. The number of correctly classified test samples typically determines the accuracy of each extracted classification rule, while the number of extracted rules and the average number of antecedents in the extracted rules determine their interpretability.

To achieve both highly accurate and concise extracted rules while maintaining the desirable framework of the Re-RX algorithm, we recently proposed supplementing the Re-RX algorithm with continuous attributes (Continuous Re-RX) [35]. In Continuous Re-RX, C4.5 [36] is employed to form a decision tree in a recursive manner, while multi-layer perceptrons (MLPs) are trained using backpropagation (BP), which allows pruning [37] and consequently generates more efficient MLPs for highly accurate rule extraction. As a result, Continuous Re-RX provides rules that are not only highly accurate, but also concise and interpretable; that is, Continuous Re-RX provides IF-THEN rules. This white box model is easier to understand than traditional black box models and is therefore preferable in the medical setting.

In this study, we proposed the use of a new rule extraction algorithm, Continuous Re-RX [35] combined with sampling selection techniques [38,39] (Sampling-Continuous Re-RX) for preprocessing. This combination is similar to Sampling Re-RX with J48graft [40]; however, based on the difficulty of extracting highly accurate rules, the use of Sampling-Continuous Re-RX algorithm allowed us to achieve high accuracy while only sacrificing slightly less conciseness because although Continuous Re-RX provides higher accuracy, it also extracts a larger number of rules [35].

The accuracy and interpretability of diagnostic rules extracted using Sampling-Continuous Re-RX were investigated based on a comparison with crisp rule extraction [41] and two previous fuzzy rule extraction techniques [42,43]. The BUPA Liver Disorders dataset from the repository of machine learning at the University of California Irvine (UCI) [44], which comprises 768 cases with two classes (disorder or non-disorder) and six continuous attributes, was used in this study, as was the Hepatitis dataset for the same glandular organ from the UCI machine learning repository [44], which comprises 155 cases with two classes (LIVE or DIE) and 19 attributes.

The performance of rule extraction algorithms for the BUPA dataset since 2006 and Hepatitis dataset since 1992 were reviewed and compared with that of the previous rule extraction techniques with Sampling-Continuous Re-RX.

In Sections 5.1.1 through 5.1.6, the concrete rule set for the BUPA dataset extracted by the proposed algorithm is compared with three kinds of previous rule extraction algorithms. In Sections 5.2.1 through 5.2.4, the concrete rule set for the Hepatitis dataset extracted by the proposed algorithm is compared with previous rule extraction algorithms.

In Section 6.1, the role of four kinds of biomarkers for the diagnosis of liver disorders is explained, and in Section 6.2, the liver enzyme and serum activity of alanine aminotransferase (ALT) is described [45]. The serum activity of gamma-glutamyl transpeptidase (transferase) (GGT) [46] is discussed in Section 6.3, GGT as an indicator of liver disease [47] in Section 6.3, and the interpretation of rules extracted by the proposed algorithm from the perspective of medical informatics in Section 6.4. In Section 6.5, the trade-offs between accuracy and the number of extracted rules is discussed using a trade-off curve for the BUPA dataset. In Section 6.6, the Child-Pugh score is described [48,49]. In Section 6.7, in consideration of the Child-Pugh score, an interpretation of the rules extracted from Hepatitis dataset is provided. In Section 6.8, the trade-off between accuracy and the number of extracted rules is discussed using a trade-off curve for Hepatitis dataset. Finally, in Section 7, a summary and conclusion are provided.

The first objective of this study was therefore to quantify the nature and magnitude of the associations between GGT, ALT, aspartate aminotransferase (AST) and ALP levels with the risk of developing liver disease using the rule extraction approach.

The second objective was to quantify the nature and magnitude of associations between the presence and severity of ascites and the levels

of several serum biomarkers with the risk of developing hepatitis using the rule extraction approach and the Child-Pugh score.

2. Related works

The BUPA Liver Disorders dataset was created by Richard S. Forsyth at BUPA Medical Research and Development Ltd. during the 1980s as part of a larger health-screening database. In 1990, the dataset was donated on his behalf to the UCI machine learning repository [44]. Since then, it has been commonly used as a benchmark for classification algorithms. The Hepatitis disease dataset was created at the Jozef Stefan Institute in Slovenia. The dataset was also donated on his behalf to the UCI machine learning repository.

Numerous methods for diagnosing the BUPA and Hepatitis datasets have been successfully applied to the classification of different tissues. These methods include the following: clustering based attribute weighting [6]; extreme learning machines [12]; support vector machines (SVMs) [9,13,14,30,31,43]; neural networks (NNs) [28,37]; fuzzy classification [42]; fuzzy decision tree [11]; fuzzy rule extraction from SVMs [43]; CART [7,8]; support vector recognition [10]; binary classification [19]; artificial immune systems [15,16,29,32]; swarm optimization [18,22,26,50]; neuro-fuzzy models [21]; fuzzy classifiers [20]; kernel nearest-neighbor [23]; feature extraction [24]; principal component analysis [25,31]; axiomatic fuzzy sets [51]; information granulation [52]; electromagnetism-like mechanisms [17]; support feature machines [27] and rough sets [28].

A brief description of four rule extraction algorithms [37–39,53] used for comparisons is provided in Section 5.

Hsieh et al. [41] proposed a particle swarm optimization (PSO)-based Fuzzy Hyper-Rectangular Composite Neural Network (PFHRCNN), which applies PSO to trim the rules generated by a trained HRCNN without downgrading (and even possibly improving) the recognition performance.

The classification methodology proposed by Gadaras and Mikhailov [42] identifies fuzzy boundaries of classes by processing a set of labeled data. Fuzzy rules are obtained by exploring the characteristics of the identified boundaries and automatically producing membership functions for each class. When new patterns require classification, their numerical attributes are tested against generated knowledge to match a patient's symptoms with an antecedent.

In order to improve the interpretability of generated rules, Chaves et al. [43] proposed FREx_SVM, a new method for fuzzy rule extraction from trained SVMs for multi-class problems that includes a wrapper feature selection algorithm.

In 2003, Tan et al. [53] proposed pioneering research to extract diagnostic rules from the Hepatitis dataset using a two-phase hybrid evolutionary classifier. In the first phase, to confine the search space by evolving a pool of good candidate rules a hybrid evolutionary algorithm is used; for example, genetic programming is applied in order to evolve nominal attributes for free structured rules, while a genetic algorithm is used to optimize the numeric attributes for concise classification rules without the need of discretization. In the second phase, these rules are used to optimize the order and number of rules to create accurate and comprehensible rule sets.

3. Methods

3.1. Recursive-Rule eXtraction (Re-RX) algorithm

Although the Re-RX algorithm can easily handle multi-group problems, it was originally developed to consider only two-group classification problems [34]. The outline of the Re-RX algorithm is as follows:

Algorithm. Re-RX (*S, D, C*)

Input: A set of data samples S having discrete attributes D and continuous attributes C.

Output: A set of classification rules.

1. Train and prune [37] an NN by using the dataset S and all of its D and C attributes.
2. LeD' and C' be the sets of discrete and continuous attributes, respectively, still present in the network, and let S' be the set of data samples correctly classified by the pruned network.
3. If $D'=\phi$, then generate a hyperplane to split the samples in S' according to the values of the continuous attributes C', and then stop.Otherwise, use only the discrete attributes D' to generate the set of classification rules R for dataset S′.
4. For each rule, R_i is generated:

$$\text{If support } (R_i) > \delta_1 \text{ and error } (R_i) > \delta_2, \text{ then} \quad (1)$$

- Let S_i be the set of data samples that satisfy the condition of rule R_i and D_i be the set of discrete attributes that do not appear in rule condition R_i.
- If $D_i=\phi$, then generate a hyperplane to split the samples in S_i according to the values of their continuous attributes C_i, and then stop.

$$\text{Otherwise, call Re−RX } (S_i, D_i, C_i). \quad (2)$$

Any NN training and pruning method can be used in Step 1 of the Re-RX algorithm, as it does not make any assumptions regarding the NN architecture; however, we have restricted ourselves to the use of backpropagation NNs with only one hidden layer because such networks have been shown to retain the universal approximation property [54].

A crucial component of any NN rule extraction algorithm is an effective NN pruning algorithm. Pruning the inputs that are not needed to solve the problem allows the extracted rule set to be more concise, and a pruned network also helps to filter noise that might be present in the data, such as that from outlying or incorrectly labeled data samples. Therefore, from Step 2 onward, the algorithm only processes training data samples that have been correctly classified by the pruned network. Previously, we developed an NN pruning algorithm that incorporates a penalty function during training and adds a positive penalty value to the sum-of-squared error function for each connection with nonzero weight [33]. Consequently, many of the connections have weights very close to zero when network training is complete, and those with very small values can typically be pruned without adversely affecting the accuracy of the network.

If all discrete attributes are pruned from the network, the algorithm generates a hyperplane in Step 3

$$\sum_{Ci \in C'} wiCi = w0 \quad (3)$$

that separates both groups of samples. Statistical and machine learning methods such as logit regression or SVMs can then be used to obtain the constant and the rest of the coefficients of the hyperplane. We employ an NN with one hidden unit in our implementation.

The support of a rule, which is the percentage of samples covered by that rule, and each rule's corresponding error rate are checked in Step 4. If the support meets the minimum threshold δ_1 and the error rate exceeds the threshold δ_2, then the subspace of the rule is further subdivided either by calling Re-RX recursively when no discrete attributes remain present in the conditions of the rule, or by generating a separating hyperplane involving only the continuous attributes. Because the Re-RX algorithm handles discrete and continuous attributes separately, it generates a set of classification rules that are more comprehensible than those with both types of attributes in their conditions.

To enable a better understanding of its underlying mechanisms, a brief overview of the Re-RX algorithm and the concept behind its design is shown in Fig. 1. C4.5 [36] was used to generate decision trees in the Re-RX algorithm. The subdivision of the Re-RX algorithm is a unique function that is inherent in its nature. Each successive subdivision allows the use of other previously unused attributes; this increases the number of extracted rules as well as their accuracy.

It should be noted that the accuracy, comprehensibility, and conciseness of extracted rules have important trade-offs. Before subdivision, extracted rules are more comprehensible and concise, yet less accurate. Conversely, after subdivision, extracted rules are less concise, yet more accurate.

3.2. Re-RX algorithm with continuous attributes (Continuous Re-RX)

Although a primary aim of the Re-RX algorithm is the strict separation of discrete and continuous attributes in the antecedent of each extracted rule, this design often results in reduced accuracy. Whereas the Re-RX algorithm prunes continuous attributes (C′) before the C4.5 decision tree is generated (Fig. 2), Continuous Re-RX uses both discrete (D′) and continuous attributes (C′) to generate the decision tree [35], which results in increased complexity. This may seem counterintuitive to the algorithm's design, but the use of both types of attributes also results in increased accuracy. An outline of Continuous Re-RX is as follows:

Continuous Re-RX (S', D', C')
Input: A set of data samples (S') having both discrete (D') and continuous (C') attributes.
Output: A set of classification rules.

1. Train and prune [37] an NN using the dataset S and all of its D and C attributes.
2. Let D' and C' be the sets of discrete and continuous attributes, respectively, still present in the network, and let S' be the set of data samples correctly classified by the pruned network.
3. Generate decision tree by using both discrete (D') and continuous (C') attributes [35].
4. For each rule, R_i is generated:

$$\text{If support } (R_i) > \delta_1 \text{ and error } (R_i) > \delta_2, \text{ then} \quad (4)$$

- Let S_i be the set of data samples that satisfies the condition of rule R_i, let D_i be the set of discrete attributes, and let C_i be the set of continuous attributes that does not appear in rule condition R_i.

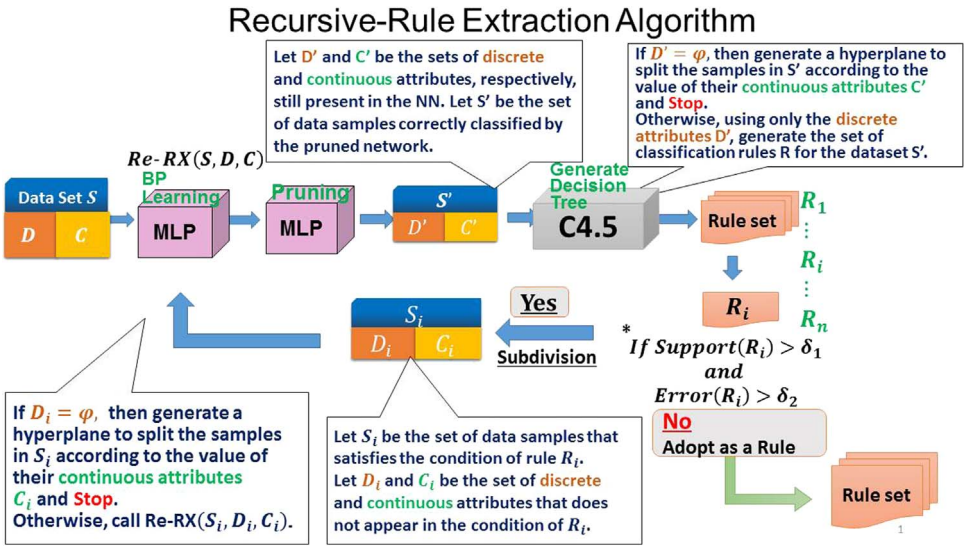

Fig. 1. Schematic overview of the Recursive-Rule eXtraction (Re-RX) algorithm.

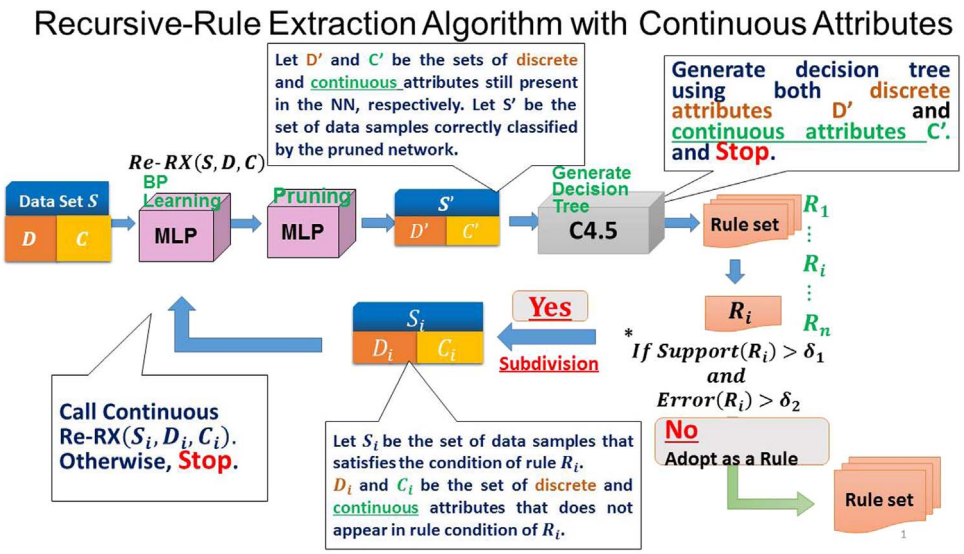

Fig. 2. Schematic overview of the Re-RX algorithm with continuous attributes.

Call Continuous Re−RX (S_i, D_i, C_i). (5)

Otherwise, Stop.

As shown in Fig. 2, to avoid such difficulties in Continuous Re-RX, we carefully set the value of subdivision rate and the values of δ_1 and δ_2 in Step 4.

3.3. Sampling selection technique

Setiono [38,39] proposed a supervised learning scheme that aimed to increase model accuracy by selecting the most appropriate training data samples. In that scheme, models for classification problems such as NNs are trained using a historical dataset. In the case of classification problems such as credit scoring, the credit risk of each sample is labeled as either good or bad.

However, some of these class labels may be incorrectly assigned, resulting in the presence of irregular data samples. Although these samples may have similar attributes, as is commonly the case for most samples in one class, they actually belong to a different class. This is problematic because the presence of irregular and/or mislabeled data samples in a training dataset is likely to adversely affect the performance of the NN.

In the sampling selection technique proposed by Setiono et al. [38,39], NNs are trained to identify potentially irregular and/or mislabeled data samples. Data samples that are consistently misclassified by a majority of NNs are then removed before a model is constructed to distinguish between good and bad credit risk.

The sampling selection technique in the present paper can be summarized as follows: 1) Ensemble creation: train an ensemble of M feedforward NNs using the available training data samples; 2) Sample selection: select training data samples based on the predictions of the NN ensemble; 3) Model generation: use the selected samples to train an NN; and 4) Rule extraction: apply the Continuous Re-RX algorithm [35] to obtain highly accurate and interpretable classification rules capable of distinguishing between disorders and non-disorders.

The selection of samples in Step 2 is a core component of the sampling selection technique. First, we employed an NN ensemble to identify outliers in the training dataset. An effective method for improving the predictive accuracy of numerous learning methods is to remove outliers and noise prior to learning. If a data sample is incorrectly classified by a proportion of NNs exceeding the threshold ρ and thereby identified as an outlier, it is discarded; otherwise, it is retained in the training dataset.

3.4. Re-RX algorithm with continuous attributes combined with a sampling selection technique (Sampling-Continuous Re-RX)

Here we propose a new highly accurate and interpretable rule extraction algorithm using Continuous Re-RX with combined with sampling selection techniques (Sampling-Continuous Re-RX) for preprocessing.

The objective of the present study was to achieve highly accurate and interpretable classification rules for the BUPA and Hepatitis datasets. However, these are medical datasets, so the focus was on decreasing the number of extracted rules and the average number of antecedents. To extract accurate rules, Sampling-Continuous Re-RX, which is better suited for achieving highly accurate and interpretable medical rules, was applied.

The BUPA and Hepatitis datasets were preprocessed using the sample selection technique [38,39] to extract a fewer number of rules and a lower average number of antecedents. We then employed Continuous Re-RX to extract a set of highly accurate and interpretable diagnostic rules for the BUPA and Hepatitis datasets. As shown in a schematic overview of Sampling-Continuous Re-RX in Fig. 3, a supplementary cross-validation (CV) loop is carried out with the sampling selection by an NN ensemble.

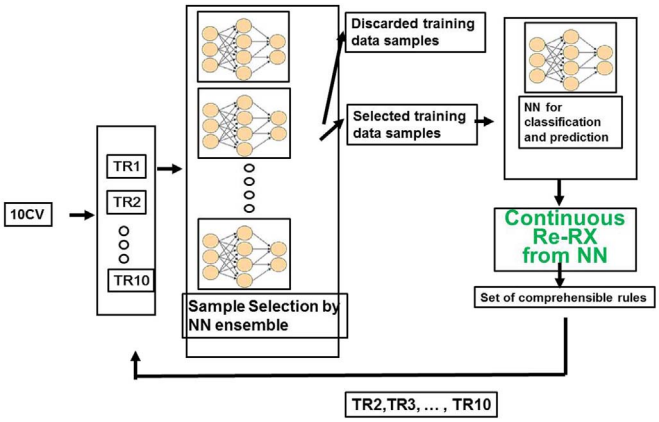

Fig. 3. Schematic overview of the sampling-continuous Re-RX algorithm.

The most important aim of Sampling-Continuous Re-RX is to improve the accuracy and interpretability of extracted rules for physicians, because the competition for achieving only better classification accuracy for the BUPA and Hepatitis datasets has appeared to plateau [6–8,28], and unless diagnostic accuracy can be considerably, i.e., closed 100%, improved, limited contributions will be made to medical informatics.

3.5. Experimental setup for the BUPA liver disorders dataset

The BUPA Liver Disorders dataset comprises 345 samples, each taken from an unmarried male, consisting of six attributes and two classes as follows: 200 of these samples are belong to one class (disorder), and the remaining 145 belong to the other (non-disorder). The first five attributes of the collected data samples are the results of blood tests, while the last attribute is daily alcohol consumption [55].

1. MCV: mean corpuscular volume (fL)
2. ALP: alkaline phosphatase (IU/L)
3. ALT: alanine aminotransferase (IU/L)
4. AST: aspartate aminotransferase (IU/L)
5. GGT: gamma-glutamyl transpeptidase (or transferase) (IU/L)
6. DRNO: number of half-pint equivalents of alcoholic beverages drunk per day.

Notes on the amount of some kinds of alcohols:
Half-pint of beer (284 mg)=12.5 g Alcohol
Wine (120 mg)=12.0 g Alcohol
Japanese sake (100 mg)=12.0 g Alcohol
Whisky (40 mg)=12.8 g Alcohol

3.6. Experimental setup for the Hepatitis dataset

The Hepatitis dataset, which consists of 155 instances, each consisting of 19 attributes, namely age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, ALP, AST, albumin, protime and histology, is summarized in Table 1. The problem with this database is that it includes both nominal and numeric attributes. The Hepatitis dataset is a complex and noisy dataset because it contains a large amount of missing data. Class is distributed with 32 (20.65%) DIE samples and 123 (79.35%) LIVE samples. The classification task is to predict whether a patient with hepatitis will live or die [53].

Table 1
Summary of the Hepatitis dataset.

Attribute	Possible values
Age, years	Integer 1–80
Sex	Male, female
Steroid	No, yes
Antivirals	No, yes
Fatigue	No, yes
Malaise	No, yes
Anorexia	No, yes
Liver big	No, yes
Liver firm	No, yes
Spleen palpable	No, yes
Spiders	No, yes
Ascites	No, yes
Varices	No, yes
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
ALP	33, 80, 120, 160, 200, 250
AST	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	No, yes
Class	Die (20.65%), Live (79.35%)

4. Results

4.1. Performance

To guarantee the validity of the results, we used k-fold CV [56] to evaluate the classification rule accuracy of test datasets. The k-fold CV method is widely applied by researchers to minimize the bias associated with random sampling.

The BUPA dataset was trained using Sampling-Continuous Re-RX, and 5 runs of 2-fold CV (repeated-randomized-hold-out-approach) accuracies for the training dataset (TR ACC), 5 runs of 2-fold CV accuracies for the test dataset (TS ACC), the number of extracted rules (# rules), the average number of antecedents (Ave. # ante.), and the area under the receiver operating characteristics curve (AUC) [57] were obtained (Table 2). In this paper, the AUC was used as an appropriate evaluator because it does not include class distribution or misclassification costs [57].

Numerous types of rules have been suggested in the literature from the perspective of the expressive power of extracted rules, including propositional rules, which take the form of IF-THEN expressions and clauses defined using propositional logic, and M-of-N rules. Breaking from traditional logic, fuzzy rules allow partial truths instead of Boolean TRUE/FALSE outcomes.

Even if all types of rules are considered, the consensus is that no matter how they are defined, an ideal measure has yet to be developed; therefore, “what is a concise and/or interpretable rule?” remains a difficult question to answer.

To answer this question, we attempted to develop a “rough indicator” of conciseness by comparing the average number of antecedents from extracted rules generated using a variety of techniques [35].

We achieved an average accuracy of 73.48% after 5 runs of 2-fold CV for the BUPA dataset, as shown in Table 2.

Although the accuracy and the number of rules extracted were slightly varied, the reliability and robustness of the accuracies and the number of rules extracted obtained by the proposed method were confirmed by varying the number of hidden units in intermediate layer, as shown in Table 3. The parameter settings are shown in Table 4. According to our experience with numerical experiments, the number of hidden units can be expected to mostly affect the accuracies and the number of rules extracted.

Regarding the complexity of Sampling-Continuous Re-RX, it took about 4.3 s to train the BUPA dataset using a standard workstation computer (3.1 GHz Intel Xeon E5-2687W, 3.5 GHz Turbo, 25 MB Cache; 64 GB RAM; 512 GB DDR3 System memory) and about 21.3 s for 5 runs of 2-fold CV. The testing time was negligible.

Based on the comparisons shown in Table 2, Sampling-Continuous Re-RX extracted more accurate, concise, and interpretable rules for the BUPA dataset. That is, Sampling-Continuous Re-RX achieved substantially better accuracy (72.44–73.48% vs. 64.33–67.33%) and considerably fewer rules and antecedents compared with the original Re-RX algorithm.

4.2. Results for the Hepatitis dataset

We achieved an average accuracy of 83.24% after 5 runs of 2-fold CV for the Hepatitis dataset, as shown in Table 5. Regarding the complexity of Sampling-Continuous Re-RX, it took about 2.43 s to

Table 3
Effect of the accuracies and number of rules extracted obtained by the proposed method by different parameter settings for the BUPA dataset.

Parameter	TR ACC (%)	TS ACC (%)	# Rules	Ave. # ante.	AUC	TR ACC (SD)	TS ACC (SD)
Parameter setting 1	75.19	**73.48**	**8.50**	**2.24**	0.69	1.66	1.74
Parameter setting 2	76.84	74.36	9.5	2.48	0.7	2.01	2.17
Parameter setting 3	75.38	74.22	9	2.39	0.7	2.15	2.22

Table 4
Summary of parameter settings for NNs for the BUPA dataset.

Parameter	Learning rate	Momentum factor	Number of epochs	Number of hidden units
Parameter setting 1	0.1	0.1	1000	1
Parameter setting 2	0.1	0.1	1000	2
Parameter setting 3	0.1	0.1	1000	3

Table 2
Average accuracies after CV for the BUPA dataset.

BUPA dataset	TR ACC (%)	TS ACC (%)	# Rules	Ave. # ante.	AUC	TR ACC (SD)	TS ACC (SD)
Sampling-Continuous Re-RX [5×2CV]	75.19	**73.48**	**8.50**	**2.24**	0.69	1.66	1.74
Sampling-Continuous Re-RX [10×10CV]	75.12	**72.44**	**11.00**	**2.54**	0.69	1.56	1.63
Re-RX with C4.5 [5×2CV]	72.44	64.33	10.31	2.78	0.60	2.28	1.43
Re-RX with C4.5 [10×10CV]	70.67	67.33	12.33	3.32	0.62	3.78	3.84

CV: cross validation; Re-RX: Recursive-Rule eXtraction; Continuous Re-RX: Re-RX algorithm with Continuous Attributes; Sampling-Continuous Re-RX: Re-RX algorithm with Continuous Attributes combined with Sampling Selection technique; TR: training dataset; TS: testing dataset; ACC: accuracy; Ave. # ante.: average number of antecedents; AUC: area under the receiver operating characteristic curve; SD: standard deviation. 10×10CV: 10 runs of 10-fold cross validation; 5×2CV: 5 runs of 2-fold cross validation (Repeated-randomized-hold-out approach).

Table 5
Average accuracies for the Hepatitis dataset after CV.

Hepatitis dataset	TR ACC (%)	TS ACC (%)	# Rules	Ave. # ante.	AUC	TR ACC (SD)	TS ACC (SD)
Sampling-Continuous Re-RX [5×2CV]	89.04	**83.24**	**3.50**	**1.90**	0.67	2.12	1.98
Sampling-Continuous Re-RX [10×10CV]	89.24	**82.08**	**5.50**	**2.22**	0.65	0.57	1.34
Re-RX with C4.5 [5×2CV]	88.42	78.73	5.50	2.63	0.63	2.55	3.01
Re-RX with C4.5 [10×10CV]	87.02	79.29	5.30	2.59	0.66	1.11	4.32

CV: cross validation; Re-RX: Recursive-Rule eXtraction; Continuous Re-RX: Re-RX algorithm with Continuous Attributes; Sampling-Continuous Re-RX: Re-RX algorithm with Continuous Attributes combined with Sampling Selection technique; TR: training dataset; TS: testing dataset; ACC: accuracy; Ave. # ante.: average number of antecedents; AUC: area under the receiver operating characteristic curve; SD: standard deviation. 10×10CV: 10 runs of 10-fold cross validation; 5×2CV: 5 runs of 2-fold cross validation (Repeated-randomized-hold-out approach).

Table 6
Performance of previous rule extraction algorithms for the BUPA dataset.

Rule extraction method [validation method] [Ref.]	TR ACC (%)	TS ACC (%)	# Rules	Rule set	Total ante. #	Ave.# ante	Year
Knowledge Acquisition via Information Granulation (KAIG) [N/A] [52]	78.2	70.0	5 (FR)	No	–	–	2006
Interpretable-Fuzzy-Rule Based-Classification [Averaged over 10 runs] [42]	–	89.9	8 (FR)	**Yes**	64 FS	6.0	2009
Axiomatic Fuzzy Sets (AFS) Approach [5CV] [51]	–	69.28	17.6 (FR)	No	–	–	2013
Fuzzy Rule Extraction from Trained SVMs [Max. ACC] [43]	–	61.19–76.13	2 (FR)	**Yes**	8	4.0	2013
Hyper-Rectangular Composite NNs [Averaged over 10 runs] [41]	90	81	24 (FR)	**Yes**	24	4.0	2014
Wind-Driven Swarm Optimization (WSO) [5*2CV] [50]	–	57.27–73.95	6–32	No	–	–	2015
Sampling-Continuous Re-RX [5×2CV] Sampling-Continuous Re-RX [10×10CV]	75.19	**73.48**	**8.5**	**Yes**	20	**2.24**	**Present study**
	75.12	**72.44**	**11.00**	**Yes**	28	**2.54**	**Present study**

Re-RX: Recursive-Rule eXtraction; TR: training dataset; TS: testing dataset; ACC: accuracy; Ave. # ante.: average number of antecedents; Total # ante.: total number of antecedents; 10CV: 10-fold cross validation; 10×10CV: 10 runs of 10-fold cross validation; 5×2CV: 5 runs of 2-fold cross validation (Repeated-randomized-hold-out approach); FR: fuzzy rule.

train the Hepatitis dataset using a standard workstation computer (3.1 GHz Intel Xeon E5-2687W, 3.5 GHz Turbo, 25 MB Cache; 64 GB RAM; 512 GB DDR3 System memory) and about 12.13 s for 5 runs of 2-fold CV. The testing time was negligible.

5. Comparisons

5.1. Rule extraction comparison for the BUPA dataset

Next, we reviewed the rule extraction algorithms used for the BUPA dataset since 2006 and tabulated their performances (Table 6). The concrete rules extracted for the BUPA dataset by Sampling-Continuous

Table 7
A review of previous classifier systems for the BUPA liver disorders dataset.

Author (Year) [Ref.]	**Method**	**Classification accuracy (%)**
Polat K (2012) [6]	SCBAW (Subtracting Clustering based Attribute Weighting) [10CV]	99.41
Seera M and Lim CP (2014) [7]	Fuzzy-Mini-Max (FMM)-CART-RF 30×2CV, 5CV, 10CV	95.01
Seera M et al. (2015) [8]	Fuzzy-ARTMAP-CART [30×10CV]	94.41
Çomak E et al. (2007) [9]	Fuzzy Weighting Pre-Processing+LS-SVM	94.29
Zangooei MH et al. (2014) [10]	Support Vector Recognition Using NSGA-II [10CV]	91.24
Fan CY et al. (2011) [11]	Case-Based-Fuzzy-Decision-Tree [100 runs Best ACC]	90.4
Mohapatra P et al. (2015) [12]	Improved-Cuckoo-Search-Based-Extreme-Learning-Machine	88.36
Li DC et al. (2010) [13]	Balance+Extension+SVM [10CV]	86.36
Peker M (2016) [14]	K-Medoids Clustering-Based Attribute Weighting+SVM [10×10CV]	86.25
Özşen S and Güneş S (2009) [15]	GA-AWAIS [10×10CV]	85.21
Polat K et al. (2007) [16]	Fuzzy-AIRS [10CV]	83.38
Wang KJ (2015) [17]	Improved Electromagnetism-Like Mechanism	77.61
Chang PC et al. (2012) [18]	Particle Swarm Optimization [Averaged over 500 runs]	76.8
Kraipeerapun P and Fung CC (2009) [19]	Ensemble NNs and Interval Neutrosophic Sets	74.64
Luukka P (2011) [20]	Fuzzy Beans Classifier [30×10CV]	73.9
Goncalves LB et al. (2006) [21]	Neuro-Fuzzy Model	73.26
Beheshti Z et al. (2014) [22]	Particle Swarm Optimization [Averaged over 10 runs]	72.32
Yu K et al. (2002) [23]	Kernel Nearest-Neighbor	71
Li DC et al. (2011) [24]	PCA+SVM	70.85
Luukka P (2009) [25]	Support Vector Machine [10×5CV]	70.25
Liu R et al. (2014) [26]	Fuzzy Robust PCA+Similarity Classifier	65.81
Fan YJ et al. (2010) [27]	Particle Swarm Optimization [Averaged over 20 runs]	63.28
Sampling-Continuous Re-RX [5×2CV]		
Present study		**73.48**
Sampling-Continuous Re-RX [10×10CV]		
Present study		**72.44**

BUPA: BUPA liver disorder; Re-RX: Recursive-Rule eXtraction; MLP: Multilayer Perceptron; 10CV: 10-fold cross validation; 10×10CV: 10 runs of 10-fold cross validation; 5×2CV: 5 runs of 2-fold cross validation; PCA: Principal Component Analysis; FNN: Fuzzy Neural Network.

Re-RX are shown in Section 5.1. The three kinds of rule sets for the BUPA dataset reported in previous studies are described in Sections 5.1.1 through 5.1.4. In Section 5.1.5, we compare Sampling-Continuous Re-RX with previous rule extraction algorithms. In Section 5.1.6, we review previous classifier systems for the BUPA dataset (Table 7).

5.1.1. Rules extracted for the BUPA dataset using Sampling-Continuous Re-RX

R1: GGT≤22 AND ALT≥21 AND AST≤27 THEN Non-disorder
R2: GGT≤22 AND ALT≥21 AND AST > 27 THEN Disorder
R3: GGT∈(22, 31] AND ALT≤21 AND AST≤11 THEN Non-disorder
R4: GGT∈(22, 31] AND ALT≤21 AND AST > 11 THEN Disorder
R5: GGT∈(22, 31] AND ALT > 34 THEN Disorder
R6: GGT∈(22, 31] AND ALT∈(21, 34] THEN Non-disorder
R7: GGT > 31 AND DRNO≤9 THEN Disorder
R8: GGT > 31 AND AST≤28 AND DRNO∈(9, 10] THEN Non-disorder
R9: GGT > 31 AND AST > 28 AND DRNO∈(9, 10] THEN Disorder

5.1.2. Rules extracted for the BUPA dataset using a PSO-based fuzzy hyper-rectangular composite NN [41]

Non-disorder			Disorder	
Rule 1		Rule 2	Rule 1	Rule 2
MCV	[82.088, 89.28]	[85.277, 89.697]	[85.277, 89.697]	[82.391, 92.501]
ALP	[80.46, 103.83]	[54.33, 132.994]	[54.33, 132.994]	[29.774, 65.203]
ALT	[4.840, 162.670]	[24.957, 87.737]	[24.957, 87.737]	[2.771, 162.015]
AST	[4.559, 30.915]	[5.191, 43.214]	[5.191, 43.214]	[2.251, 21.598]
GGT	[15.063, 30.574]	[26.863, 41.596]	[26.863, 41.596]	[22.290, 184.61]
DRNO	[1.697, 2.874]	[0.428, 1.589]	[0.428, 1.589]	[1.766, 13.116]
Confidence	0.999	0.809	0.809	0.810
Scale	0.360	0.008	0.008	0.858

5.1.3. Rules extracted for the BUPA dataset using evolving fuzzy rule-based classification [42]

Rule no.	MCV	ALP	ALT	AST	GGT	DRNO	Class
R1	High	Low	Low	Low	Low	High	Non-disorder
R2	Low	High	High	High	High	Low	Disorder
R3.1	Med—high	Med—high	Low—med	Low—med	Low—med	Low—med	Non-disorder
R3.2	Low—med	Low—med	Med—high	Med—high	Med—high	Med—high	Disorder
R3.3.1	Med—med High	Med—med high	Med—med high	Med—med high	Med—med high	Med—med high	Non-disorder
R3.3.2	Low—med	Low—med	Low—med	Low—low	Low—low	Low—med	Disorder
	med	med	med	med	med	med	
R3.3.3.1	Med— Med— Med high	Med— med— med high	Med— med— med low	Med— med— med low	Med— med— med high	Med— Med— med low	Non-Disorder
R3.3.3.2	Med— med— med— low	Med— med— med— low	Med— med— med— high	Med— med— med— high	Med— med— med— low	Med— med— med— high	Disorder

5.1.4. Rules extracted for the BUPA dataset using fuzzy rule extraction from trained SVMs [43]

R1: IF ALT is *low* and AST is *low* and GGT is *low* and DRNO is *medium*, THEN Class 1—Accuracy: 0.6119 for 5 fuzzy sets.

R2: IF ALT is *medium* and AST is *low* and GGT is *low* and DRNO is *low*, THEN Class 2—Accuracy: 0.7613 for 3 fuzzy sets.

5.1.5. Comparison of rules extracted in the present study with those from three previous algorithms

In Section 5.1.2, among 24 rules extracted, four rules were extracted by the PSO-based fuzzy hyper-rectangular composite network [41]. However, the actual number of extracted rules and antecedents were much higher compared with Sampling Continuous Re-RX.

Furthermore, extracted closed intervals for non-disorder and disorder were considerably overlapped, which made classification more difficult. For example, the first rule, i.e., the closed interval of non-disorder for ALP and GGT, was [80.46, 103.83] and [15.063, 30.574], respectively; the second rule of non-disorder for ALP and GGT was [54.33, 132.994] and [26.863, 41.596], respectively.

On the other hand, the first rule, i.e., the closed interval of disorder for ALP and GGT, was [54.33, 132.994] and [26.863, 41.596], respectively; the second rule of non-disorder for ALP and GGT was [29.774, 65.203] and [22.290, 184.61], respectively.

In Section 5.1.3, regarding accuracy, six rules extracted by fuzzy classifiers [42] achieved an accuracy of 89.9% based on an average of 10 runs. Therefore, the accuracy is expected to decrease considerably based on the 10 CV accuracy measure.

In addition, not all membership functions for MCV, ALP, ALT, AST, GGT, and DRNO in the antecedent were accurately depicted. Moreover, the rules extracted in Section 5.1.3 appeared to be too complicated. The default number of attributes in antecedents of rules extracted by fuzzy classifiers is six, which is identical to the number of attributes in the BUPA dataset; therefore, the rules extracted in Section 5.1.3 may be intuitively interpretable, but also overly subjective.

At a glance, two fuzzy rules extracted in Section 5.1.4 look quite simple. However, the procedure to derive these two rules is not straightforward. FREx_SVM combined with the feature selection method [43] reduces the number of rules. The best rules obtained are two fuzzy rules with six features. Although replacing all attributes with their real names and assuming that the set of labels {*very low*, *low*, *medium*, *high*, *very high*} are associated with each of the five fuzzy sets. However, membership functions for ALT, AST, GGT, and DRNO in the antecedent were not accurately depicted.

In the present form, the average number of antecedents was considerably larger (4.0), and the accuracies were less than that of the proposed algorithm. Therefore, these extracted rules are less interpretable than the rules obtained using the proposed algorithm.

Consequently, we believe that the present rules extracted using the Sampling-Continuous Re-RX algorithm achieved excellent performance (73.48% with an average of 2.24 antecedents).

5.1.6. Comparison of the classification accuracy in the present study with another classifier system for the BUPA dataset

In this section, the classification accuracy obtained using Sampling Continuous Re-RX is compared with that obtained using previous classifier systems [6–27], as tabulated in Table 7. We reviewed the classifier systems reported since 2002 and tabulated their performances. Table 7 shows a comparison of studies that carried out k-fold CV to measure classification accuracy. Generally, rule extraction algorithms attempt to achieve both highly accurate and highly concise extracted rules with a well-balanced trade-off. Strictly in terms of classification accuracy, Sampling Continuous Re-RX may not be superior to recent high performance classifiers.

5.2. Rule extraction comparison for the Hepatitis dataset

5.2.1. Performance of previous rule extraction algorithms for the Hepatitis dataset

See Table 8.

5.2.2. Rules extracted for the Hepatitis dataset using Sampling-Continuous Re-RX

R1: Ascites=No THEN LIVE
R2: Ascites=Yes AND Albumin≤3.5 THEN DIE
R3: Ascites=Yes AND Albumin > 3.5 THEN LIVE

5.2.3. Rules extracted for the Hepatitis dataset using EvoC [53]

R1: IF Fatigue=Yes AND Age≥30.0 AND ALP≤280.0
AND Albumin 4.3 AND Protime≤46.0
THEN Class=DIE

R2: IF Anorexia=No AND Bilirubin≤1.8 AND AST≤420.0
THEN Class=LIVE

R3: IF Spiders=Yes AND Age≥30.0 AND 62.0≤ALP≤175.0
AND Albumin≤4.3 AND Protime≤85.0
THEN Class=DIE

R4: ELSE Class=LIVE

Note: AST: SGOT (aspartate transaminase)
ALK phosphatase: ALP (alkaline phosphatase)

5.2.4. Comparison of extracted rules obtained by Sampling-Continuous Re-RX and EvoC

As shown in Table 8, the concrete rule set for diagnosis of the Hepatitis dataset is the only one obtained using a two-phase hybrid evolutionary classifier [53]. Therefore, we compared the rule set obtained using Sampling-Continuous Re-RX, shown in Section 5.2.2, with that obtained using EvoC, shown in 5.2.3 [53].

The number of rules and average number of antecedents of the rule set obtained using Sampling-Continuous Re-RX was 5.0 and 1.67, respectively. Furthermore, only two attributes were used in the rule set: ascites and albumin.

On the other hand, the number of rules and average number of antecedents of the rule set obtained using EvoC was 4.0 and 3.25, respectively. Nine attributes were used in the rule set is nine. Regarding the accuracy of the test dataset, as shown in Table 8, Sampling-Continuous Re-RX achieved 83.24% using 5 runs of 2-fold CV (repeated-randomized-hold-out approach), while EvoC achieved 83.92% using a average of more than 100 runs. Therefore, in terms of statistical validity, our proposed method achieved considerably better accuracy.

Moreover, the ELSE condition used in R4 obtained by EvoC is itself a black box. One of the major objectives of rule extraction is to provide a clear understanding and explanation of a predictive model. The ELSE part blindly assigns a class label to the samples. This black box part is thereby unable to describe the entire corresponding data space.

In summary, Sampling-Continuous Re-RX achieved considerably better accuracy with a much more concise rule set in terms of the number of rules, the average number of antecedents, and the number of attributes used in the extracted rule set than that of EvoC for the Hepatitis dataset.

6. Discussion

In Section 6.1, we explain the role of four kinds of biomarkers in the diagnosis of liver disorders, and in Section 6.2, we describe liver enzymes and the risk of liver disease. Next, in Section 6.3, we discuss the medical informatics interpretation of the rules extracted in the present study, and in Section 6.4, we address an important trade-off issue between the accuracy and the number of extracted rules.

6.1. Role of four different kinds of biomarkers in the diagnosis of liver disease

Assays for GGT, ALT, AST and ALP are the most common laboratory tests used for the detection of liver disease. Circulating GGT can be found in serum and on the external surfaces of most cells, especially hepatocytes; it is used as a biological marker for excessive alcohol intake. ALT and AST, which are abundantly present within hepatocytes, catalyze the transfer of amino groups to generate products

Table 8
Performance of previous rule extraction algorithms for the Hepatitis dataset.

Rule extraction method [validation method] [Ref.]	TR ACC (%)	TS ACC (%)	# Rules	Rule set	Total ante. # .	Ave. # ante.	Year
Two-phase Hybrid Evolutionary Classifier [Averaged over 100 runs] [53]	85.04	83.92	2.93	**Yes**	–	–	2003
C4.5 [Averaged over 100 runs] [36]	–	78.94	5.85	No	–	–	1992
PART [Averaged over 100 runs] [58]	–	80.02	6.64	No	–	–	1998
Wind-Driven Swarm Optimization (WSO) [5*2CV] [50]		63.58–76.92	14–52	No	–	–	2015
Decision Table [10CV] [59]	–	81.93	28	No	–	–	2011
Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [10CV] [60]	–	78.06	4	No	–	–	2011
Partial C4.5 Decision Tree [10CV] [61]	–	84.51	8	No	–	–	2011
Ripple Down Rule Learner [10CV] [62]	–	78.71	2	No	–	–	2011
Sampling-Continuous Re-RX [5×2CV]	89.04	**83.24**	**3.50**	**Yes**	7	**1.90**	**Present study**
Sampling-Continuous Re-RX [10×10CV]	89.24	**82.08**	**5.50**	**Yes**	13	**2.22**	**Present study**

Re-RX: Recursive-Rule eXtraction; TR: training dataset; TS: testing dataset; ACC: accuracy; Ave. # ante.: average number of antecedents; Total # ante.: total number of antecedents; 10CV: 10-fold cross validation; 10×10CV: 10 runs of 10-fold cross validation; 5×2CV: 5 runs of 2-fold cross validation (Repeated-randomized-hold-out approach).

in gluconeogenesis and amino acid metabolism. ALP is a hydrolase enzyme that catalyzes the hydrolysis of inorganic pyrophosphate, a vascular calcification inhibitor. Serum ALP is commonly used as a marker of liver or bone disease in clinical practice [63].

6.2. Serum ALT activity

Physicians, predominantly hepatologists and gastroenterologists, treating patients with liver disease have long known that the measurement of liver enzyme activities (serum aminotransferases, including ALT and AST) is critical in the diagnosis and assessment of liver disease [45].

6.2.1. Serum ALT blood test

ALT measurement is a low-cost, readily available blood test utilized throughout the US to detect liver disease. It is a valuable screening test to detect largely undiagnosed liver diseases such as asymptomatic viral hepatitis and nonalcoholic fatty liver disease (NAFLD).

ALT levels differ according to sex, with men having higher values than women. Additional factors that affect serum ALT levels include body mass index and triglyceride levels, regardless of sex. In men, total cholesterol levels and alcohol consumption have a positive correlation with ALT levels, whereas smoking, physical activity, and age have a negative correlation [45].

6.2.2. ALT as an indicator of liver disease

Since serum ALT levels increase in disease states that cause hepatocellular injury, they are effective for identifying ongoing liver disease. If elevated ALT levels are associated with symptoms such as fatigue, anorexia, or pruritus, the probability of clinically significant liver disease increases. The effectiveness of additional evaluation in patients with asymptomatic elevation of ALT depends on the results of physical examinations and the length of time and degree to which ALT levels have been elevated [45].

6.3. Serum GGT activity

GGT is a sensitive marker of hepatobiliary disorders, although nonspecific to its cause, found in hepatocytes and biliary epithelial cells [46]. In the clinical setting, blood GGT is used to indicate liver injury.

6.3.1. GGT as an indicator of liver disease

GGT is regarded as less specific than ALT for liver injury, and is used less frequently for the detection and monitoring of liver disease. However, as a prognostic indicator, it may be as discriminating as ALT for liver disease, and more discriminating for other diseases. The increased all-cause mortality found in this study associated with elevated GGT is supported by a number of previous studies in terms of disease association and mortality.

We confirmed the association between GGT and liver injury by showing a mortality risk at least as great as that for elevated ALT, even after adjusting for several known liver disease risk factors. GGT has been strongly associated with both alcoholic and NAFLD. However, surprisingly, less evidence has been reported for elevated GGT and liver disease outcomes, including mortality [47].

6.4. Medical informatics interpretation of the rules extracted from the BUPA dataset

We explained the reason why the present extracted rules achieved very good results in Section 5.1.1. In this section, we attempt to interpret how nine rules play a role in the diagnosis of non-liver or liver disorder.

We consider abnormal liver enzyme values as a serum concentration of ALT and AST greater than 30 U/L or GGT greater than 51 U/L, based on prior publications with liver enzyme and fatal outcomes. Additionally, combinations of liver abnormalities, including in any one enzyme elevation (ALT > 30 U/L, AST > 37 U/L, or GGT > 51 U/L) or in all three enzymes (ALT > 30 U/L, AST > 37 U/L, and GGT > 51 U/L), or an AST/ALT ratio greater than 1.0, were assessed [64].

Interpretations of extracted rules are as follows:

R1 and R2 state that a GGT of 22 and an ALT of 21 are critical cutoff points. In the same range values of GGT and ALT, if AST > 27, then disorder is diagnosed, as shown in R2.

R3 and R4 state that a GGT$\in$(22, 31] and an ALT of 21 are critical cutoff points. In the same range values of GGT and ALT, if AST > 11, then disorder is diagnosed, as shown in R4.

R5 and R6 state that a GGT$\in$(22, 31] and an ALT of 34 are critical cutoff points. In the same range values of GGT and ALT, if ALT > 34, then disorder is diagnosed, as shown in R6.

R7 states the upper limit of GGT for non-disorder. If GGT > 31, then it is definitely disorder, regardless of the DRNO$\leq$9.

R8 and R9 state that a GGT of 31 and a DRNO$\in$(9, 10] are critical points. In the same range values, if AST > 28, then disorder is diagnosed, as shown in R9.

We think that these rules can be applied to the diagnosis of the BUPA dataset. We hope that the proposed algorithm could also be adapted to similar liver disorder datasets to extract diagnostic rules.

6.5. Trade-off between the accuracy and number of extracted rules for BUPA dataset

In the case of medical rule extraction, a trade-off is apparent between high diagnostic accuracy and interpretability. Thus, if a physician wishes to extract rules with high diagnostic accuracy from medical datasets, they can choose the algorithm with high diagnostic accuracy, but reduced interpretability. However, in other situations, a physician may want to obtain extracted diagnostic rules with more interpretability but reduced accuracy [40].

Needless to say, if the best trade-off can be found, then the best extracted rules can be obtained. Ideally, we hope to extend the trade-off curve to obtain a wider viable region that provides improvements in both diagnostic accuracy and interpretability [40].

Recently, Fortuny, and Martens [65] expressed the same opinion: Rule extraction is a technique that attempts to find compromise between both requirements by building a simple rule set that mimics how the well-performing complex model (black box) makes decisions.

As shown in Table 6, five fuzzy rule extraction algorithms were proposed for the BUPA dataset. In contrast, only one concise rule extraction algorithm was proposed for the same dataset.

To allow a better understanding of our claim, the best trade-off curve between the accuracy and number of rules extracted is shown in Fig. 4. The reciprocal of the number of rules extracted is shown on the x-axis. The red dot, which is located at the trade-off curve, shows the performance of the proposed algorithm. This demonstrates that the present algorithm provided extracted rules for the BUPA dataset that were both accurate and concise.

The three green dots obtained by hyper-rectangular composite NNs [41], interpretable-fuzzy-rule-based classification [42], and fuzzy rule extraction from trained SVMs [43] may provide better accuracy and/or fewer rules compared with the algorithm used in the present student.

However, in general, fuzzy rules involve strong expressive power by linguistic and intuitive expressions. Thus, the number of fuzzy rules is not equivalent to the same number of concise rules in terms of expressive power. On the contrary, the number of fuzzy rules should be considered much more important than the number of concise rules.

Considering the potential for the more expressive power of fuzzy rules, all of the green dots for fuzzy rules should be shifted horizontally to the left, which result in being beyond the trade-off curve [40].

Consequently, the red dot obtained by the proposed algorithm is the

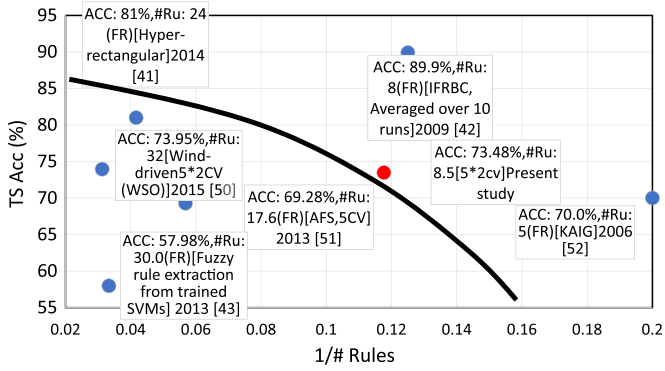

Fig. 4. Trade-off curve between the accuracy and number of rules extracted for the BUPA dataset. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

closest to the trade-off curve, and shows a well-balanced performance between accuracy and number of rules.

6.6. Child-Pugh score

In gastroenterology, the Child-Pugh score (sometimes referred to as the Child-Turcotte-Pugh score) [48,49] is used to assess the prognosis of chronic liver disease, primarily cirrhosis. Although the Child-Pugh score was originally used to predict mortality during surgical procedures, it is presently used to determine prognosis, as well as the required strength of treatment and the necessity of liver transplantation. The Child-Pugh score employs five clinical measures of liver disease, each of which is scored 1–3, with 3 indicating the most severe derangement [66], as shown in Table 9. A numerical score is assigned for patients in each parameter (albumin, bilirubin, prothrombin time, ascites, encephalopathy), after which, patients are categorized into Child A (5–7 points), B (8–11 points), or C (12–15 points), with class C patients presenting with the most abnormalities.

6.7. Medical informatics interpretation of the rules extracted from hepatitis dataset

Apparently, the rule set for diagnosis of the Hepatitis dataset is quite concise, as it only consists of three rules and two attributes, i.e., ascites and albumin.

On the other hand, although the Child–Pugh score is widely used to evaluate hepatic reserve function and related problems, such as non-objective factors (ascites and hepatic encephalopathy). Child–Pugh A (a patient with a Child–Pugh score of 5 points) includes chronic hepatitis with normal hepatic function and early liver cirrhosis accompanied by slightly abnormal hepatic function [67]. A simple and objective method for evaluation of hepatic reserve function using only albumin and total bilirubin measurements was recently proposed as albumin–bilirubin (ALBI) grade [68].

Considering reports in the literature [64,67,68], rule R1 extracted by the proposed algorithm is concise and reasonable. Rules R2 and R3 sharply derived an important cut-off point for serum albumin (3.5) for Child-Pugh A. Since the presence and severity of ascites is quite serious overall, rule R3 seems a little bit optimistic; however, we believe that meaningful rules were extracted from the Hepatitis dataset, which has been widely used for medical benchmarks.

6.8. Trade-off between the accuracy and number of extracted rules for the Hepatitis dataset

As shown in Table 8, eight rule extraction algorithms were proposed for the Hepatitis dataset. To allow a better understanding of our claim, the best trade-off curve between the accuracy and number of rules

Table 9
Classification measure of the Child-Pugh score.

Measure	1 Point	2 Points	3 Points
Total bilirubin (mg/dL)	< 2	2–3	> 3
Serum albumin (g/dL)	> 3.5	2.8–3.5	< 2.8
Prothrombin time, prolongation (%)	> 60	40–60	< 40
Ascites	None	Mild (or suppressed with medication)	Moderate to severe (or refractory)
Hepatic encephalopathy	None	Grade I–II	Grade III–IV

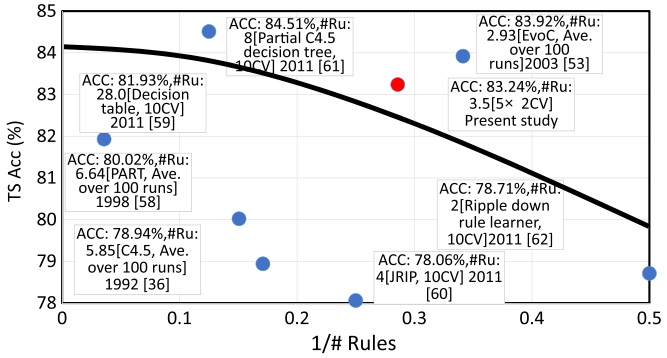

Fig. 5. Trade-off curve between the accuracy and number of rules extracted for the Hepatitis dataset. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

extracted is shown in Fig. 5. The reciprocal of the number of rules extracted is shown on the x-axis. The red dot, which is located at the trade-off curve, shows the performance of the proposed algorithm. This demonstrates that the present algorithm provided extracted rules that were both accurate and concise for the Hepatitis dataset.

As described in Section 5.2.4, the proposed algorithm provided considerably better performance than that of EvoC [53]. Since the rule extraction algorithm by partial C4.5 decision tree [61] did not demonstrate a rule set in the literature, we cannot compare those performances directly. The rule extraction algorithm results were close to the trade-off curve and showed slightly higher accuracy (84.51%) than that of the proposed algorithm (83.24%). However, interpretability was considerably lower (the number of rules was 8.0) than that of proposed algorithm (the number of rules was 3.5).

Consequently, the red dot obtained by the proposed algorithm is the closest to the trade-off curve and shows well-balanced performance between accuracy and the number of rules.

7. Conclusions

In this paper, we proposed Sampling-Continuous Re-RX as a new algorithm for extracting highly accurate and interpretable rules for the BUPA and Hepatitis datasets. We showed a rule set extracted from the BUPA dataset and provided a medical informatics interpretation of the extracted rules.

We also demonstrated that the extracted rules using the proposed method were close to the trade-off curve, meaning that they were more accurate and interpretable, and therefore more suitable for medical decision making. Actually, high accuracy and interpretability were achieved simultaneously using the proposed Sampling-Continuous Re-RX algorithm for the BUPA and Hepatitis datasets.

We have previously developed the Re-RX family [69], which can deal with various rule extraction situations in the medical setting; that is, accuracy-priority types [35] and interpretability-priority types [70].

In the clinical setting, the measurement of liver enzymes, particularly GGT and ALP, may serve as a prognostic tool for the long-term

prediction of mortality from liver disease [63]. Therefore, the rules extracted for the BUPA dataset in the present study suggest the presence of an association between GGT and ALP levels and liver disease.

Additional research is needed to clarify this association. However, in the absence of such data, slightly elevated levels of these enzymes, even within normal ranges, may indicate a risk of liver disease and suggest the need for further clinical evaluation.

Only three rules were extracted to diagnose the Hepatitis dataset. These rules sharply derived the cut-off point (3.5) of albumin for the Child-Pugh score. As previously reported [68], the albumin level is very important and useful; therefore, in the future, we hope to successfully extract interpretable rules from hepatic function datasets for the diagnosis of hepatitis based on the presence and severity of ascites and the levels of several biomarkers, such as albumin and bilirubin, which may be indicative of various types of hepatitis.

Regardless, the complex problem of diagnosing liver disease with practical diabetes datasets needs to be recognized. Based on the findings in the present study, we hope to extract even more meaningful diagnostic rules for more recent liver disease datasets in the future.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Stewart WB, Wild CP. World cancer report 2014. Geneva, Switzerland: WHO Press; 2014, ISBN 978-92-832-0432-9.
[2] Blachier M, Leleu H, Peck-Radosavljevic M, Valla DC, Roudot-Thoraval F. The burden of liver disease in Europe. Geneva, Switzerland: European Association for the Study of the Liver (EASL); 2013, ISBN 978-2-8399-1176-4.
[3] Acharya UR, Faust O, Molinari F, Sree SV, Junnarkar SP, Sudarshan V. Ultrasound-based tissue characterization and classification of fatty liver disease: a screening and diagnostic paradigm. Knowl Based Syst 2015;75:66–77.
[4] European Association for the Study of Liver . EASL clinical practical guidelines: management of alcoholic liver disease. J Hepatol 2012;57:399–420.
[5] Rehm J, Samokhvalov AV, Shield KD. Global burden of alcoholic liver diseases. J Hepatol 2013;59:160–8.
[6] Polat K. Application of attribute weighting method based on clustering centers to discrimination of linearly non-separable medical datasets. J Med Syst 2012;36:2657–73.
[7] Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Syst Appl 2014;41:2239–49.
[8] Seera M, Lim CP, Tan SC, Loo CK. A hybrid FAM–CART model and its application to medical data classification. Neural Comput Appl 2015;26:1799–811.
[9] Çomak E, Polat K, Güneş S, Arslan A. A new medical decision making system: least square support vector machine (LSSVM) with fuzzy weighting pre-processing. Expert Syst Appl 2007;32:409–14.
[10] Zangooei MH, Habibi J, Alizadehsani R. Disease diagnosis with a hybrid method SVR using NSGA-II. Neurocomputing 2014;136:14–29.
[11] Fan CY, Chang PC, Lin JJ, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 2011;11:632–4.
[12] Mohapatra P, Chakravarty S, Dash PK. An improved cuckoo search based extreme learning machine for medical data classification. Swarm Evol Comput 2015;24:25–49.
[13] Li DC, Liu CW, Hu SC. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010;40:509–18.
[14] Peker M. A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. J Med Syst 2016;40:116.
[15] Özşen S, Güneş S. Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. Expert Syst Appl 2009;36:386–92.
[16] Polat K, Şahan S, Kodaz H, Güneş S. Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. Expert Syst Appl 2007;32:172–83.
[17] Wang K-J, Adrian AM, Chen K-H, Wang K-M. An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. J Biomed Inform 2015;54:220–9.
[18] Chang P-C, Lin J-J, Liu C-H. An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. Comput Methods Prog Biomed 2012;107:382–92.
[19] Kraipeerapun P, Fung CC. Binary classification using ensemble neural networks and interval neutrosophic sets. Neurocomputing 2009;72:2845–56.
[20] Luukka P. Fuzzy beans in classification. Expert Syst Appl 2011;38:4798–801.
[21] Gonçalves LB, Vellasco MMBR, Pacheco MAC, Souza FJ. Inverted hierarchical neuro-fuzzy BSP System: a novel neuro-fuzzy model for pattern classification and rule extraction in databases. IEEE Trans Syst Man Cybern-Part C: Appl Rev 2006;36:236–48.
[22] Beheshti Z, Shamsuddin SMH, Beheshti E, Yuhaniz SS. Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis. Soft Comput 2014;18:2253–70.
[23] Yu K, Ji L, Zhang X. Kernel nearest-neighbor algorithm. Neural Process Lett 2002;15:147–56.
[24] Li DC, Liu CW, Hu SC. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med 2011;52:45–52.
[25] Luukka P. Classification based on fuzzy robust PCA algorithms and similarity classifier. Expert Syst Appl 2009;36:7463–8.
[26] Liu R, Chen Y, Jiao L, Li Y. A particle swarm optimization based simultaneous learning framework for clustering and classification. Pattern Recognit 2014;47:2143–52.
[27] Fan Y-J, Chaovalitwongse WA. Optimizing feature selection to improve medical diagnosis. Ann Oper Res 2010;174:169–83.
[28] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput Math Methods Med 2015(10):1155.
[29] Polat K, Gunes H. Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. Digit Signal Process 2006;16:889–901.
[30] Sartakhti JS, Zangooei MH, Mozafari K. Heapatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Comput Methods Program Biomed 2012;108:570–9.
[31] Calisir D, Dogantekin E. A new intelligent hepatitis diagnosis system: PCA-LSSVM. Expert Syst Appl 2011;38:10705–8.
[32] Polat K, Gunes S. Medical decision support system based on artificial immune recognition immune system (AIRS): fuzzy weighted pre- processing and feature selection. Expert Syst Appl 2007;33:484–90.
[33] Napierala K, Stefanowski J. BRACID: a comprehensive approach to learning rules from imbalanced data. J Intell Inf Syst 2012;39:335–73.
[34] Setiono R, Baesens B, Mues C. Recursive neural network rule extraction for data with mixed attributes. IEEE Trans Neural Netw 2008;19:299–307.
[35] Hayashi Y, Nakano S, Fujisawa S. Use of the recursive-rule extraction algorithm with continuous attributes to improve diagnostic accuracy in thyroid disease. Inform Med Unlocked 2015;1:1–8.
[36] Quinlan JR. C4.5: programs for machine learning, Morgan Kaufmann series in machine learning. San Mateo, California: Morgan Kaufman Inc.; 1993.
[37] Setiono R. A penalty-function approach for pruning feedforward neural networks. Neural Comput 1997;9:185–204.
[38] Setiono R. Sampling selection and neural network rule extraction for credit scoring. In: Proceedings of the 43rd decision sciences institutes annual meeting; 2012. pp. 1280–90
[39] Setiono R, Azcarraga A, Hayashi Y. Using sample selection to improve accuracy and simplicity of rules extracted from neural networks for credit scoring applications. Int J Comput Intell Appl 2015;14, 1550021-1–1550021-20.
[40] Hayashi Y, Yukita S. Rule extraction using recursive-rule extraction algorithm with J48graft with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the pima indian dataset. Inform Med Unlocked 2016;2:92–104.
[41] Hsieh Y-Z, Su MC, Wang PC. A PSO-based rule extractor for medical diagnosis. J Biomed Inform 2014;49:53–60.
[42] Gadaras I, Mikhailov L. An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif Intell Med 2009;47:25–41.
[43] Chavas ADF, Vallasco MMBR, Tanscheit R. Fuzzy rules extraction from support vector machines for multi-class classification. Neural Comput Appl 2013;22:1571–80.
[44] University of California, Irvine Learning Repository. ⟨http://archive/ics.uci/edu/m/⟩; 2016 [accessed 25.09.16]
[45] Kim WR, Flamm SL, Di Bisceglie AM, Bodenheimer HC. Public policy committee of the American association for the study of liver disease, serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. Hepatology 2008;47:1363–70.
[46] Beek JHDA, Moor MHM, Geus EJC, Lubke GH, Vink JM, Willemsen G, Boomsma DI. The genetic architecture of liver enzyme levels: GGT, ALT and AST. Behav Genet 2013;43:329–39.
[47] Ruhl CE, Everhart JE. Elevated serum alanine aminotransferase and gamma-glutamyltransferase and mortality in the United States population. Gastroenterology 2009;136:477–85.
[48] Child CG, Turcotte JG. Surgery and portal hypertension. The liver and portal hypertension. Saunders, Philadelphia; 1964. pp. 50–64
[49] Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. Br J Surg 1973;60:646–9.
[50] Christopher JJ, Nehemiah HK, Kannan A. A swarm optimization approach for clinical knowledge mining. Comput Methods Programs Biomed 2015;121:137–48.
[51] Liu X, Feng X, Pedrycz W. Extraction of fuzzy rules from fuzzy decision trees: an axiomatic fuzzy sets (AFS) approach. Data Knowl Eng 2013;84:1–25.
[52] Su CT, Chen LS, Yih Y. Knowledge acquisition through information granulation for imbalanced data. Expert Syst Appl 2006;31:531–41.
[53] Tan KC, Yu Q, Heng CM, Lee TH. Evolutionary computing for knowledge discovery in medical diagnosis. Artif Intell Med 2003;27:129–54.
[54] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal

approximators. Neural Netw 1989;2:359–66.
[55] MacQueen B. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley; 1967; 1. pp. 281–97
[56] Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1997;1:317–28.
[57] Marqués AI, García V, Sánchez JS. On the suitability of resampling techniques for the class imbalance problem in credit scoring. J Oper Res Soc 2013;64:1060–70.
[58] Frank E, Witten IH. Generating accurate rule sets without global optimization. In: Proceedings of the 15th international conference machine learning. Madison, WI, USA. San Francisco, CA: Kaufmann, Morgan; 1998. pp. 144–51
[59] Kohavi Ron. The power of decision tables. In: Proceedings of the European conference on machine learning; 1995; 912. pp. 174–189
[60] Cohen William W, Fast effective rule induction. In: Proceedings of the twelfth international conference on machine learning; 1995. pp. 115–23
[61] Frank Eibe, Witten Ian. Generating accurate rule sets without global optimization. In: Proceedings of the fifteenth international conference on machine learning; 1998. pp. 144–51
[62] Gaines Brian R, Compton Paul. Induction of ripple-down rules applied to modeling large databases. Int Intell Inf Syst 1995;5(3):211–28.
[63] Kunutsor SK, Apekey TA, Seddoh D, Walley J. Liver enzymes and risk of all-cause mortality in general populations: a systematic review and meta-analysis. Int J Epidemiol 2014;43:187–201.
[64] Hernaez R, Yeh H-C, Lazo M, Chung H-M, Hamilton JP, Koteish A, et al. Elevated ALT and GGT predict all-cause mortality and hepatocellular carcinoma in Taiwanese male: a case-cohort study. Hepatol Int 2013;7:1040–9.
[65] Fortuny EJ, Martens D. Active learning-based pedagogical rule extraction. IEEE Trans Neural Netw Learn Syst 2015;26:2664–77.
[66] Cholongitas E, Papatheodoridis GV, Vangeli M, Terreni N, Patch D, Burroughs AK. Systematic review: the model for end-stage liver disease – should it replace Child-Pugh's classification for assessing prognosis in cirrhosis?. Aliment Pharmacol Ther 2005;22:1079–89.
[67] Hiraoka A, Kumada T, Michitaka K, et al. Usefulness of albumin–bilirubin grade for evaluation of prognosis of 2584 Japanese patients with hepatocellular carcinoma. J Gastroenterol Hepatol 2016;31:1031–6.
[68] Johnson PJ, Berhane S, Kagebayashi C, et al. Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach-the ALBI grade. J Clin Oncol 2015;33:550–8.
[69] Hayashi Y. Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk dataset assessment from a Pareto optimal perspective. Oper Res Perspect 2016;3:32–42.
[70] Hayashi Y, Nakano S Satoshi. Use of a recursive-rule extraction algorithm with J48graft to archive highly accurate and concise rule extraction from a large breast cancer dataset. Inform Med Unlocked 2016;1:9–16.