A Novel Model based on Non Invasive Methods for Prediction of Liver Fibrosis

Mahmoud Nasr*, Khaled El-Bahnasy†, M. Hamdy‡ and Sanaa M. Kamal§
*†‡ Information Systems Department
Faculty of Computer and Information Sciences
§Hepatology Department
Faculty of Medicine
* Ain Shams University
Egypt, Cairo
* Email: *ma7moud.asp@gmail.com, †khaled.bahnasy@cis.asu.edu.eg, ‡m.hamdy@cis.asu.edu.eg, §sanaakamal@gmail.com

***Abstract*—Serial liver biopsies are typically the gold standard for diagnosis of liver fibrosis progression. However, It is associated with serious complications, inconvenient to patients and expensive, the challenge is to substitute the liver biopsy with non-invasive method. The proposed technique is employed to resolve this issue with average accuracy 99.48% for 5-folds cross validation. This accuracy pave the way to utilize classification models as a clinically non-invasive and reliable method to assess the degree of liver fibrosis.**

***Keywords*—Subsumption, Minimal unique rules, Classification, Knowledge representation**

I. Introduction

Hepatitis C virus (HCV) infection affects more than 170 million people worldwide[1]. Egypt has the highest prevalence of hepatitis C in the world with prevalence rates reaching 13%-15%. Thus, HCV represents a major public health and economic problem in Egypt[2]. HCV infection is marked by high tendency to persistence and evolution to chronic hepatitis with development of serious consequences such as cirrhosis and liver cancer in some patients[3]. To date, there are no indicators or reliable criteria that identify who would develop cirrhosis and when given that the rate of progression of fibrosis is highly variable. The treatment for HCV is both difficult and expensive. Egypt has launched a nationwide government sponsored campaign to treat patients with chronic hepatitis. The large pool of patients and the financial constraints necessitate prioritizing therapy to those most likely to progress rapidly to liver fibrosis[4]. Serial liver biopsies are typically the gold standard for diagnosis of liver fibrosis progression[5]. Liver biopsies are invasive, associated with serious complications, inconvenient to patients and expensive[6].Recently, several non-invasive serum markers and imaging techniques have emerged as tools for diagnosis of fibrosis. However, to date such biomarkers and imaging procedure have not be adequately validated as reliable alternatives for liver biopsy[7], [8]. Therefore, We aim to develop, evaluate and validate a prediction model that replaces the invasive techniques, and to be a measurement to liver fibrosis progression. Also, developing and validating computerized clinical decision-support system (CDSS) to support identification of individuals at higher risk of accelerated liver fibrosis progression.

The Clinical decision support systems (CDSS) use decision support system theory and technology to assist clinicians in the evaluation and treatment process. Using historical clinical data and the relationship processed by Artificial Intelligence (AI) techniques to aid physicians in their decision making process is the goal of CDSS [9], From a computational point of view, by a decision support system (DSS) we understand a computer-based information system assisting the decision-making process, and used to solve a large variety of real-life problems. Basically, DSSs are developed to support the solution of unstructured management issues in order to improve the decision-making process [10]. Recent advances in artificial intelligence (AI) and statistical learning (SL) enhanced these systems, giving rise to intelligent decision systems (IDS) [11]. Among the most popular approaches, one can mention the expert systems and well-known AI models, such as neural networks (NNs), genetic algorithms (GAs), support vector machines (SVMs), cluster analysis, intelligent agents, swarm intelligence, random forests, etc. [12], [13], [14]. The IDSs development has been encouraged by their effectiveness when applied in a large variety of real-world decision issues, such as: medical decision-making, business intelligence, customers relationship management, etc. The majority of IDSs based on machine learning (ML) techniques is built around a single algorithm and solves a specific problem only. There are various such approaches based on natural computing algorithms applied to different real-life problems. Recent studies propose the use of structured frameworks, usually known as committees of machines, involving more than one algorithm (e.g., NNs, SVMs) working together to solve a given problem [15], [16], [17]. The medical decision-making is nowadays one of the most promising fields to use IDSs. Thus, NNs, SVMs and classification trees have been proposed as standalone algorithms to solve medical decision problems, such as: prediction of severe acute pancreatitis at admission to hospital[18].

In medical field, Liver biopsy is an invasive procedure associated with some complications. Thus, different biomarkers and imaging techniques have been developed for non-invasive diagnosis of liver fibrosis. However, the equivalence of such non-invasive procedures to liver biopsy in diagnosis of liver fibrosis has not been proven. Furthermore, there are no tools to predict the risk and rate of liver fibrosis progression. Therefor we developed two algorithms , the first algorithm designed for searching the whole dataset to produce the full possible unique subset combinations that provide the domain expert

978-1-5386-4266-5/17/$31.00 ©2017 IEEE

with knowledge base of unique rules and the second algorithm is a rule based classifier designed to evaluate unique rules.

II. Related work

Numerous studies have shown that machine learning techniques are powerful tool in the medical sector with great prediction of liver fibrosis due to its ability of discovering the hidden predictive patterns from medical databases [19]. Some of them are black box like NN [20] and on the other hand, there are the distance measuring like KNN and rule based techniques like DT. Many related works varied in using these techniques upon the used datasets nature. Linear projection (LP) and Bayesian Networks (BN), were used to assess and identify associations between the HCV sequences and rate of fibrosis progression (RFP) which uses biological dataset with 90.38% accuracy [21]. The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C using C4.5 decision tree using clinical data with 73% accuracy [22]. A framework is built up on a weighted voting system, NN (RBF) and SVM designed to provide an automatic liver fibrosis progression for optimizing the decision-making process with 83% accuracy for 722 instances [23]. a classification model based on uses K-Nearest Neighbor and Neural Networks for liver fibrosis prediction with accuracy 66% for 771 instances [24]. Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C achieved 77% for clinical data with 722 instances [25]. A major challenge with medical datasets is the huge amount of data that are often referred to as high-dimensional data.

III. MATERIAL AND METHODS

A. Patients

This study includes 1741 Hepatitis C virus patients who have genotype 4. Patient's data is collected at Ain Shams University , Faculty of Medicine, El Demerdash Hospital. Patients were treated with a combined therapy interferon-Alfa and ribavirin for more than 15 months. The study shows patients who response to the treatment and others who does't show a clearance of the virus were considered as non responder.

B. Dataset analysis

This study demonstrates a HCV Liver Fibrosis dataset. It includes data for 1741 Egyptian patients who underwent treatment dosages. The collected data had several forms and structures. Therefore, a preprocessing stage of refinements has been applied based on expert recommendations. From a diagnosis perspective, the refined data includes 31 features which are described in Table [I]. The data in the previously mentioned dataset is labeled. The "Baseline histological staging" is the class label with values {F0, F1, F2, F3, F4}. These labels represent different prognosis levels of Liver Fibrosis as follows: No Fibrosis (F0), Portal Fibrosis (F1), Few Septa (F2), Many Septa (F3), Cirrhosis (F4).

IV. Minimal Unique Rules and Subsumption

Individual values of features and their combinations may discriminate class labels. This work applies an exhaustive search technique that generates all possible combinations of the feature values. These combinations are then be matched to the class labels. Therefore any combination which uniquely indicates only one class label is called a unique rule. The technique uses a Sequential Forward Generation (SFG). It starts searching from empty set of feature values and sequentially adds features until it reaches to a full set like the lattice in Figure **??**. A generated rule will contain a non-empty vector of feature values with a class label. For example, a rule, $Rule_x : < gender =' Male', Fever =' 1', Diarrhea =' 1' >\rightarrow F1$, has a combination of some feature values and associated to a specific class label $F1$. This rule is a unique rule, if its feature values combination appears with a distinct class label. The subsumption concepts indicates a containment relation between a combination and a previously discovered unique rule is called a subsumption. Figure 2 shows a subsumption relation between two rules (subsets) $\mathbf{s}_1 = \{B, C\}$ and $\mathbf{s}_2 = \{A, B, C\}$, As: $\mathbf{s}_1 \cap \mathbf{s}_2 = \mathbf{s}_1$, In this situation $\mathbf{s}_2$ is substituted with $\mathbf{s}_1$ and $\mathbf{s}_2$ is considered as a redundant subset. So, it should be pruned. This works consider the resultant set of unique rules can classify the whole dataset.

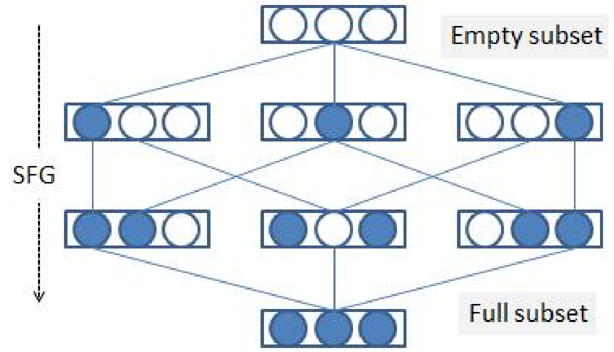


Fig. 1. A lattice for 3 features (each has binary feature values) with the SFG search combinations

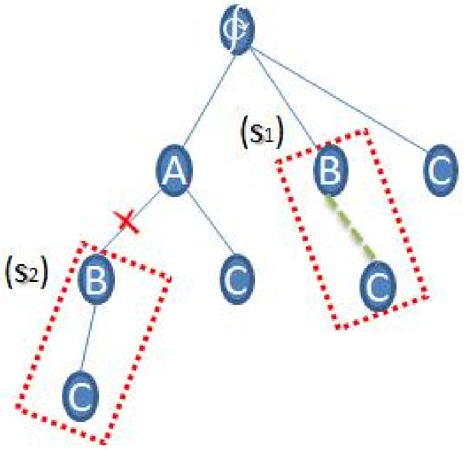


Fig. 2. A subsumption relation between two rules $s_1 ands_2$

V. The proposed Related Cases Skipper Algorithm (RCS)

Related Objects Skipper **RCS** algorithm is a complete search algorithm. It uses **SFG** technique for rules generation.

TABLE I. DESCRIPTION FOR THE DATASET FEATURES

Features Names	Values	Decritization
Age	32:61	$<= 32,> 32\& >= 37,> 37\& <= 42,< 42\& >= 47,< 47\& >= 52, < 52\& >= 57,< 57\& >= 62$
Gender	{Male,Female}	-
BMI(Body Mass Index)	22:35	$> 18.5, >= 18.5\& < 25, >= 25\& < 30, >= 30\& < 35, >= 35\& < 40$
Fever	1=absent,2=present	-
Nausea/Vomiting	1=absent,2=present	-
Headache	1=absent,2=present	-
Diarrhea	1=absent,2=present	-
Fatigue and generalized	1=absent,2=present	-
Bone ache	1=absent,2=present	-
Jaundice	1=absent,2=present	-
Epigastria pain	1=absent,2=present	-
WBC(White blood cell)	2991:12101	$< 4000, >= 4000\& < 11000, >= 11000$
RBC(red blood cells)	3816422:5018451	$< 3000000, >= 3000000\& < 5000000, >= 5000000$
HGB(Hemoglobin)	10:15	$If(Gender = 1) < 14, >= 14\& < 17.5, > 17.5 If(Gender = 2) < 12.3, >= 12.3\& < 15.3, >= 15.3$
Plat(Platelet)	93013:226464	$< 100000, >= 100000\& < 255000, > 255000$
AST 1(aspartate transaminase ratio 1 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT1(alanine transaminase ratio 1 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT1(alanine transaminase ratio 1 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT4(alanine transaminase ratio 4 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT12(alanine transaminase ratio 12 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT24(alanine transaminase ratio 24 week)	39:128	$< 20, >= 20\& < 40, >= 40$
ALT36(alanine transaminase ratio 36 week)	5:128	$< 20, >= 20\& < 40, >= 40$
ALT48(alanine transaminase ratio 48 week)	5:128	$< 20, >= 20\& < 40, >= 40$
ALT after 24 w	5:45	$< 20, >= 20\& < 40, >= 40$
RNA Base	11:1201086	$<= 5, > 5$
RNA 4	5:1201715	$<= 5, > 5$
RNA 12	5:3731527	$<= 5, > 5$
RNA EOT	5:808450	$<= 5, > 5$
RNA EF (Elongation Factor)	5:808450	$<= 5, > 5$
Baseline histological Grading	3:16	-
Baseline histological Staging(Class Label)	1:4	-

```
procedure SRBC
    for all t in T do                          ▷ Test data (T)
        for all ur in MUL do
            if ur is subsumed with t then
                Calculate W for each c         ▷ class label (c)
            end if
        end for
        Return c with max W
    end for
end procedure
```

Fig. 3. Subsumption Rule Based Classifier SRBC

And based on the subsumption relation, the algorithm ignores generating longer rules of any subsumed rules in a pruning process, This pruning process is skipping the generated unnecessary rules, therefore it is named as skipping process. Therefore **RCS** two skipping lists: Skipped-Indexes-List (**SIL**) which contains all indexes of related cases (ie. all cases which have the same values for a feature combination) and the Minimum-Unique-List (**MUL**) which contains all unique combinations and their related indexes. During the generation process , Suppose we have a combination $< \mathbf{s}, i >$, Where **s** is the feature values and (i) is the case index, **RCS** skips **s** while i is existed in **SIL**. Otherwise, it skips **s** which subsumed with **MUL** rules and all related cases as well. Finally, **RCS** checks **s** for uniqueness against the database. If (**s**) is unique then **RCS** adds **s** to **MUL** and adds all related cases indexes to the **SIL** if **s** is unique or not. After searching all the dimensional space **RCS** produces an **XML** Rule-based format as mentioned in Section VI. Figure 4 explains the proposed RCS algorithm.

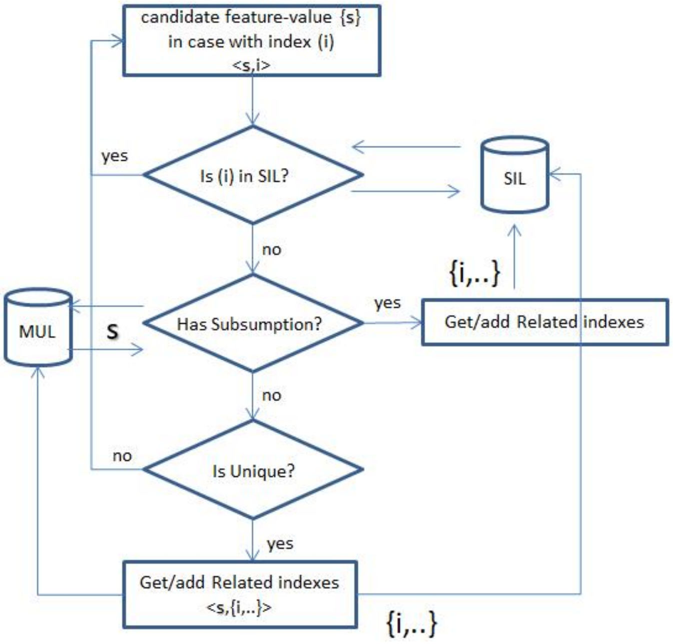
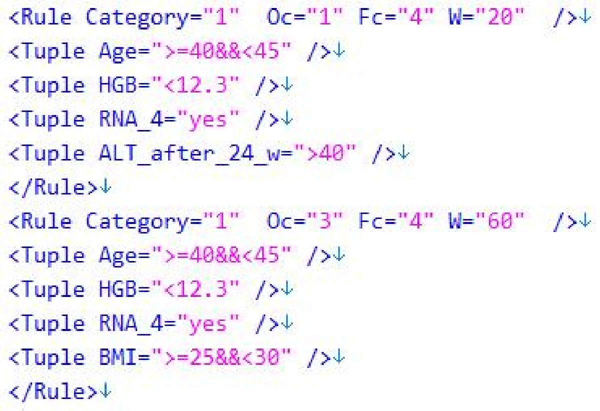


Fig. 4. RCS algorithm

VI. RULE BASED FORMAT AND KNOWLEDGE REPRESENTATION

RCS algorithm produces rules based on xml format like in Figure5. In this work the xml file plays as the model for **SRBC** which will be explained in the next section. Rules representation include two parts(tags): condition/premise part(Tuple tag) and action/conclusion part (Category tag). In the diagnosis domain, rules can be understood easily and help physicians in



```
<Rule Category="1"  Oc="1" Fc="4" W="20"  />↓
<Tuple Age=">=40&&<45" />↓
<Tuple HGB="<12.3" />↓
<Tuple RNA_4="yes" />↓
<Tuple ALT_after_24_w=">40" />↓
</Rule>↓
<Rule Category="1"  Oc="3" Fc="4" W="60"  />↓
<Tuple Age=">=40&&<45" />↓
<Tuple HGB="<12.3" />↓
<Tuple RNA_4="yes" />↓
<Tuple BMI=">=25&&<30" />↓
</Rule>↓
```

Fig. 5. XML for Rule based format

decision making. It includes meta-data to save the thumb of original dataset. During rule extraction, additional attributes are added to calculate many weight (**W**) criterion. The Two attributes: (Oc) and (Fc) represents the extracted rule coverage while the other is number of features in it. This knowledge representation can be easily used as HCV standard model for interchange on the Web or other semantic systems.

TABLE II. COMPARISON AMONG SRBC AND OTHER TREE BASED AND RULE BASED CLASSIFIERS

classifier	1-fold	2-fold	3-fold	4-fold	5-fold	Accuracy-Avg
ADTree	97.69	97.69	96.83	97.12	93.95	96.66
BFTree	95.68	98.85	98.56	95.68	95.68	96.89
DecisionStump	95.68	95.68	95.68	95.68	95.68	95.68
FT	99.71	98.27	98.27	99.14	95.97	98.27
J48	97.12	95.68	96.25	98.56	95.68	96.66
LADTree	97.98	96.83	98.56	98.27	96.83	97.69
LMT	97.12	97.12	97.12	97.69	95.68	96.95
NBTree	99.71	95.97	96.54	99.42	93.37	97.00
RandomForest	100.00	100.00	100.00	100.00	95.39	99.08
RandomTree	100.00	100.00	100.00	100.00	95.39	99.08
REPTree	97.41	95.68	95.68	95.68	95.68	96.02
SimpleCart	95.68	95.68	95.68	95.68	95.68	95.68
conjunctiveRule	95.68	95.68	95.68	95.68	95.68	95.68
DecisionTable	95.68	95.68	95.68	95.68	95.68	95.68
DTNP	95.68	95.68	95.68	95.68	95.68	95.68
JRIP	95.68	95.68	96.25	99.14	95.68	96.48
Nnge	100.00	100.00	100.00	100.00	93.95	98.79
OneR	95.68	95.68	95.68	95.68	95.68	95.68
PART	97.98	99.14	98.56	99.42	97.41	98.50
Ridor	98.56	95.10	96.25	96.54	95.10	96.31
ZeroR	65.42	65.42	65.42	65.42	65.42	65.42
SRBC	100.00	100.00	100.00	100.00	97.41	99.48

VII. SUBSUMPTIONS-RULE-BASED CLASSIFIER

Subsumptions Rule Based Classifier (**SRBC**) is a sequential classifier which uses an XML Rule-based format files as model. First, **SRBC** check subsumption for the test cases against the model(unique rules of the XML file). As a consequence, one or many rules are subsumed. Therefore **SRBC** calculates a weight ***W*** for each subsumed rule for final evaluation. **SRBC** gets max ***W*** which is calculated according to many criterion. The following equation represents the coverage criterion which is applied.

$\mathbf{W}_i = (\sum nOc_i * C_i)/T_i$

Where nOc_i is the number of cases that shares the same values of the features combination, C_i is the total number of cases with a class label ***i*** of the original dataset and T_i is the total number of all rules. Figure 3 represents a pseudopod for the SRBC algorithm.

And Figure[6] represents a pipeline which shows the interaction or dependency between **RCS** and **SRBC** algorithms.

VIII. EXPERIMENTAL RESULTS

By applying the first RCS algorithm against HCV dataset, it produced (98002) minimal unique rules. The following table shows the relation between rules count and sizes :

TABLE III. UNIQUE RULES DISTRIBUTION

Rule Size	Count
1	5
2	438
3	4793
4	17189
5	27351
6	25533
7	14323
8	5988
9	1722
10	566
11	70
12	23
13	1

And the graph[7] shows the distribution of that table **SRBC** is applied after dividing the dataset to 5-folds validation. It achieves average accuracy 99.48%. For benchmarking, the same HCV dataset is applied against some of rule based algorithms built in the **WEKA**[1] tool, **SRBC** outperformed them as shown in Table II.

IX. CONCLUSION

The **RCS** algorithm scans the database for same combination of feature-value only one time. It skips all related cases combinations what makes it faster and more efficient. In addition being a complete search algorithm it promising to gain all possible optimum solutions. In conjunction with **SRBC** , They beat other state of art algorithms

REFERENCES

[1] Mohamed, Amal Ahmed et al. Hepatitis C Virus: A Global View. World Journal of Hepatology 7.26 (2015): 26762680. PMC. Web. 4 Nov. 2017.

[2] Elgharably, Ahmed et al. Hepatitis C in Egypt Past, Present, and Future. International Journal of General Medicine 10 (2017): 16. PMC. Web. 4 Nov. 2017.

[3] Bourliere, M., et al. ”Validation and comparison of indexes for fibrosis and cirrhosis prediction in chronic hepatitis C patients: proposal for a pragmatic approach classification without liver biopsies.” Journal of viral hepatitis 13.10 (2006): 659-670.

[4] Trooskin, Stacey B., Helen Reynolds, and Jay R. Kostman. ”Access to costly new hepatitis C drugs: medicine, money, and advocacy.” Clinical Infectious Diseases 61.12 (2015): 1825-1830.

[5] Gebo, Kelly A., et al. ”Role of liver biopsy in management of chronic hepatitis C: a systematic review.” Hepatology 36.S1 (2002).

[6] Chalasani, Naga, et al. ”The diagnosis and management of nonalcoholic fatty liver disease: Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association.” Hepatology 55.6 (2012): 2005-2023.

[7] Cobbold, J. F. L., et al. ”Optimal combinations of ultrasoundbased and serum markers of disease severity in patients with chronic hepatitis C.” Journal of viral hepatitis 17.8 (2010): 537-545.

[1] https://www.cs.waikato.ac.nz/ml/weka/

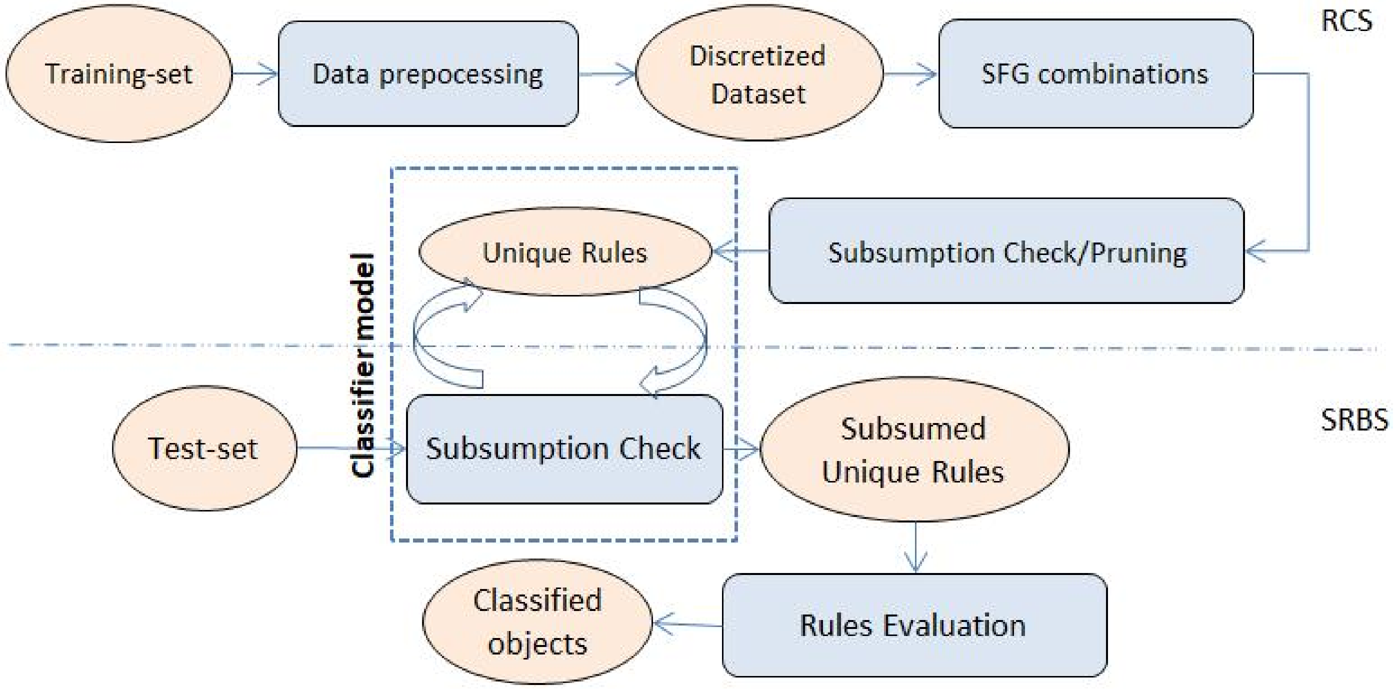


Fig. 6. a pipeline for the two algorithms

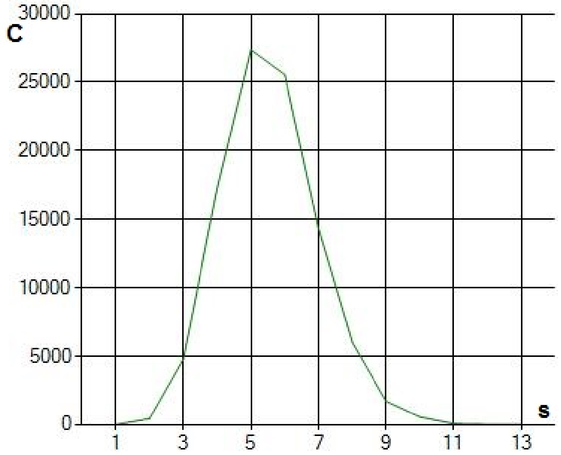


Fig. 7. Rules distribution against their size

[8] Kamal, Sanaa M., et al. ”Progression of fibrosis in hepatitis C with and without schistosomiasis: correlation with serum markers of fibrosis.” Hepatology 43.4 (2006): 771-779.

[9] Keltch, Brian, Yuan Lin, and Coskun Bayrak. ”Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients.” Journal of medical systems 38.8 (2014): 60.

[10] Power, Daniel J. ”A brief history of decision support systems.” DSSResources. COM, World Wide Web, http://DSSResources. COM/history/dsshistory. html, version 4 (2007).

[11] S. Guerlain, D.E. Brown and C. Mastrangelo, Intelligent Decision Support Systems, Proc. IEEE International Conference on Systems, Man, and Cybernetics 3 (2000), 193438

[12] E. Turban, J.E. Aronson and T-P. Liang, Decision Support Systems and Intelligent Systems (7th ed.), Prentice-Hall of India, 2006.

[13] T. Hastie, R. Tibshirani and R. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), Springer, 2009.

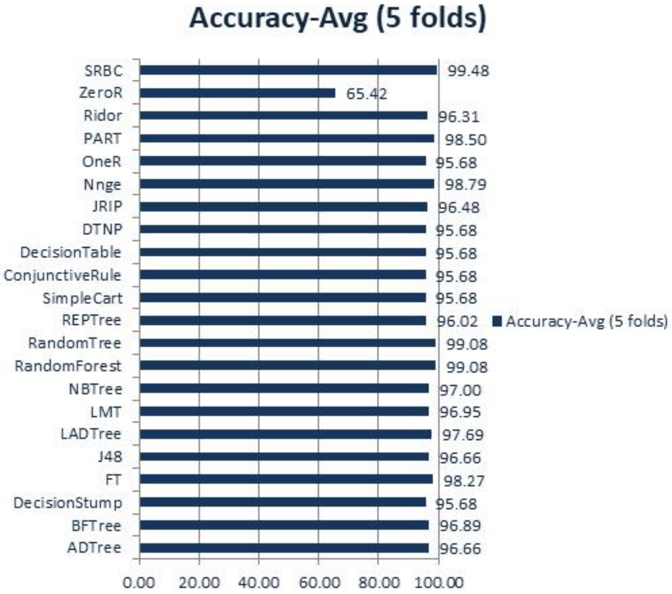


Fig. 8. SRBC accuracy vs other Rule based classifiers

[14] F. Gorunescu, Data Mining: Concepts, Models and Techniques, Springer-Verlag Berlin Heidelberg, 2011

[15] C. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

[16] S. Haykin, Neural networks. A comprehensive foundation (2nd ed.), Prentice-Hall. Inc., 1999.

[17] D. Valinicius, A. Verikas, M. Bacauskiene and A. Gelzinis, Evolving committees of support vector machines, LNCS Machine Learning and Data Mining in Pattern Recognition, vol. 4571 (2007), 263275.

[18] B. Andersson, R. Andersson, M. Ohlsson and J.J. Nilsson, Prediction of Severe Acute Pancreatitis at Admission to Hospital Using Artificial

Neural Networks, Pancreatology 11, (2011), 328335

[19] Hashem, Ahmed M., et al. ”Prediction of the degree of liver fibrosis using different pattern recognition techniques.”Biomedical Engineering Conference (CIBEC), 2010 5th Cairo International. IEEE, 2010.

[20] Cortez, Paulo, and Mark J. Embrechts. ”Opening black box data mining models using sensitivity analysis.” Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, 2011.

[21] Lara, James, et al. ”Computational models of liver fibrosis progression for hepatitis C virus chronic infection.” BMC bioinformatics 15.8 (2014): S5.

[22] Zayed, Naglaa, et al. ”The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C.”Clinics and research in hepatology and gastroenterology37.3 (2013): 254-261.

[23] Belciug, Smaranda, et al. ”Evolutionary-based intelligent decision model to optimize the liver fibrosis stadialization.”Annals of the University of Craiova-Mathematics and Computer Science Series40.2 (2013): 237-248.

[24] Mayer, Benjamin W., et al. ”Feature mining for prediction of degree of liver fibrosis.”AMIA Annual Symposium Proceedings. Vol. 2005. American Medical Informatics Association, 2005.

[25] Stoean, Ruxandra, et al. ”Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C.”Artificial intelligence in medicine51.1 (2011): 53-65.