Draft. Full paper at: http://dx.doi.org/10.1109/ICCP.2011.6047870

Texture Analysis as a Noninvasive Tool for Fibrosis Assessment in Chronic Hepatitis C. Influence of Expert Dependent Variability over the Performance of Texture Analysis

Cristian Vicas[1], Monica Lupsor[2], Mihai Socaciu[2], Radu Badea[2], Sergiu Nedevschi[1]
[1)] Computer Science dept., Technical University Cluj Napoca, Romania
[2)] Regional Institute of Gastroentherology and Hepatology Cluj Napoca, Romania
cristian.vicas@cs.utcluj.ro

***Abstract*—Texture analysis is viewed as a method to enhance the diagnosis power of classical B-mode ultrasound image. Present study aims to evaluate the dependence between the human expert and the performance of such a texture analysis system in predicting the cirrhosis in chronic hepatitis C patients. 125 consecutive chronic hepatitis C patients were included in this study. All the patients had positive HCV-RNA in serum and had undergone percutaneous liver biopsy for disease staging using Metavir score. Ultrasound images were acquired from each patient and 4 experts established regions of interest. Textural analysis software generated 234 features from each region of interest. Relevant textural features were identified and a classification schema was evaluated. Texture analysis can discriminate between F0 and F4 fibrosis stages (AUROC=0.64). The performance of this approach depends highly on the human expert that establishes the regions of interest ($p<0.05$). The relevant textural features were identified and it was shown that the detection performance didn't depend on the particular feature selection ($p=0.8$). In classical form met in literature non invasive diagnosis through texture analysis has limited utility in clinical practice because of the user variability introduced by the expert who establishes the regions of interest.**

Keywords non invasive diagnosis; texture analysis; user variability; liver fibrosis

I. INTRODUCTION

Non invasive detection and staging of liver fibrosis have received more and more attention in scientific literature. One approach involves simple B-mode ultrasound in conjunction with textural analysis. The main assumption of the textural analysis approach is that fibrosis alterations at liver lobule level can induce significant changes in the speckle pattern of the ultrasound image [1]. Even if these alterations are not visible with naked eye, a texture analysis system can detect and learn these alterations. In image processing literature, textural analysis is viewed as a method to enhance the diagnosis power of B-mode ultrasound by providing the physician with new information. This data can be otherwise inferred only by invasive methods.

The methodology presented in most of the papers [1-9] approaching textural analysis on B-mode ultrasound follows four general steps. First, a physician acquires a liver ultrasound image. Then, on the ultrasound image, another physician (or the same) establishes a rectangular region of interest (ROI). In the third step several textural algorithms produce a feature vector. This vector is labeled according to biopsy findings. The fourth step implies the training of a classification schema. The resulting classifier can be used to predict fibrosis stages to unknown ultrasound images.In the first two steps there is a human expert that introduces an operator dependent variability.

This paper addresses the user variability introduced by the second step, the establishment of the ROI. To our knowledge this problem has not been addressed before. We included almost all the textural algorithms proposed in the literature as means of detecting liver fibrosis stages. We selected only healthy and cirrhosis patients.

Present study aims to evaluate the dependence between the human expert and the performance of the texture analysis system in predicting cirrhosis in chronic hepatitis C patients. In the next chapter we present the textural analysis system and the methodology to evaluate the user variability. In the third chapter we present the experimental results, including patient lots and ultrasound images. The discussions are presented in chapter four and the conclusions to this paper in chapter five.

II. PROPOSED METHODS

A. Regions of interestfor textural analysis

The textural analysis is applied only on a rectangular region of the ultrasound image. This region is called region of interest and is manually established on the ultrasound image by a trained radiologist.

The main objective of this paper is to evaluate the dependence between the quality of the human expert and the performances of texture analysis in fibrosis detection. In order to evaluate this variability one has to employ a number of human experts.

The region of interest (ROI) establishment procedures followed the guidelines presented in literature [1, 10]. The experts were instructed to choose one region of interest for each patient. The ROI had to be placed as close as possible to the vertical axis of the ultrasound image and at 1 cm below the liver capsule. The ROI had to avoid artifacts and anatomical

features like blood vessels, liver capsule, shadowing, etc. The dimensions of the ROI were 64x64 pixels representing an area of 2.62x2.62 cm. In Figure 1 is shown a right lobe ultrasound image with a ROI.

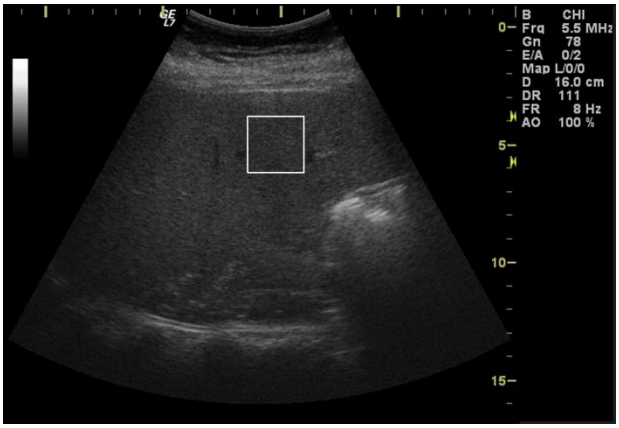

Figure 1. Right lobe ultrasound image. Squared white rectangle represents the region of interest.

In order to evaluate the user variability of the textural system, on the saved images, the ROIs were established by four experts with different skill level. The first two experts are trained radiologists with experience in gastro-intestinal ultrasound investigation. First expert have more than 20 years in ultrasound investigation and the second more than 10 years. The third expert is a radiology intern with 2 years of experience. The fourth expert is a general practitioner trained in ultrasound examination. The order of the patients and the order of the images for a patient were randomized. With this step we tried to avoid the influence of the image order over the performance detection.

The following algorithm was used to ensure independent samples:

Algorithm A1. ROI establishment
Input: **patients** – a set with patients and ultrasound images
Output: **DS** – a list with 20 sets DS_{ij} with regions of interest
For i =1 to 5
    D_i = Perform a randomization on **patients**
    For j=1 to 4
        Sequentially present to expert j dataset D_i
        DS_{ij} = established ROI's

B. Textural analysis

In texture analysis there are two main steps [11]. First step is the computation of several textural attributes that numerically describe the texture (using dedicated algorithms). The second step involves the training and evaluation of a classifier using the previously computed textural features.

Each texture description algorithm has a certain number of parameters that control the feature extraction process. For each algorithm implemented in present study we used the same proposed set of parameters found in corresponding fibrosis detection papers. These algorithms are: first order statistics [4, 12], gray tone difference matrix [11], gray level co-occurrence matrix [1, 4, 12, 13], multiresolution fractal dimension [1], differential box counting [6, 14], morphological fractal dimension estimators [15], Fourier power spectrum [1, 10], Gabor filters [16], Law's energy measures [1], texture edge co-occurrence matrix [6], phase congruency based edge detection [17] and texture feature coding matrix [18].

These 12 algorithms processed the entire ROI and computed 234 features per patient. Each feature vector was labeled with the corresponding histopathological finding as healthy or cirrhotic.

The classification schema employed here was the Support Vector Machines (SVM). The classifier was used with an Gaussian (RBF) kernel. The two main meta-parameters cost C and γ exponent were searched exhaustively using grid search strategy [19]. A value set was proposed for each parameter and the algorithm searched all possible combinations. Each combination was evaluated using 2-fold cross validation (CV) [19]. From this, the best found pair, along with its neighbors was evaluated using 5-fold CV. Again, the best pair was selected. The search schema returned a classifier trained on the entire input dataset. Cost C meta parameter takes values 2^{-3}, 2^{-2}, 2^{-1}… 2^{8}. Exponent γ takes values 2^{-15}, 2^{-14}… 2^{4}. The cost is weighted in order to compensate for imbalanced class distribution. The class with fewer instances will have a greater weight than the more populated class. The feature values were normalized in [0,1] interval prior to classification. Care was taken that the test subset was normalized with the same coefficients as the train set.

The trained classifier can then be used in fibrosis detection. Usually one wants to know the performance of such a classifier on unknown set of instances. Because a new set of instances is expensive or very hard to acquire (especially in medical field) there were developed certain algorithms that estimate the real performance of a classification schema using the available dataset. There are various algorithms available in literature [19]. In present paper the performance estimation was performed using 10 fold stratified CV technique. All available data was split in 10 folds. At each iteration one fold was kept for testing while the other folds were used for training the classifier using the procedure described above. The predictions on test folds were collected into a vector. The whole process was iterated 10 times, for each fold. At each step the algorithm selects a different test fold, such that, after 10 iterations each fold acted as test fold exactly one time. In order to better estimate the average performance the 10 fold CV procedure was iterated 10 times with random fold splitting [20].

Textural analysis generated 234 features for each instance. We used a feature selection method to reduce the number of features to a smaller group by keeping the most significant ones. The feature selection procedure is described below.

From each dataset we extracted 10 instances per class. With the extracted instances we built a separate dataset. In this dataset we performed a univariate logistic regression for each feature and noted the p-value. These p-values were adjusted for multiple testing [21]. We kept only features with false discovery rate lower than 0.2. Another logistic regression model was fitted using these features and using stepwise regression we selected the final features. The goodness of the

fit for the stepwise regression was Akaike's information criterion (AIC) [22].

From each of the initial datasets we retained only features selected by stepwise regression.

These datasets then entered the classification and performance evaluation loops. SVM algorithm allows the computation of *a posteriori* probabilities for each prediction. These probabilities were collected along with the predictions. The performance criterion was Area Under Curve (AUROC) computed on the collected predictions using Mann-Whitney-Wilcoxon U statistic [23].

In short, the methodology is structured as following:

1) The human experts establish the ROIs on the images using **algorithm A1.** There were 20 datasets generated.
2) Each ROI was processed by 12 textural algorithms producing 234 element feature vector. Each vector was labeled with the corresponing Metavir findings.
3) Iterate ten times:
    3.1) Select features using algorithm A2
    3.2) Evaluate a classifier using algorithm A3 and 10 fold CV. Collect predictions.
    3.3) Compute AUROC on collected predictions
    3.4) Repeat from step (3.2) 10 times and report average performance

For each combination of expert, ROI set and feature set we obtained a mean AUROC. There were computed 4x5x10=200 performance estimations.

Algorithm A2. Feature selection
Input: **DS** – 20 sets of labeled feature vectors. Each vector has 234 elements
Output: **featset** – a list of selected features; **DS'** – 20 sets of labeled feature vectors. Each feature vector contains only features listed in **featset**.
S=empty
DS'=empty
For each DS_{ij} in **DS**
    inst = set of randomly selected 20 instances, 10 from each class.
    $DS_{ij}' = DS_{ij} - inst$
    $S = S + inst$
Perform feature selection on *S*. Let **featset** be the set of selected features
For each DS_{ij}'
    Reduce the dataset DS_{ij}' by keeping ony features in **featset**
    DS' = **DS'**+DS_{ij}'

Algorithm A3. Training a SVM instance.
Input: **dataset** – a set of labeled feature vectors.
Output: **ClassInstance** – an instance of trained classifier
For *i* in $2^{-15}, 2^{-14} \dots 2^{4}$
    For *j* in $2^{-3}, 2^{-2}, 2^{-1} \dots 2^{8}$
        $\gamma=i; C=j$
        $perf_{ij}$=CV_evaluate(SVM(γ,*C*), **dataset**,2)
Select (*i*,*j*) that maximizes $perf_{ij}$
For (*i'*,*j'*) in all of neighbors of (*i*,*j*)
    $Perf_{i'j'}$=CV_evaluate(SVM(γ',*C'*), **dataset**,5)
Select (*i'j'*) that maximizes $perf_{i'j'}$
return train_instance(SVM(γ',*C'*),**dataset**)

where CV_evaluate(**alg**(**param**), **data**, **f**) represents a cross validation loop for the algorithm **alg** having metaparameters **param**. The dataset is divided in **f** folds.

The whole performance evaluation process was iterated 10 times. Each step involved a feature selection. We ranked each feature by the number of times it was selected. The texture analysis system was validated using a set of known textures from Brodatz [24] library. Each image was divided in 100 non overlapping regions of interest. Each region has 64x64 pixels area. The textural analysis system was trained to predict the original image from where the region originated. The images were chosen following the guidelines in [11].

C. Statistical Analysis

Two way ANOVA test was used to evaluate the performance variability. The dependent variable was set to be the average AUROC and the independent variables were the expert that established ROIs and the feature set obtained after the feature selection step. Tukey post hoc analysis was used to identify the source of variation when the ANOVA test was statistically relevant.

When the assumption of normal distribution with equal variances could not be met we used Kruskal-Wallis one way analysis of variance. The significance threshold was set to p=0.01.

Textural algorithms were implemented in a custom made software system developed at Technical University of Cluj Napoca, Romania. Classification schema used the libsvm implementation [25] (public domain, ver. 2.89) integrated in Weka framework [19] (public domain, ver. 3.7) . Statistical analysis was performed in R (public domain, ver. 2.10).

III. Experimental Results

A. Validation of the textural analysis system

The texture analysis system was validated using three sets of images. First dataset contained regions from D77, D84, D55, D53 and D24 Brodatz [24] textures. Second dataset consisted of D4 and D84 textures. The third set had regions from D5 and D92. The classification accuracy was 98.9 for the first set, 98.4 for the second set and 97.9 for the third set.

B. Patients

This study was approved by the local Ethical Committee of the University of Medicine and Pharmacy Cluj-Napoca. The patients provided written informed consent before the beginning of the study, in accordance to the principles of the Declaration of Helsinki (revision of Edinburgh, 2000). We prospectively included in this study 125 patients with hepatitis C infection having fibrosis stage 0 or 4 according to Metavir scoring system. The fibrosis stages were determined by liver biopsy. This lot was selected from 1200 patients, prospectively examined in 3rd Medical Clinic, Cluj-Napoca, Romania,

between May 2007 and August 2009. All patients had positive HCV-RNA and underwent percutaneous liver biopsy (LB), in order to stage and grade their condition.

The exclusion criteria were: presence of ascites at clinical or ultrasound examination, co-infection with HBV and/or HIV, other active infectious diseases, and pregnancy.

Alongside the epidemiological data, certain biological parameters were determined on a blood sample taken 12 hours after overnight fasting: alanine aminotransferase (ALT), aspartate aminotransferase (AST), gama-glutamyl transferase (GGT), total cholesterol, triglycerides, total bilirubin and glycemia (Konelab 20i – Thermo Electron Corp., Finland).

C. Histopathological analysis

A liver biopsy specimen was prelevated using the TruCut technique with an 1.8 mm (14G) diameter automatic needle device - Biopty Gun (Bard GMBH, Karlsruhe, Germany). The LB specimens were fixed in formalin and embedded in paraffin. The slides were evaluated by a single expert pathologist unaware of the clinical data. Only biopsy specimens with more than 6 intact portal tracts were eligible for evaluation [26]. The liver fibrosis and necroinflammatory activity were evaluated semi quantitatively according to the Metavir scoring system [27].

Fibrosis was staged on a 0-4 scale as follows: F0 – no fibrosis; F1 – portal fibrosis without septa; F2 - portal fibrosis and few septa; F3 – numerous septa without cirrhosis; F4 – cirrhosis. The necroinflammatory activity was graded as A0 – none; A1 – mild; A2 – moderate; A3 – severe.

In present study only patients having fibrosis stage 0 or 4 were included.

D. Ultrasound examination

Each patient included in this study undergoed an ultrasound examination using a GE Logiq 7 ultrasound machine (General Electric Company, Fairfield, England) with a 5.5 MHz convex phased array probe one day prior to liver biopsy. From each patient were acquired right lobe ultrasound images with liver tissue without blood vessels or other artifacts with a depth setting of 16 cm using the same pre established machine protocol. The acquisition protocol was established in such a way that we obtained a maximum amount of information from underlying tissue and in the same time keeping the noise level down. All post processing settings were set to minimum. The frame rate was kept as high as possible in order to avoid movement artifacts. The time gain compensation curve was set to neutral position. Once the device settings were established they were used to examine all the patients. Captured images were saved in DICOM format on the equipment's local hard drive. They were later transferred and processed on a personal computer.

E. Statistical Results

Clinical and biochemical characteristics of the study patients are summarized in Table 1. The median length of the LB samples was 11.38 mm, and the mean number of portal spaces was 11.6. The fibrosis stage distribution in our patients was as follows: F0 –51 (40.8%), F4 – 74 (59.2%).

Each expert was instructed to select one region of interest for each patient. The process was iterated five times. Expert 1 established in average 121.6 regions (min=121, max=122), expert 2 – 120.8 (115-123), expert 3 – 122 (122-122) and expert 4 – 113 (112-115). There were three patients that had poor quality images and no physician was able to establish a ROI. Two were healthy patients and one cirrhotic. The cost weights for SVM algorithm were set to 1.7 for F0 class and 1 for F4 class.

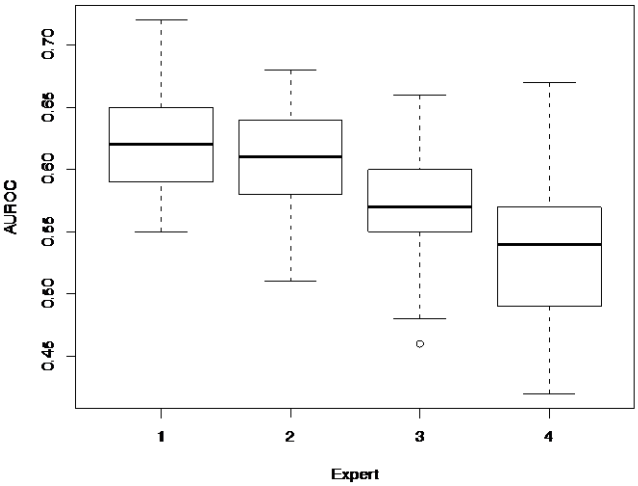

Figure 2. Box plot representing the dependency between the estimated performance and the human expert that established the regions of interest. The top and the bottom of the boxes are the first and third quartiles, respectively. The length of the box thus represents the interquartile range within which 50% of the values were located. The line through the middle of each box represents the median.

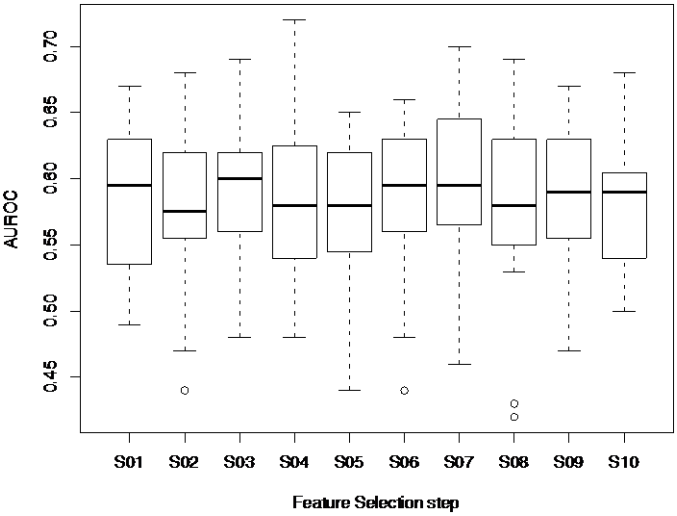

Figure 3. Box plot representing the dependency between the estimated performance and the feature selection process. Each label on the horizontal axis represents a separate feature selection step. The top and the bottom of the boxes are the first and third quartiles, respectively. The length of the box thus represents the interquartile range within which 50% of the values were located. The line through the middle of each box represents the median.

We recorded the mean, standard deviation, minim and maximum AUROC for each of the experts: Expert 1 – 0.64±0.04 (0.55-0.72); Expert 2 – 0.63±0.04 (0.51-0.68) ;

Expert 3 – 0.58±0.04 (0.46-0.66) and Expert 4 – 0.56±0.05 (0.42-0.67).

We investigated the role of feature selection and the user expertise in the performance of the system using two way ANOVA. The only relevant factor was the human expert ($p<0.0001$) as shown in Figure 2. The other factor, feature selection, was not relevant ($p=0.8$). In Figure 3 are shown the corresponding box plots.

Post hoc analysis using Tukey method revealed that the differences between experts are significant ($p<0.001$) with one exception, the difference between the expert 1 and 2.

TABLE I. CHARACTERISTICS OF THE STUDY GROUP.

Characteristics of patients	Entire lot	Patients with fibrosis stage 0	Patients with cirrhosis
	Mean ± SD (interval or %)		
Number	125 (100%)	51 (40.8%)	74 (59.2%)
Sex (male)	50 (40%)	16 (31.4%)	34 (45.9%)
Age (years)	47.45 ± 12.13 (22-77)	53.39 ± 8.93 (33-77)	38.82 ± 10.97 (22-66)
BMI (kg/m^2)	26.41 ± 5.15 (18.56-46.48)	28.29 ± 5.33 (18.83-46.48)	23.9 ± 3.65 (18.56-33.87)
AST (U/I)	58.54 ± 47.67 (12-387)	82 ± 49.57 (23-387)	25.79 ± 13.47 (12-71)
ALT (U/I)	75.68 ± 55.66 (8-270)	102.25 ± 53.94 (21-270)	38.58 ± 31.87 (8-163)
GGT (U/I)	77.83 ± 107.77 (13-993)	105.47 ± 133.33 (27-993)	39.83 ± 28.13 (13-130)
Total bilirubin (mg/dl)	0.88 ± 0.64 (0.27-4.27)	1.09 ± 0.73 (0.4-4.27)	0.59 ± 0.28 (0.27-1.72)
Alkaline phosphatase (U/I)	263.13 ± 188.34 (127-1781)	286.98 ± 215.81 (127-1781)	201.5 ± 45.61 (142-307)
Glucose (mg/dl)	106.73 ± 27.75 (72-266)	113.81 ± 32.78 (72-266)	96.86 ± 13.72 (72-129)
Cholesterol (mg/dl)	195.29 ± 45.8 (97-331)	174.22 ± 36.31 (97-299)	223.83 ± 41.92 (149-331)
Triglycerides (mg/dl)	124.11 ± 57.67 (51-349)	123.85 ± 50.08 (53-316)	124.46 ± 67.16 (51-349)
Platelet count ($10^9/L$)	166.06 ± 70.32 (42-373)	142.81 ± 65.35 (42-373)	226.52 ± 40.94 (151-314)
INR	1.12 ± 0.2 (0.83-1.84)	1.17 ± 0.2 (0.89-1.84)	0.99 ± 0.12 (0.83-1.3)
Right lobe images per patient	12.97 ± 6.06 (2-33)	13.02 ± 5.03 (3-24)	12.94 ± 6.69 (2-33)

Abbreviation: body mass index (BMI), aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyl-transpeptidase (GGT), international normalized ratio (INR)

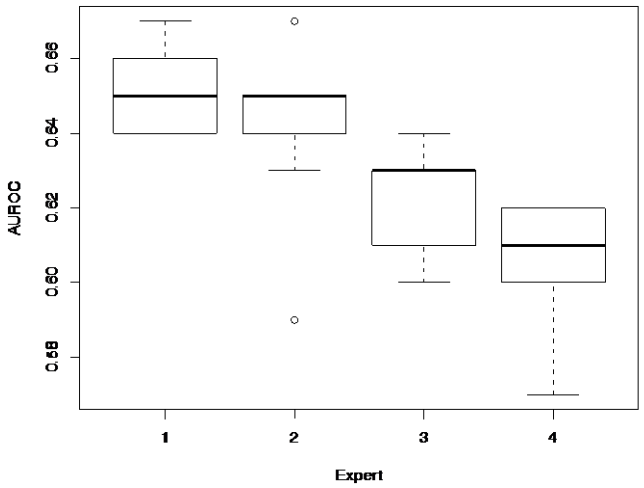

Figure 4. Box plot representing the estimated performance obtained when the same expert trains and uses the texture analysis tool in clinical practice. The top and the bottom of the boxes are the first and third quartiles, respectively. The length of the box thus represents the interquartile range within which 50% of the values were located. The line through the middle of each box represents the median.

In practice, a classifier is trained with data gathered from an expert but it can be used by other physicians. We identified two cases. First case, the expert that trained the classifier uses it in the current practice. In this scenario, the same expert that first established the ROIs establishes the ROIs for the new, unknown images. In the second scenario the expert that establishes the ROIs on the new images is different from the inital expert

The first scenario was simulated here by training a classifier with each dataset from each expert. Resulting classifier was evaluated using the other datasets from the same expert obtained at different ROI establishment step. Kruskal-Wallis test revealed that there is a significant variation due to the human expert ($p<0.001$), as seen in Figure 4. Again, most experienced experts provided best performance.

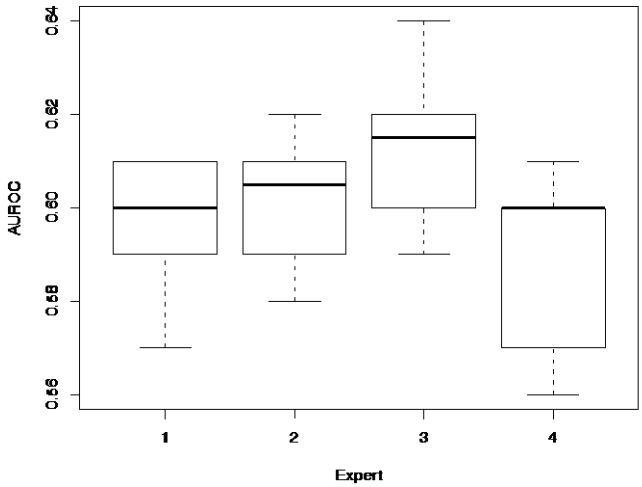

Figure 5. Box plot representing the estimated performance obtained when the texture analysis system is trained with datasets provided by one expert and used with ROIs established by a different expert. The top and the bottom of the boxes are the first and third quartiles, respectively. The length of the box thus represents the interquartile range within which 50% of the values were located. The line through the middle of each box represents the median.

The second scenario followed a similar setup, with the difference that the test sets are the all the other datasets from different experts. Kruskal-Wallis analysis revealed an interesting fact; there is no significant variance due to experts ($p=0.0506$) as shown in Figure 5. In both scenarios the analysis did not revealed significant variance due to the feature selection step.

F. Relevant textural features in fibrosis detection

The relevant features are presented in Table 2. For each algorithm we specified the parameters used to compute that feature. The reader is advised to read corresponding references for the meaning of the parameters.

IV. DISCUSSIONS

Liver biopsy is an imperfect golden standard in fibrosis staging. It is an invasive procedure and even if the method allows direct examination of the liver tissue there is a certain variability due to the reduced tissue volume and due to the fact that a human expert qualitatively evaluates the biopsy [28-30].

There are numerous research directions involving non invasive fibrosis staging and non invasive diagnosis of liver diseases in general [31, 32]. Papers [8, 17, 18, 33, 34] studying texture analysis as a non invasive staging tool reported high performances in cirrhosis detection [33] and even in fibrosis staging [8]. In these papers there are variations in terms of studied pathology and classification evaluation methodology. We believe that these factors might have positively biased the results reported by other authors.

TABLE II. RELEVANT FEATURES ORDERED BY THEIR RANK.

Rank	Algorithm name	Parameters
10	Texture edge co-occurrence matrix	(dx,dy)=(0,1); Computed statistic: Entropy
8	Gray tone difference matrix	K=5; Computed statistic: Busyness
6	Gray level co-occurrence matrix	(dx,dy)=(0,1); Computed statistic: Entropy
5	Law's texture energy	Kernel: W5L5; Computed stastistic: SPI
4	Fourier power spectrum	Computed statistic: First Moment
4	Law's texture energy	Kernel: S5S5; Computed stastistic: SPI
4	Law's texture energy	Kernel: W5E5; Computed stastistic: SVAI
4	Phase congruency based edge detection	(dx,dy)=(0,1); Computed statistic: Sum Variance
4	Phase congruency based edge detection	Sum Variance
4	Texture edge co-occurrence matrix	(dx,dy)=(0,2); Computed statistic: Entropy

Present study aims to evaluate the dependence between the human expert and the performance of such a texture analysis system in predicting the cirrhosis in chronic hepatitis C patients. In the same time present paper brings the following contributions to the non invasive fibrosis detection field: it includes only patients with chronic hepatitis C, excluding other pathologies; it integrates almost all textural algorithms met in fibrosis detection and it proposes a more rigorous performance evaluation methodology that gives results closer to the real performance of a classifier.

In present study we included only patients with chronic hepatitis C etiology. Other papers that address the non invasive detection of cirrhosis include patients having different pathologies like fatty infiltration[12]. Another important highlight of this paper is the volume of patients. There are few papers that study more than 100 patients but not all the patients included in these studies have chronic hepatitis C, or the etiology for cirrhosis is not specified [4, 12, 35].

Another factor that could be responsible for low estimated performance of texture analysis system (AUROC=0.64) is the performance estimation method. In literature, the performance estimation methods vary. Most of the authors divided the available dataset in training and test lot[4, 34]. Training set is used to train the classifier and test set to evaluate it. Some papers employ a form of cross validation [13, 14]. The parameters for the classifier or the parameters for textural algorithms are manually searched in order to improve detection rates. There is no mention if this optimization is performed on an independent dataset or on the same dataset that is used for training [6, 13, 16]. The dominant performance measure is accuracy even if the datasets are not balanced [17, 33, 34].

Most classification schemas can easily overfit the data and give strongly positive biased performance estimations. This positive bias is accentuated when the feature selection or the parameter search is performed on the same dataset that is used to estimate the performance. A rigorous performance evaluation methodology is required in order to obtain unbiased results.

Performance estimation algorithm proposed in this paper ensures that each time the classifier is tested the test data are new and unseen at the training or feature selection phase. The meta parameter sets are evaluated on unseen data to ensure that we do not select a classificator instance that overfits the training data. The outer cross validation loop ensures that even this search procedure doesn't overfit the data. The 10 times repetition of the evaluation phase ensures a better estimation of the mean performance. No other papers employed repeated performance estimation on their classification schemas. When performing one iteration the data might get partitioned in such a way that by accident the performance estimation is very high. For example, in some iterations the performance reached levels as high as 0.79. Of course, the mean performance estimated over 10 runs is smaller. The same phenomenon of increased variance can be noted when the performance measure is computed on each test fold and not on the entire prediction vector. In 2 fold CV a "lucky" splitting might give a very high performance reading.

In present paper, the CV predictions are collected and the performance is measured on a vector that has the same dimensions as the initial dataset.

Textural feature selection is performed on an independent dataset. This dataset is obtained by randomly sampling the original datasets. It is important to note that each instance that is included in the feature selection dataset is excluded from the original dataset. As a result, the feature selection process has less chances of overfitting.

There are other algorithms [19] in literature that can be applied to feature selection. The method proposed here has the advantage that it uses the well known univariate analysis instead of other, less known methods. On the other hand, other methods might give a better feature selection (for example, Correlation based feature selection [36]). The evaluation of such algorithms is beyond the scope of this paper.

The particular set of features does not influence the detection rates. The subset of features selected at each step has

a high variability. High ranking features cover large spectra of algorithms, from statistical algorithms to multirezolution analysis. This indicates that the specific algorithm used to numerically describe the texture has its importance but there are fewer chances that new textural algorithms will make a great impact over cirrhosis detection and fibrosis staging.

The design of the experiment, where each expert establishes 5 sets of ROI's ensures that the samples are independent and normally distributed. Each set of patients have different randomizations in order to minimize the effect of patient/image succession over the experiment. Moreover, for each patient the order of the images is altered. It is important to note that the order of images is the same for all the experts. Expert x viewed the patients and images in the same order as the Expert y when establishing ROI for the same dataset z.

The main finding of this paper is that the performance of the studied software diagnosis tool depends on the expert that employs this tool. In the results section we have shown that there is a significant performance variation between experts. The results presented here showed that more experienced experts tend to capture the same aspects of the ultrasound image, aspects that are consistent with the histological findings. If this tool is trained and employed by an experienced physician it might give some extra information about the underlying pathology.

The results from the second scenario, when the expert that uses the texture analysis tool is different from the expert that provided the data for training, revealed the fact that there is little use for texture analysis tool in screening processes.

The classical methodology has a severe drawback. It requires a human expert to establish a representative area where the texture will be analyzed. Replacing the human expert with a computerized solution will improve the usefulness of such a software analysis tool. Studies performed on the other major diffuse disease, liver steatosis revealed that such a replacement is possible [37].

V. CONCLUSIONS

Texture analysis can enhance the diagnosis power of the B-mode ultrasound image. The performance of this approach depends highly on the human expert that establishes the regions of interest. In classical form met in literature non invasive diagnosis through texture analysis has limited utility in clinical practice. Further work in this domain has to be focused on reducing or eliminating the role of human expert and not on finding new textural features.

ACKNOWLEDGMENT

The authors would like to thank to Vlad Popovici from Swiss Institute of Bioinformatics for the priceless help in designing the experiments presented here. Part of this work was funded by National Council for Scientific Research in Higher Education, Grant nr. 41-071/2007: SONOFIBROCAST.

REFERENCES

[1] C. M. Wu, et al., "Texture Features for Classification of Ultrasonic Liver Images," Ieee Transactions on Medical Imaging, vol. 11, pp. 141-152, Jun 1992.

[2] K. J. Taylor, et al., "Quantitative US attenuation in normal liver and in patients with diffuse liver disease: importance of fat," Radiology, vol. 160, pp. 65-71, Jul 1986.

[3] B. Oosterveld, et al., "Ultrasound attenuation and texture analysis of diffuse liver disease: methods and preliminary results," Phys. Med, vol. 36, pp. 1034-1064, 1991.

[4] Y. M. Kadah, et al., "Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images," Ieee Transactions on Medical Imaging, vol. 15, pp. 466-478, Aug 1996.

[5] D. Gaitini, et al., "Feasibility study of ultrasonic fatty liver biopsy : texture vs. attenuation and backscatter," Ultrasound in Medicine and Biology, vol. 30, pp. 1321-1327, Oct 2004.

[6] G. T. Cao, et al., "Liver fibrosis identification based on ultrasound images captured under varied imaging protocols," J Zhejiang Univ Sci B, vol. 6, pp. 1107-1114, Nov 2005.

[7] D. Gaitini, et al., "Computerised analysis of liver texture with correlation to needle biopsy," Ultraschall Med, vol. 26, pp. 197-202, Jun 2005.

[8] H. Yamada, et al., "A pilot approach for quantitative assessment of liver fibrosis using ultrasound: preliminary results in 79 cases," Journal of Hepatology, vol. 44, pp. 68-75, Jan 2006.

[9] J. W. Jeong, et al., "The echotextural characteristics for the diagnosis of the liver cirrhosis using the sonographic images," 2007 Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-16, pp. 1343-1345, 2007.

[10] C. Abe, et al., "Computer-aided detection of diffuse liver disease in ultrasound images," Invest Radiol, vol. 27, pp. 71-77, Jan 1992.

[11] A. Materka and M. Strzelecki, "Texture analysis methods a review," Technical University of Lodz, Institute of Electronics,1998.

[12] A. M. Badawi, et al., "Fuzzy logic algorithm for quantitative tissue characterization of diffuse liver diseases from ultrasound images," Int J Med Inform, vol. 55, pp. 135-147, Aug 1999.

[13] W. C. Yeh, et al., "Liver fibrosis grade classification with B-mode ultrasound," Ultrasound in Medicine and Biology, vol. 29, pp. 1229-1235, Sep 2003.

[14] W. Lee, et al., "A study of ultrasonic liver images classification with artificial neural networks based on fractal geometry and multiresolution analysis," Biomedical Engineering - Applications, Basis&Communications, vol. 16, pp. 17-25, 2004.

[15] Y. Xia, et al., "Morphology-based multifractal estimation for texture segmentation," Ieee Transactions on Image Processing, vol. 15, pp. 614-623, Mar. 2006.

[16] A. Ahmadian, et al., "A texture classification method for diffused liver diseases using Gabor wavelets," 2005 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1-7, pp. 1567-1570, 2005.

[17] G. Cao, et al., "Ultrasonic Liver Characterization Using Phase Congruency," Proc. 27th Ann. Int. Conf. IEEE-EMBS pp. 6356-6359, Jan. 2005.

[18] M.-H. Horng, et al., "Texture feature coding method for classification of liver sonography," Computerized Medical Imaging and Graphics, vol. 26, pp. 33-42, Jan-Feb 2002.

[19] I. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann Pub, 2005.

[20] V. Popovici, "Assessment of classification models for medical applications," presented at the Workshop on Computers in Medical Diagnoses, Cluj-Napoca, Romania 2009.

[21] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," Annals of Statistics, vol. 29, pp. 1165-1188, Aug 2001.

[22] H. Akaike, "A new look at the statistical model identification," IEEE transactions on automatic control, vol. 19, pp. 716-723, 1974.

[23] M. Pepe, The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.

[24] P. Brodatz, Textures: A Photographic Album for Artists and Designers.: Dover, New York, 1966.

[25] C. Chih-Chung and L. Chih-Jen. LIBSVM: a library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[26] P. Bedossa and T. Poynard, "An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group," Hepatology, vol. 24, pp. 289-293, Aug 1996.

[27] "Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. The French METAVIR Cooperative Study Group," Hepatology, vol. 20, pp. 15-20, Jul 1994.

[28] A. Regev, et al., "Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection," Am J Gastroenterol, vol. 97, pp. 2614-2618, Oct 2002.

[29] P. Bedossa, et al., "Sampling variability of liver fibrosis in chronic hepatitis C," Hepatology, vol. 38, pp. 1449-1457, 2003.

[30] M. Guido and M. Rugge, "Liver biopsy sampling in chronic viral hepatitis," Semin Liver Dis, vol. 24, pp. 89-97, Feb 2004.

[31] I. Guha and W. M. Rosenberg, "Noninvasive Assessment of Liver Fibrosis: Serum Markers, Imaging, and Other Modalities," Clinics in Liver Disease, vol. 12, pp. 883-900, Nov. 2008.

[32] S. Bonekamp, et al., "Can imaging modalities diagnose and stage hepatic fibrosis and cirrhosis accurately?," Journal of Hepatology, vol. 50, pp. 17-35, Jan 2009.

[33] G. T. Cao, et al., "Liver fibrosis identification based on ultrasound images," 2005 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vols 1-7, pp. 6317-6320, 2005.

[34] A. Mojsilovic, et al., "Characterization of visually similar diffuse diseases from B-scan liver images using nonseparable wavelet transform," Ieee Transactions on Medical Imaging, vol. 17, pp. 541-549, Aug 1998.

[35] A. Szebeni, et al., "Correlation of ultrasound attenuation and histopathological parameters of the liver in chronic diffuse liver diseases," European Journal of Gastroenterology & Hepatology, vol. 18, pp. 37-42, Jan 2006.

[36] M. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," Department of Computer Science, University of Waikato, 1998.

[37] C. Vicas, et al., "Automatic detection of liver capsule using Gabor filters. Applications in steatosis quantification," IEEE Proc.Intelligent Computer Comm. Process., pp. 133-140, Aug. 2009.