machine learning, multiple classifier systems,
compound pattern recognition, medical informatics,
liver fibrosis, hepatitis C

Bartosz KRAWCZYK[1], Michał WOŹNIAK[1], Tomasz ORCZYK[2] , Piotr PORWIK[2],
Joanna MUSIALIK[3], Barbara BŁOŃSKA-FAJFROWSKA[4]

CLASSIFICATION TECHNIQUES FOR NON-INVASIVE RECOGNITION OF LIVER FIBROSIS STAGE

Contemporary medicine should provide high quality diagnostic services while at the same time remaining as comfortable as possible for a patient. Therefore novel non-invasive disease recognition methods are becoming one of the key issues in the health services domain. Analysis of data from such examinations opens an interdisciplinary bridge between the medical research and artificial intelligence. The paper presents application of machine learning techniques to biomedical data coming from indirect examination method of the liver fibrosis stage. Presented approach is based on a common set of non-invasive blood test results. The performance of four different compound machine learning algorithms, namely Bagging, Boosting, Random Forest and Random Subspaces, is examined and grid search method is used to find the best setting of their parameters. Extensive experimental investigations, carried out on a dataset collected by authors, show that automatic methods achieve a satisfactory level of the fibrosis level recognition and may be used as a real-time medical decision support system for this task.

1. INTRODUCTION

Early detection and stage identification of liver fibrosis, especially in a chronic type C hepatitis, is very important in the further therapy. Commonly used method for fibrosis stage determination is liver biopsy, but it is an invasive method which may cause risk of severe health complications. Also a single biopsy does not guarantee a required confidence about the fibrosis stage thus it is required to retrieve samples from more than one region of liver [1]. Some non-invasive tests methods are also available, but they are rather expensive and their availability is low – examples of such methods are: FibroTest by BioPredictive [2] and ELF Test by Siemens [3].

Analysis of the medical data is often a complex and time-consuming tasks. Therefore new filed known as medical decision support have risen [4]. It utilizes statistical and soft-computing methods for aiding the work of physicians. Countless successful real-life applications have proven that this approach not only saves the so-valuable time of the medical experts but additionally may lead to an improvement of the diagnostic quality [5]. Yet one should bore in mind that this field cannot be fully automatized – physicians are required not only as a source of the expert knowledge but also as a final link in the decision support chain – to evaluate the suggested diagnosis and to exploit fully the conclusions that can be drawn from the artificial intelligence algorithm's output.

Among many popular decision support techniques machine learning has gained a significant attention in last years. It allows, on the basis of previously gathered samples of data, to generate models that can generalize the attained knowledge on new, unseen objects [6]. In recent years more and more attention is being paid to the branch of machine learning known as Multiple Classifier Systems (MCS) [7]. They may be viewed as compound pattern recognition methods and are considered one of the most promising research directions in this field. Instead of relying on a single classifier MCS utilize a pool of available predictors and fuse their outputs to receive the final decision. This has been shown to usually improve the overall accuracy, as an ensemble of classifiers may outperform any single classifier from

[1] Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland.
e-mail : {bartosz.krawczyk , michal.wozniak}@pwr.wroc.pl
[2] University of Silesia, Institute of Computer Science, 41-200 Sosnowiec, Będzińska 39, Poland,
email: {tomasz.orczyk, piotr.porwik}@us.edu.pl,
[3] Dept. of Gastroenterology and Hepatology, Medical University of Silesia, Katowice, Poland,
[4] Dept. of Basic Biomedical Science, Medical University of Silesia, Sosnowiec, Poland.

a pool. It happens so due to the fact that several classifiers may contribute with different areas of competence and their fusion may lead to a significant decrease of their individual drawbacks. It must be pointed out that individual classifiers for the MCS should be selected carefully. Adding classifiers that differ little from each other (i.e., having a small diversity in the ensemble) shall lead only to the increase of the computational complexity of the system. In certain situations (such as voting algorithms for output fusion) it may even lead to a drop in the accuracy. On the other hand adding classifiers with high diversity but poor quality will produce a weak ensemble. Ideally MCS should consist of classifiers with high accuracy and high diversity [8].

In the field of MCS two main approaches may be distinguished – off-the-shelf [9] and custom-designed [10]. Latter methods need to be carefully designed. User need to prepare a pool of classifiers, taking care of their quality. Next a classifier selection method must be applied to choose only the most proper predictors. Finally a fusion method must be specified to combine the individual outputs. On the other hand off-the-shelf algorithms are ready to use classifiers in form of a black-box – user needs only to input values several parameters and the method will handle itself. In many cases the customized algorithms deliver better performance than black-box ones. Yet they are very sensitive to the proper selection of their components e.g., good classifier selection will be diminished by poorly designed fuser. Off-the-shelf ones are much more easy to use and therefore are recommended to the end-users without deeper insight into the machine learning field. That is why they are often a popular choice in medical decision support.

The article investigates the performance of four different black-box MCS applied to the liver fibrosis stage recognition. Experimental investigations concentrate on the proper selection of their parameters to maximize the final accuracy. Comparing these methods with the aid of a statistical significance testing we assess their usefulness as a data mining part of a real-life medical decision support system.

2. SOURCE DATA DESCRIPTION

The results of routine liver function tests from 127 patients mainly infected with chronic hepatitis type C virus but also with hepatitis type B virus and other chronic hepatitis were analysed. A total of 34 parameters of the blood were chosen. In case of all patients, a standard liver biopsy was performed and liver specimens were evaluated according to the METAVIR classification (Fibrosis score: F0 = no fibrosis; F1 = portal fibrosis without septa; F2 = portal fibrosis with few septa; F3 = numerous septa without cirrhosis; F4 = cirrhosis) [4]. The clinical characteristics of these patients are presented in Table 1.

3. EXPERIMENTAL INVESTIGATION

3.1. EXPERIMENTS SETUP

All experiments were carried out in the R environment, with classification algorithms taken from the dedicated packages, thus ensuring that the results achieved the best possible efficiency and that the performance was not decreased by a bad implementation. All tests were done by a 5x2 cv combined F-test [11] to assess if the differences between the tested methods are statistically significant.

In case of missing feature values the Expectation-Maximization (EM) [6] algorithm was applied to fill the gaps.

Table 1. Clinical and biological characteristics of patients.
* mean (std. deviation).

Age* (years)	50 (13)
Male, n(%)	75 (59%)
Biopsy result, n(%)	
F0	2 (2%)
F1	35 (28%)
F2	5 (4%)
F3	16 (13%)
F4	67 (53%)
HCV/HBV/other	70% / 9% / 21%
HB* (g/L)	14 (1.91)
RBC* (10^6/UL)	5 (0.74)
WBC* (10^3/UL)	6 (2.31)
PLT* (10^3/UL)	161 (70.75)
PT* (sec.)	13 (9.04)
PTP* (%)	90 (17.82)
APTT* (sec.)	38 (12.53)
INR*	1 (0.26)
ASPT* (IU/L)	65 (51.01)
ALAT* (IU/L)	72 (61.81)
ALP* (IU/L)	104 (55.11)
BIL* (mg/dL)	2 (2.69)
GGTP* (IU/L)	89 (94.43)
KREA* (mg/dL)	1 (0.23)
GLU* (mg/dL)	95 (19.02)
Na* (mmol/L)	138 (3.48)
K* (mmol/L)	5 (5.16)
Fe* (mmol/L)	104 (70.23)
CRP* (IU/L)	4 (25.38)
TG* (mg/dL)	107 (50.83)
CHO* (mg/dL)	189 (51.04)
Ur. acid* (mg/dL)	6 (1.39)
TP* (g/dL)	7 (0.81)
TIBC*	322 (120.47)
Neutr* (10^3/UL)	3 (1.35)
Lymph* (10^3/UL)	2 (0.55)
Mono* (10^3/UL)	1 (0.19)
Eos* (10^3/UL)	0 (0.13)
Baso* (10^3/UL)	0 (0.02)
Albu* (%)	58 (7.79)
Glb. α_1* (%)	3 (1.33)
Glb. α_2* (%)	8 (2.52)
Glb. β* (%)	11 (2.43)
Glb. γ* (%)	19 (7.21)

3.2. MACHINE LEARNING METHODS

For the experiment we applied four different off-the-shelf classification methods: Bagging [12], Boosting, Random Forest [13] and Random Subspaces [14].

For Bagging and Multi-class version of Boosting - AdaBoost.M2 [16] the C4.5 decision tree [15] was used as a weak classifier. was coupled with C4.5 classifier. As an individual classifier for Random Subspace the Support Vector Machine with polynomial kernel is applied [6].

3.3. TUNING THE CLASSIFIER PARAMETERS

Each of these algorithms has several parameters to be tuned. In this section we present a study conducted using a grid search with aim to establish an optimal setting for each of the compound classifiers.

The performance of Bagging algorithm strongly depends on the number of bags B used for the ensemble creation and the size of each of the bags n, describe as the percentage of the original training sample size. For the experimental evaluation we have selected $B = \{10;20;30;40;50\}$ and $n = \{0.3;0.7;1.0\}$. The results of the grid search are presented in the Fig. 1.

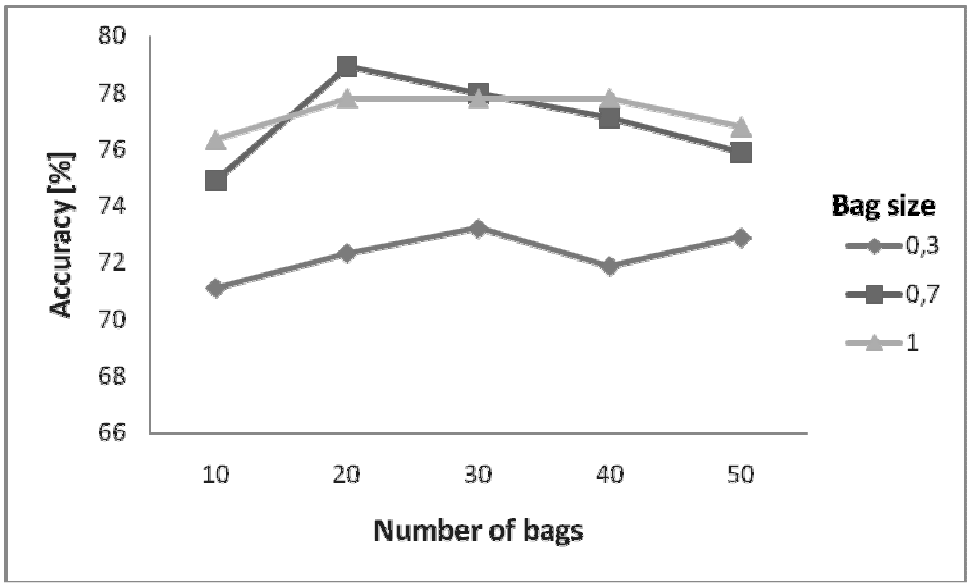

Fig. 1. Evaluation of Bagging parameters according to the size and the number of bags.

Interestingly the best performance is achieved for a bigger size of the bags. This may be due to the fact that larger subsamples lead to the creation of more stable individual classifiers. In case of the number of subsamples smaller ensembles tend to perform better – probably larger number of classifiers, coupled with the large single bag size, tend to lack in diversity.

For Boosting algorithm one must establish the optimal number of iterations I for which the algorithm will create new classifiers for the ensemble. For the experimental evaluation we have selected $I = \{10;20;30;40;50;60;70;80;90;100\}$. The results of the grid search are presented in the Fig. 2.

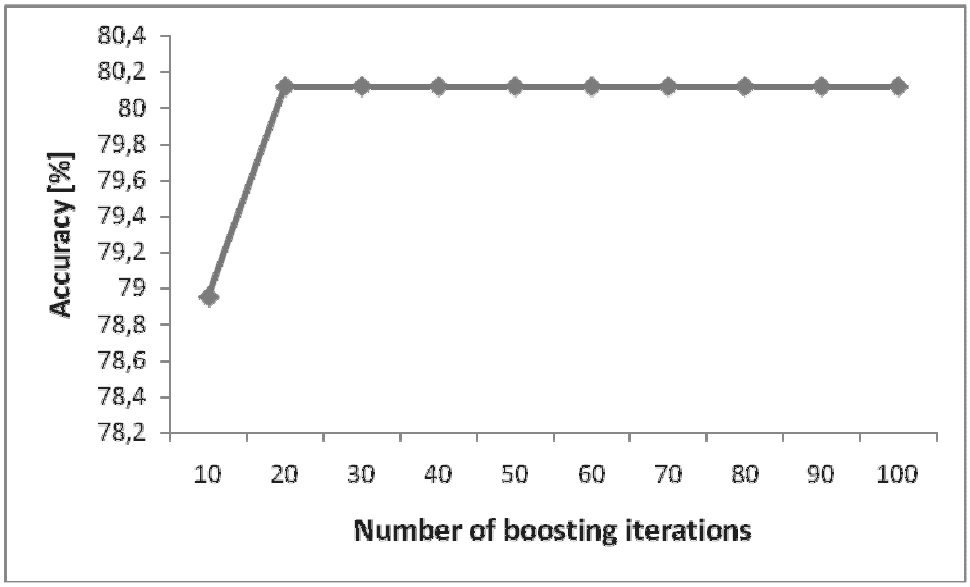

Fig. 2. Evaluation of the number of iterations for the Boosting approach.

Boosting reaches optimal accuracy in small number of iterations. With the increase of cycles no improvement can be reached – only the increase of the execution time.

The performance of the Random Forest algorithm relies strongly on two most important parameters – number of decision trees *N* used for the process of creation of the ensemble and their maximum depth *D*. For the experimental evaluation we have selected $N = \{20;40;60;80;120\}$ and $D = \{5;6;7;8\}$. The results of the grid search are presented in the Fig. 3.

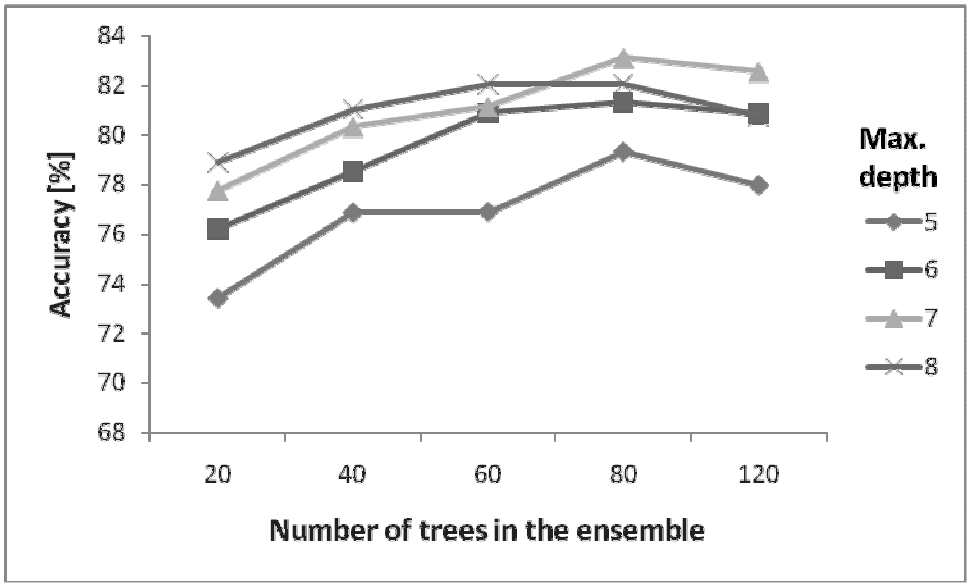

Fig. 3. Evaluation of Random Forest parameters with the respect to the size of the ensemble and maximum depth of a single tree.

Random Forest behaves better with larger number of trees, independently of the maximum depth of a single classifier. This confirms previous observations that for this method it is better to create a larger pool of weaker classifiers, improving the overall diversity this way.

In case of the Random Subspace method one needs to establish the number of subspaces *R* on which the classifiers for the ensemble will be constructed and the size of each of the subspaces *S*. In this paper the second parameter is expressed as a percentage of the original feature space that is included in the subspace after a random projection. For the experimental evaluation we have selected $R = \{10;20;30;40;50\}$ and $S = \{0.5;0.6;0.7;0.8\}$. The results of the grid search are presented in the Fig. 4.

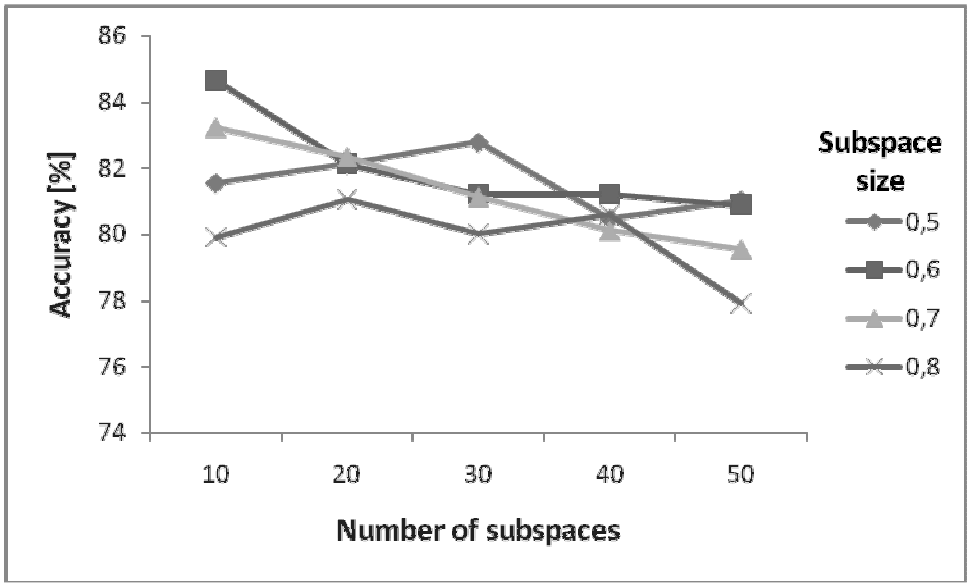

Fig. 4. Evaluation of Random Subspace with the respect to the number and size of subspaces.

Contrary to the previous method, the Random Subspace works best with small number of base classifiers. This may be caused by the SVM selected as a base model – as it is a strong classifier therefore it is hard to maintain a high diversity in the pool when operating on a large number of predictors.

The presented tests allowed us to select optimal parameter setting for compound classifiers tuned to the considered problem of liver fibrosis. They are presented in Tab.2.

Table 2. Selected parameter values for classification algorithms.

Model	Parameter #1	Parameter #2
Bagging	$B = 20$	$n = 0.7$
Boosting	$I = 20$	-
Random Forest	$N = 80$	$D = 7$
Random Subspace	$R = 10$	$S = 0.6$

These setting are used in the experimental investigations for comparison between classifiers.

3.4. EXPERIMENTAL RESULTS

The results of the experiment are presented in Tab. 3. Subscript numbers below the classification accuracy indicates the indexes of classifiers from which the tested method was significantly better.

Table 3. Performances of the examined machine learning algorithms.

Bagging[1]	Boosting[2]	Random Forest[3]	Random Subspace[4]
78.94% -	80.12% -	83.11% 1,2	84.65% 1,2,3

Four examined methods delivered varied performance on the liver fibrosis dataset. Best results were achieved by the Random Subspace method coupled with SVM classifier. It was statistically better from all other tested approaches. Random Forest ended on the second position with small but statistically significant difference form the former method. Bagging and Boosting have produced the models with the lowest quality – additionally there were no statistical difference between their outputs.

4. DISCUSSION AND CONCLUSION

The presented paper shows that, despite some problems (like the fact that it is not easy to get blood test results of patients with diagnosed chronic hepatitis C, infected with genotype1 HCV that have no other medical conditions and are not under any medical therapy, or that blood test results which were available for research were inconsistent, i.e. some patients have one set of blood tests, while other patients have a set of other blood tests) it is possible to reach similar or even lower error level than commercial tests [17]. It is also worth to mention that liver biopsy result, according to the other research, is also only a prediction with classification error varying from 35% up to 45% [18], depending on the sample size and count. Our method coupled with the state-of-the-art MCS classifiers outperformed significantly these commercial tests.

Presented methods have better accuracy than our previous method [19] based on multiple linear regression (79%), but they operate on a full set of features, without determining their importance. In future we would like to concentrate on the problem of feature selection to reduce the complexity of our model and to identify the most important prognostic factors for this task. Additionally we intend to propose a complex and customized MCS dedicated to this problem.

BIBLIOGRAPHY

[1] CASTÉRA L., Non invasive assessment of liver fibrosis in chronic hepatitis C. Liver International , Vol. 5, No. 2, 2011, pp. 625-634.

[2] BioPredictive, http://www.biopredictive.com/intl/physician/fibrotest-for-hcv/view?set_language=pl.

[3] Simens Medical, http://www.medical.siemens.com/webapp/wcs/stores/servlet/PSGenericDisplay~q_catalogId~e_-111~a_langId~e_-111~a_pageId~e_103713~a_storeId~e_10001.htm.

[4] EOM S, KIM E. A survey of decision support system applications (1995-2001). *J Oper Res Soc* 2006;57(11):1264-78.

[5] GARG AX, ADHIKARI NKJ, MCDONALD H, ROSAS-ARELLANO MP, DEVEREAUX PJ, BEYENE J, SAM J, HAYNES RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *J Am Med Assoc* 2005;293(10):1223-38.

[6] ALPAYDIN E., *Introduction to Machine Learning. Second edition*, The MIT Press, Cambridge, MA, USA, London, UK, 2010.

[7] WOŹNIAK M. Combining classifiers - concept and applications. *Journal of Medical Informatics & Technologies.* 2010, Vol. 15, pp. 19-27.

[8] KRAWCZYK B., WOŹNIAK M. Analysis of Diversity Assurance Methods for Combined Classifiers. *Image Processing and Communications Challenges 4*, pp. 179-186.

[9] KRAWCZYK B., WOŹNIAK M. Hypertension diagnosis using compound pattern recognition methods. *Journal of Medical Informatics & Technologies.* 2011, Vol. 18, pp. 41-50.

[10] KRAWCZYK B. Classifier committee based on feature selection method for obstructive nephropathy diagnosis; *Semantic Methods for Knowledge Management and Communication* 2011, Katarzyniak R. et al. (Eds.), Springer, Studies in Computational Intelligence 381:115-125.

[11] ALPAYDIN, E.: Combined 5 x 2 cv F Test for Comparing Supervised Classification Learning Algorithms, Neural Computation, 11:1885-1892,1998.

[12] BREIMAN L.: *Bagging predictors*, Technical Report 421, Department of Statistics, University of California, Berkeley, 1994.

[13] BREIMAN L.: *Random forests*, Machine Learning, Volume 45:5–32, 2001.

[14] BRYLL, R.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition 20 (6): 1291–1302, 2003.

[15] QUINLAN, J.R., Induction of decision trees. *Machine Learning*, 1(1), pp. 81-106, 1986.

[16] SCHAPIRE R. E., The boosting approach to machine learning: An overview. *Proc. Of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, 2001.

[17] Enhanced Liver Fibrosis Test (ELF) for evaluating liver fibrosis, The National Horizon Scanning Centre, Department of Public Health and Epidemiology, University of Birmingham, 2008.

[18] BEDOSSA P., DARGÈRE D., PARADIS V., Sampling variability of liver fibrosis in chronic hepatitis C, Hepatology 2003, Vol. 38, pp. 1449–1457.

[19] ORCZYK T., PAŁYS M., PORWIK P., MUSIALIK J., BŁOŃSKA-FAJFROWSKA B., Simple and non-invasive liver fibrosis stage prediction method, *Journal of Medical Informatics & Technologies*, 2011, Vol. 17, pp. 227-232.