International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:2, No:8, 2008

Artificial Intelligence Support for Interferon Treatment Decision in Chronic Hepatitis B

Alexandru George Floares

***Abstract*—**Chronic hepatitis B can evolve to cirrhosis and liver cancer. Interferon is the only effective treatment, for carefully selected patients, but it is very expensive. Some of the selection criteria are based on liver biopsy, an invasive, costly and painful medical procedure. Therefore, developing efficient non-invasive selection systems, could be in the patients benefit and also save money. We investigated the possibility to create intelligent systems to assist the Interferon therapeutical decision, mainly by predicting with acceptable accuracy the results of the biopsy. We used a knowledge discovery in integrated medical data - imaging, clinical, and laboratory data. The resulted intelligent systems, tested on 500 patients with chronic hepatitis B, based on C5.0 decision trees and boosting, predict with 100% accuracy the results of the liver biopsy. Also, by integrating the other patients selection criteria, they offer a non-invasive support for the correct Interferon therapeutic decision. To our best knowledge, these decision systems outperformed all similar systems published in the literature, and offer a realistic opportunity to replace liver biopsy in this medical context.

***Keywords*—**Interferon, chronic hepatitis B, intelligent virtual biopsy.

I. Introduction and Biomedical Background

HEpatitis B is one of the major diseases of mankind and is a serious global public health problem. Of the two billion people who have been infected with the hepatitis B virus (HBV), more than 350 million have chronic infections. These persons are at high risk of death from cirrhosis of the liver and liver cancer, diseases that kill about one million persons each year.

Activity (necroinflammation) and fibrosis are two major histologic features of chronic hepatitis B included in the most used scoring systems, METAVIR and Ishak. These systems assess histologic lesions in chronic hepatitis B using two separate scores, one for necroinflammatory grade - METAVIR A (A for activity) or Ishak NI (NI for necroinflammatory) and another for the stage of fibrosis (F) - METAVIR F or Ishak F.

Liver biopsy is the gold standard for grading the severity of disease and staging the degree of fibrosis and the grade of necroinflammation. permanent architectural damage. Liver biopsy is invasive and usually painful; complications severe enough to require hospitalization can occur in approximately 4% of patients [1]. In a review of over 68,000 patients recovering from liver biopsy, 96% experienced adverse symptoms during the first 24 hours of recovery. Hemorrhage was the most common symptom, but infections also occurred. Side effects of the biopsies included pain, tenderness, internal bleeding, pneumothorax, and rarely, death [2].

Alexandru Floares is with: 1) SAIA - Solutions of Artificial Intelligence Applications, Cluj-Napoca, Romania, email: saia4ai@yahoo.com, 2) IOCN - Institute of Oncology Cluj-Napoca, Cluj-Napoca, Romania, email: alexandru.floares@gmail.com

Transient elastography (FibroScan®) is an ultrasound imaging technique used to quantify hepatic fibrosis in a totally non-invasive and painless manner. It performs well in identifying severe fibrosis or cirrhosis, but is less accurate in identifying lower degrees of fibrosis.

Chronic hepatitis B in some patients is treated with drugs called interferon or lamivudine, which can help some patients. However, interferon or lamivudine therapy costs thousands of dollars, and the patients' selection criteria include fibrosis and necroinflammation assessed by liver biopsy, an invasive medical procedure.

As an example, the Romanian Ministry of Health's criteria, for selecting the patients with chronic hepatitis B, who will benefit from Interferon treatment, are:

1) Chronic infection with HBV:
   a) the hepatitis B surface antigen (HBsAg) is present for at least 6 months, *or*
   b) the hepatitis B e antigen (HBeAg) is present for at least 10 weeks.
2) The cytolytic syndrome: the transaminases level is increased or normal.
3) Pathology (*biopsy*): the Ishak NI ≥ 4 and Ishak F ≤ 3.
4) The virus is replicating, with the following possible situations:
   a) HBsAg is present, and HBeAg is present, and DNA HBV $> 10^5$ copies/mililiter,
   b) HBsAg is present, and HBeAg is absent, antibodies against HBsAg (anti-Hbs) are present, and DNA-HBV $> 10^5$ copies/mililiter (mutant viruse infection),
   c) anti-HBs are present, and DNA-HBV $> 10^5$ copies/mililiter.

Developing an efficient selection system, based on non-invasive medical procedures, is important for the patients' benefit and could also save money. To this goal, it is important to investigate if it is possible:

- To extract and integrate information from various (non-invasive) sources, e.g. imaging, clinical, and laboratory data, to build systems capable to predict the biopsy results - fibrosis stage and necroinflammation grade - with an acceptable 90%-100% accuracy.
- To integrate these predictions with other selection criteria, in a system capable to support the correct interferon treatment decision.

- To quantify the end results of the Interferon treatment and use them in a system capable to identify the important selection criteria and their cutoff values.

As it will be shown, the extraction and integration of information from various data sources is indeed possible, using a knowledge discovery in data or data mining approach, based on computational intelligence tools, and the prediction accuracy of the resulted intelligent systems could even reach 100%. Also, the intelligent system for Interferon treatment decision support can be built and is effective.

In this way, an important medical protocol or workflow for patients management - Interferon treatment decision in chronic hepatitis B - is integrated with intelligent agents or modules. By letting this agents to learn the prediction of the end results of the Interferon treatment, they could reveal the biomedical variables correlated to various degree of treatment response, and also their cutoff values, delimiting the response patients'subgroups (work in progress). We developed a similar intelligent system for Interferon treatment decision support in chronic hepatic C (Floares, 2008 - submitted to "Intelligent Data Analysis in Biomedicine and Pharmacology, November, 7th, 2008, Washington, DC, USA).

By far the most difficult problem of these investigations consists in predicting the results of liver biopsy [3], [4], (Floares, 2008 - accepted at "Intelligent Systems for Medical Decisions Support", CIBB 2008, 3-4 October, 2008, Vietri sul Mare, Salerno, Italy). We used several non-invasive approaches - routine laboratory tests and basic ultrasonographic features - with and without liver stiffness measurement by transient elastography (FibroScan®), to build intelligent systems for staging liver fibrosis and the grade of necroinflammation in chronic hepatitis B.

To the best of our knowledge, this is the first intelligent system, to support Interferon treatment decision in chronic hepatitis B, developed by integrating intelligent agents (modules) in the medical workflow, capable to predict the fibrosis stage and necroinflammatory degree with the highest published accuracy (100%). The fact that we reached similar results for hepatitis C and also in a different but similar problem - predicting prostate biopsy results in prostate cancer to support surgical treatment decisions (Floares et al., 2008 - accepted at Workshop on Computers in Medical Diagnoses, IEEE International Conference on Intelligent Computer Communication and Processing, August 28 - 30, 2008, Cluj-Napoca, Romania), corroborate our believe that this approach can become a standard one.

II. Intelligent Systems for Interferon Treatment Decision Support in Hepatitis B

A. Data Integration and Preprocessing

One of the key aspect of intelligent data analysis is in our opinion the integrating various medical data sources: clinical, imaging and lab data. Our experiments showed that isolated data sources do not usually contain enough information for building accurate intelligent systems. The main problems we found, in mining the medical data bases, were the small number of patients relative to the number of features, and the large extent of missing data. However, comparing to similar medical studies our dataset was quite large, with hundreds of patients.

The order of the pre-processing steps is important. Due to the above mentioned problems, one should avoid as much as possible the elimination of patients form the analysis during data pre-processing, and try to eliminate uninformative features first. If feature selection is performed first, even without using sophisticated methods for missing data imputation, the number of eliminated cases is smaller. For a recent exhaustive collection of feature selection methods see [5].

Feature selections was performed in three steps:

1) Cleaning. Unimportant and problematic features and patients were removed.
2) Ranking. The remaining features were sorted and ranks were assigned based on importance.
3) Selecting. The subset of features to use in subsequent models was identified.

In data cleaning, we always removed or excluded from the analysis the following variables:

- variables that have all missing values,
- variables that have all constant values,
- variables that represent case ID.

The following cases were always removed:

- cases that have missing target values,
- cases that have missing values in all its features.

The following variables were also removed:

1) Variables that have more than 70% missing values.
2) Categorical variables that have a single category counting for more than 90% cases.
3) Continuous variables that have very small standard deviation (almost constants).
4) Continuous variables that have a coefficient of variation CV < 0.1 (CV = standard deviation/mean).
5) Categorical variables that have a number of categories greater than 95% of the cases.

For ranking the features, "predictor" an important step of feature selection, also important for understanding the biomedical problem, we used a simple but effective method which considers one feature at a time, to see how well each feature alone predicts the target variable. For each feature, the value of its importance is calculated as (1 - p), where p is the p value of the corresponding statistical test of association between the candidate feature and the target variable. The target variable was categorical with more than two categories for all investigated problems, and the features were mixed, continuous and categorical.

For categorical variables, the p value was based on Pearson's Chi-square altfel, fara Pearson test of independence between X, the feature under consideration with I categories, and Y target variable with J categories. The Chi-square test involves the difference between the observed and expected frequencies. Under the null hypothesis of independence, the expected frequencies are estimated by $\widehat{N} = N_i \cdot N_j/N$. Under the null hypothesis, Pearson's chi-square converges asymptotically to a chi-squared distribution χ^2_d with degree of freedom $d = (I-1)(J-1)$, and the p value is equal with the probability

that $\chi_d^2 > X^2$, where $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(N_{ij} - \widehat{N}_{ij})^2/\widehat{N}_{ij}$. The categorical variables were sorted first by p value in the ascending order, and if ties occurred they were sorted by chi-squared in descending order. If ties still occurred, they were sorted by degree of freedom d in ascending order.

For the continuous variables, p values based on the F statistic are used. For each continuous variable a one-way ANOVA F test is performed to see if all the different classes of Y have the same mean as X. The p value based on F statistic is calculated as the probability that $F(J-1, N-J) > F$, where $F(J-1, N-J)$ is a random variable that follows and F distribution with degrees of freedom $J-1$ and $N-J$, and

$$F = \frac{\sum_{j=1}^{J} N_j(\overline{x} - \overline{\overline{x}})^2/(J-1)}{\sum_{j=1}^{J}(N_j - 1s_j^2/(N-J)} \quad (1)$$

If the denominator for a feature was zero, the p value of that feature was set to zero. The features were ranked first by sorting them by p value in ascending order, and if ties occurred, they were sorted by F in descending order. If ties still occurred, they were sorted by N in descending order.

Based on the features' importance $(1-p)$, with p calculated as explained above, we ranked and grouped features in three categories:

1) important features, with $(1-p)$ between 0.95 and 1,
2) moderately important features, with $(1-p)$ between 0.90 and 0.95, and
3) unimportant features, with $(1-p)$ less than 0.90.

Some of the categorical features and also the target categorical variable have imbalanced distributions, and this can cause some modeling algorithms to perform poorly. We tested the influence on the prediction accuracy of several methods for dealing with imbalanced data (see [6] for a recent comprehensive review). Because the number of patients is small relative to the number of features, a very common situation in biomedical data bases, we only used oversampling methods and not undersampling methods. We also found that simple techniques such as random oversampling perform better than the "intelligent" sampling techniques. An exhaustive comparison of these methods ca be found in [6].

B. Intelligent Systems as Ensemble of Classifiers

For modeling, we first tested the fibrosis and necroinflammation prediction accuracy of various methods:

1) Neural Networks
2) C5.0 decision trees
3) Classification and Regression Trees
4) Support Vector Machines, expresia consacrata
5) Bayesian Networks.

Because physicians prefer white-box algorithms, we have chosen C5.0 decision trees, the last and improved version of the C4.5 algorithm [7], with 10-fold cross-validation.

Breiman's bagging [8] and Freund and Schapire's boosting [9] are examples of methods for improving the predictive power of classifier learning systems. Both form a set of classifiers that are combined by voting, bagging by generating replicated bootstrap samples of the data, and boosting by adjusting the weights of training cases. While both approaches improve predictive accuracy, boosting showed sometimes greater benefit. Unfortunately, boosting doesn't always help, and when the training cases are noisy, boosting can actually reduce classification accuracy. Naturally, it took longer to produce boosted classifiers, but the results often justified the additional computation. Boosting should always be tried when peak predictive accuracy is required, especially when unboosted classifiers are already quite accurate.

Boosting combines many low-accuracy classifiers (weak learners) to create a high-accuracy classifier (strong learner). We used a boosting version called *AdaBoost*, with reweighting; AdaBoost comes from ADAptive BOOSTing [9].

Suppose we are given the training set data (X_1, F_1),..., (X_n, F_n), where n is the number of patients, the input $X_i \in \Re^p$ represents the p selected features in the preprocessing steps (image, laboratory data, etc.), and the categorical output F_i is the fibrosis stage (things are similar if necroinflammation is the output) according to one of the two scoring systems Metavir F and Ishak F, and assumes values in a finite set {F0, F1,..., Fk}, were k = 5 for Metavir F (from Metavir F0 to Metavir F4) and k = 7 for Ishak F (from Ishak F0 to Ishak F6). The goal is to find a classification rule $F(\mathbf{X})$ from the training data, so that given a new patient's input vector $\mathbf{X}$, we can assign it a fibrosis degree F from {F0, F1,..., Fk} according to the corresponding scoring systems.

Moreover, we want to find the best possible classification rule achieving the lowest misclassification error rate. We assumed that the patients' training data are independently and identically distributed samples from an unknown distribution. Starting with the unweighted training sample, the AdaBoost builds a classifier which can be a neural network, decision tree, etc., that produces class labels - fibrosis degree. If a training data point (patient) is misclassified, the weight of that training patient is increased (boosted). A second classifier is built using the new weights, which are now different. Again, misclassified training patients have their weights boosted and the procedure is repeated. Usually, one may build hundred of classifiers this way. A score is assigned to each classifier, and the final classifier is defined as the linear combination of the classifiers from each stage.

With the above notations, and noting with I an indicator function, a compact description of the AdaBoost algorithm used is the following:

1) Initialize the patient weights $\omega_i = 1/n, i = 1, 2, \ldots, n$.
2) For $m = 1$ to M:
   a) Fit a classifier $F^{(m)}(\mathbf{x})$ to training patients using weights ω_i.
   b) Compute

$$err^m = \sum_{i=1}^{n} \omega_i I(F_i \neq F^{(m)}(\mathbf{X}_i))/\sum_{i=1}^{n}\omega_i. \quad (2)$$

   c) Compute

$$\alpha^{(m)} = \log\frac{1 - err^{(m)}}{err^{(m)}}. \quad (3)$$

d) Set

$$\omega_i \leftarrow \omega_i \cdot \exp(\alpha^{(m)} \cdot I(F_i \neq F^{(m)}(\mathbf{X}_i)) \quad i = 1, 2, \ldots, n. \tag{4}$$

e) Re-normalize ω_i.

3) Output

$$F(\mathbf{X}) = \arg\max \sum_{m=1}^{M} \alpha^{(m)} \cdot I(F^{(m)}(\mathbf{X}) = k). \tag{5}$$

For two-class classification problems AdaBoost could be very successful in producing accurate classifiers. The multi-class classification is more involved, and some technical criteria must be satisfied and experiments need to be done. While fibrosis stage and necroinflammation degree prediction are multi-class classification problems, the Interferon treatment decision is a binary one.

Therefore, it will be advantageous to take into account the largely accepted cutoff values for (see also section I):

- fibrosis stage, e.g., Ishak F ≤ 3, and to build an intelligent system capable to predict if the fibrosis stage is either Ishak F ≤ 3 or Ishak F > 3,
- necroinflammatory degree, e.g., Ishak NI ≥ 4, and to build an intelligent system capable to predict if the necroinflammatory degree is either Ishak NI ≥ 4 or Ishak $F < 4$.

The intelligent system for the Interferon treatment decision support takes as inputs the outputs of the above systems. The decision is again binary, recommending or not the Interferon treatment. For the positive decision a series of other criteria, presented in section I, must be satisfied. The proposed methodology is by no means restricted to the Romanian Ministry of Health's criteria, or even to this problem. On the contrary, we believe that this is a rather general methodology for building intelligent systems for medical decisions support. The final intelligent systems are the result of a more detailed data mining predictive modeling strategy which is patented now, consisting mainly in:

- Extracting and integration information from various medical data sources, after a laborious preprocessing:
  - cleaning features and patients,
  - various treating of missing data,
  - ranking features,
  - selecting features,
  - balancing data.
- Testing various classifiers or predictive modeling algorithms.
- Testing various methods of combining classifiers.

C. Intelligent Virtual Biopsy and Intelligent Scoring Systems

Replacing painful, invasive, and/or costly procedures with intelligent systems, taking as inputs integrated data from non-invasive, usual, or cheap medical procedures, techniques and tests, and producing as output 90-100% similar results with the replaced techniques, is an important medical goal. We outline some general ideas, terms and concepts to characterize this new exciting enterprize.

The central new concept is *Intelligent Virtual Biopsy* (IVB), which designates an intelligent system capable to predict, with an acceptable accuracy (e.g., 90-100%), the results given by a pathologist, examining the tissue samples from real biopsies, expressed as scores of a largely accepted scoring system. As an alternative term we suggest *intelligent biopsy* or *i-biopsy*, were the term intelligent indicates that the system is based on artificial intelligence. To predict the pathologist's scores, the intelligent systems take as inputs and integrate various non-invasive biomedical data.

Also, to distinguish between the scores of the scoring systems of the real biopsy, and their counterparts predicted by the i-biopsy, we proposed the general term of *i-scores* belonging to *i-scoring systems*. In the gastroenterological context of these investigations, we have the following correspondences:

1) Liver intelligent virtual biopsy (IVB), or *liver i-biopsy* is the intelligent system corresponding to the real liver biopsy.
2) The *i-Metavir F* or *A* and *i-Ishak F* or *NI* correspond to the two liver fibrosis or necroinflammation scoring systems Metavir F or A, and Ishak F or NI respectively.
3) The *i-scores* are the values predicted by the intelligent systems for the fibrosis scores.

From a biomedical point of view, the most important general characteristics of the i-scores are exemplified for the Metavir F and Ishak F scores:

1) I-Metavir F or A and i-Ishak F or NI scores have exactly the same biomedical meaning as Metavir-F or A and Ishak-F or NI, scoring the same pathological features.
2) I-Metavir F or A and i-Ishak F or NI scores are obtained in a non-invasive and painless manner, as opposed to Metavir-F and Ishak-F.
3) The estimation of i-Metavir F or A and i-Ishak F or NI does not have the risks related to Metavir-F or A and Ishak-F or NI estimation via biopsy.

III. Results

We have built the following modules, components of the intelligent system for Interferon treatment decision support:

1) Module for liver fibrosis prediction,
   a) according to Metavir F scoring system
      i) with liver stiffness (FibroScan®),
      ii) without liver stiffness (FibroScan®)
   b) according to Ishak F scoring system
      i) with liver stiffness (FibroScan®),
      ii) without liver stiffness (FibroScan®).
2) Module for the grade of necroinflammation (activity) prediction, according to Ishak NI scoring systems

The fibrosis prediction module was first built using a dataset of 381 chronic hepatitis C patients and the METAVIR scoring system [3]. Now, it was tested on 700 chronic hepatitis C patients and the fibrosis is predicted according to METAVIR F or Ishak F scoring system. As we previously mentioned, in the interferon treatment decision system we used the binary version of the fibrosis and necroinflammation classifiers. For the version with liver stiffness, at the end of

the preprocessing stage, besides liver stiffness, the relevant features for predicting liver fibrosis, according to Metavir scoring system, were: age, aspartate aminotransferase, gamma-glutamyl-transpeptidase, cholesterol, triglycerides, thickness of the gallbladder wall, spleen area and perimeter, left lobe and caudate lobe diameter, liver homogeneity, posterior attenuation of the ultrasound, liver capsule regularity, spleen longitudinal diameter, the maximum subcutaneous fat, perirenal fat. Combining all these features, the intelligent system was able to predict each fibrosis stage with 100% accuracy.

In the mean time we have tried to reduce the number of features to at most ten, without sacrificing the accuracy, because some of our investigations showed that this is possible [4]; the results are very encouraging (manuscript in preparation).

We also wanted to investigate if it is possible to build intelligent systems, capable to predict fibrosis scores according to Metavir F and Ishak F scoring system, without using apparently a key source of information - the liver stiffness measured with FibroScan®. Such intelligent systems could be useful to those gastroenterology clinics having ultrasound equipment but not the expensive FibroScan®. After feature selection, the relevant features for Metavir F prediction without Fibroscan®were: cholesterol, caudate lobe diameter, thickness of the abdominal aortic wall, aspartate aminotransferase, preperitoneal fat thickness, splenic vein diameter, time averaged maximum velocity in hepatic artery, time averaged mean velocity in hepatic artery, flow acceleration in hepatic artery, hepatic artery peak systolic velocity. Combining these 10 attributes, the boosted C5.0 decision trees were able to predict each fibrosis stage, according to Metavir F scoring system, with 100% accuracy, even without liver elastography.

The relevant features for predicting liver fibrosis according to Ishak F scoring system were: caudate lobe diameter, left lobe diameter, liver capsule regularity, liver homogeneity, thickness of the abdominal aortic wall, steatosis (ultrasonographic), cholesterol, sideremia, and liver stiffness. The boosted C5.0 decision trees were able to predict each Ishak fibrosis stage with 100% accuracy.

We also built a module for predicting the grade of necroinflammation according only to Ishak NI scoring systems, because Metavir A scoring system is less used. The selected features were: aspartate aminotransferase (ASAT), alanine aminotransferase (ALAT), left liver lobe diameter, hepatic artery acceleration time, hepatic vein Doppler waveform, liver capsule regularity, posterior attenuation of the ultrasound, liver parenchymal echogenity, and hepatic arterial pulsatility index. Combining these 9 attributes, the boosted C5.0 decision trees were able to predict each fibrosis stage, according to Metavir F scoring system, with 100% accuracy, even without liver elastography.

All these models have 100% accuracy, and at the moment of writing this paper, the intelligent systems were tested on 528 patients with chronic hepatitis C.

IV. Discussions

The reasons for the relative disproportion between the number of patients and the number of features is that, at the beginning of these investigations, our multi-disciplinary team tried to define a large number of potentially important features. We intended to use a *data-driven* approach avoiding as much as possible restrictive a priori assumptions. Usually, this opens the door for potential surprises, e.g., previously unknown and unexpected relationships between fibrosis and various other biomedical features. There were some unexpected findings (results not shown) but they need further investigations.

While the data related problems are not so serious as in mining genomics or proteomics data, the fact that the difficulties are not so evident could be a trap. This apparent simplicity was responsible for some initially poor results, but a careful pre-processing increased the accuracy of the predictions with 20% to 25%.

A short comment about the meaning of 100% diagnostic accuracy seems to be necessary, because it confused many physicians who say that 100% accuracy is not possible in medicine. The meanings will be made clear more easy by means of examples. We have proposed intelligent systems predicting the fibrosis scores resulted from liver biopsy with 100% accuracy. Usually, an invasive liver biopsy is performed and a pathologist analyzes the tissue samples and formulates the diagnostic, expressed as a fibrosis score. The pathologist may have access to other patient's data, but usually these are not necessary for the pathological diagnostic. Moreover, in some studies it is required that the pathologist knows nothing about the patient. His or her diagnostic can be correct or wrong for many reasons, which we do not intend to analyze here. On the contrary, for the intelligent system some of the clinical, imaging and lab data of the patient are essential, because they were somehow incorporated in the system. They were used like features to train the system, and they are required for a new, unseen patient, because the i-biopsy is in fact a relationship between these inputs and the fibrosis scores.

Intelligent systems do not deal directly with diagnostic *correctness*, but with diagnostic prediction accuracy. In other words, the intelligent system will predict, in a non-invasive and painless way, and without the risks of the biopsy, a diagnostic which is 100% identic with the pathologist diagnostic, if the biopsy is performed. While the accuracy and the correctness of the diagnostic are related in a subtle way, they are different concepts. An intelligent system will use the information content of the non-invasive investigations to predict the pathologist diagnostic, without the biopsy. The correctness of the diagnostic is a different matter, despite the fact that a good accuracy is almost sure related with a correct diagnoses, but we will not discuss this subject.

The accuracy of the diagnosis, as well as other performance measures like the area under the receiver operating characteristic (AUROC), for a binary classifier system [10], are useful for intelligent systems comparison. From the point of view of accuracy, one of the most important medical criterions, to our best knowledge the proposed liver intelligent virtual biopsy or i-biopsy system outperformed the most popular and accurate system, FibroTest [11] commercialized by Biopredictive company. The liver i-biopsy presented in this paper is based on a five classes classifier, more difficult to build than binary classifiers; we also build binary classifiers as decision

trees with 100% accuracy and mathematical models (work in progress, results not shown). Despite the fact that AUROC is only for binary classifiers, loosely speaking a 100% accuracy n classes classifier is equivalent with n binary classifiers with AUROC = 1 (maximal). In [11], a total of 30 studies were included which pooled 6,378 subjects with both FibroTest and biopsy (3,501 chronic hepatitis C). The mean standardized AUROC was 0.85 (0.82-0.87).

Moreover, in some circumstances the result of the liver IVB could be superior to that of real biopsy. When building the intelligent system, the results of the potentially erroneous biopsies, which are not fulfilling some technical requirements, were eliminated from the data set. Thus, the IVB predicted results correspond only to the results of the correctly performed biopsies, while some of the real biopsy results are wrong, because they were not correctly performed. Due to the invasive and unpleasant nature of the biopsy, is very improbable that a patient will accept a technically incorrect biopsy to be repeated. Unlike real biopsy, IVB can be used to evaluate fibrosis evolution, which is of interest in various biomedical and pharmaceutical studies, because, being non-invasive, painless and without any risk, can be repeated as many time as needed. Also, in the early stages of liver diseases, often the symptoms are not really harmful for the patient, but the treatment is more effective then in more advanced fibrosis stages. The physician will hesitate to indicate an invasive, painful and risky liver biopsy, and the patients is not so worried about his or her disease to accept the biopsy. However, IVB can be performed and an early start of the treatment could be much more effective. Moreover, we have obtained high accuracy results for other liver diseases, like chronic hepatitis B and steatohepatitis, for other biopsy findings, like necroinflammatory activity and steatosis (results not shown), and also for prostate biopsy in prostate cancer. These corroborate our believe that this approach can become a standard one.

REFERENCES

[1] A. Lindor, "The role of ultrasonography and automatic-needle biopsy in outpatient percutaneous liver biopsy," *Hepatology*, vol. 23, pp. 1079–1083, 1996.

[2] A. Tobkes and H. J. Nord, "Liver biopsy: Review of methodology and complications.," *Digestive Disorders*, vol. 13, pp. 267–274, 1995.

[3] A. G. Floares, M. Lupsor, H. Stefanescu, Z. Sparchez, A. Serban, T. Suteu, and R. Badea, "Toward intelligent virtual biopsy: Using artificial intelligence to predict fibrosis stage in chronic hepatitis c patients without biopsy," *Journal of Hepatology*, vol. 48, no. 2, 2008.

[4] A. Floares, M. Lupsor, H. Stefanescu, Z. Sparchez, R. Badea, and Romania, "Intelligent virtual biopsy can predict fibrosis stage in chronic hepatitis c, combining ultrasonographic and laboratory parameters, with 100% accuracy," *Proceedings of The XXth Congress of European Federation of Societies for Ultrasound in Medicine and Biology*, 2008.

[5] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing, Springer, August 2006.

[6] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24 th International Conference on Machine Learning*, (Corvallis, OR), 2007.

[7] J. Quinlan, *C4.5 : Programs for Machine Learning*. Morgan Kaufmann, 1993.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[9] Y. Freund and R. E. Schapire, "A decisiontheoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[10] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers. technical report," tech. rep., Palo Alto, USA: HP Laboratories, 2004.

[11] T. Poynard, R. Morra, P. Halfon, L. Castera, V. Ratziu, F. Imbert-Bismut, S. Naveau, D. Thabut, D. Lebrec, F. Zoulim, M. Bourliere, P. Cacoub, D. Messous, M. Muntenau, and V. de Ledinghen, "Meta-analyses of fibrotest diagnostic value in chronic liver disease," *BMC Gastroenterology*, vol. 7, no. 40, 2007.