Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers

Stavroula G. Mougiakakou\textsuperscript{a,1,*}, Ioannis K. Valavanis\textsuperscript{a,1}, Alexandra Nikita\textsuperscript{b,2}, Konstantina S. Nikita\textsuperscript{a,1}

\textsuperscript{a}National Technical University of Athens, Faculty of Electrical and Computer Engineering, Biomedical Simulations and Imaging Laboratory, 9 Heroon Polytechneiou Str., 15780 Zografou, Athens, Greece
\textsuperscript{b}University of Athens, Medical School, Department of Radiology, 20 Papadiamantopoulou Str., 15228 Athens, Greece

Received 9 January 2007; received in revised form 16 May 2007; accepted 22 May 2007

Summary

Objectives: The aim of the present study is to define an optimally performing computer-aided diagnosis (CAD) architecture for the classification of liver tissue from non-enhanced computed tomography (CT) images into normal liver (C1), hepatic cyst (C2), hemangioma (C3), and hepatocellular carcinoma (C4). To this end, various CAD architectures, based on texture features and ensembles of classifiers (ECs), are comparatively assessed.

Materials and methods: Number of regions of interests (ROIs) corresponding to C1—C4 have been defined by experienced radiologists in non-enhanced liver CT images. For each ROI, five distinct sets of texture features were extracted using first order statistics, spatial gray level dependence matrix, gray level difference method, Laws' texture energy measures, and fractal dimension measurements. Two different ECs were constructed and compared. The first one consists of five multilayer perceptron neural networks (NNs), each using as input one of the computed texture feature sets or its reduced version after genetic algorithm-based feature selection. The second EC comprised five different primary classifiers, namely one multilayer perceptron NN, one probabilistic NN, and three \textit{k}-nearest neighbor classifiers, each fed with the combination of the five texture feature sets or their reduced versions. The final decision of each EC was extracted by using appropriate voting schemes, while

\* Corresponding author. Tel.: +30 210 772 2968; fax: +30 210 772 3557.
E-mail address: smougia@cc.ece.ntua.gr (S.G. Mougiakakou).
1 Tel.: +30 210 772 2968; fax: +30 210 772 3557.
2 Tel.: +30 210 722 7488; fax: +30 210 729 2280.

0933-3657/$ — see front matter \copyright\ 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2007.05.002
1. Introduction

One of the most common and robust imaging techniques for the detection of hepatic lesions is computed tomography (CT) [1]. Although the quality of CT images has been significantly improved during the last years, it is difficult in some cases, even for experienced doctors, to make a 100% accurate diagnosis. In these cases, the diagnosis has to be confirmed by administration of contrast agents, which is related with renal toxicity and allergic reactions, or invasive procedures (biopsies). During the last years, along with the developments in image processing and artificial intelligence, computer-aided diagnosis (CAD) systems, aiming at the characterization of liver tissue, attract much attention, since they can provide diagnostic assistance to clinicians, and contribute to reduction of the number of required biopsies.

Various approaches, most of them using ultrasound B-scan and CT images, have been proposed based on different image characteristics, such as texture features, estimated from first- and second-order gray level statistics, and fractal dimension estimators combined with various classifiers [2—5]. Texture analysis of liver CT images based on spatial gray level dependence matrix (SGLDM), gray level run length method (GLRLM), and gray level difference method (GLDM) has been proposed by Mir et al. [6], in order to discriminate normal from malignant hepatic tissue. Chen et al. [7] have applied SGLDM texture features to a probabilistic neural network (P-NN) for the characterization of hepatic tissue (hepatoma and hemangioma) from CT images. Additionally, SGLDM-based texture features fed to a system of three sequentially placed neural networks (NNs) have been used by Gletsos et al. [8] for the classification of hepatic tissue into four categories. Although a lot of effort has been devoted to liver tissue characterization, the developed systems are most of the times limited to two or three classes of liver tissue and/or do not gain from the interaction of different texture characterization methods, or the combination of different classifiers.

In order to select the most robust characteristics from an initial high-dimensional feature set, that might be derived from different feature extraction techniques, feature selection methods can be applied. Deterministic, or stochastic feature selection methods decrease the feature extraction costs of the classification system, and may also enhance its performance [9]. During the last years, an increasing number of researchers are using genetic algorithms (GAs) for dimensionality reduction. The use of GAs for feature selection was first introduced in 1989 [10], and since then, GAs have been successfully applied to a broad spectrum of dimensionality reduction studies [11—13]. According to Ref. [8], the use of GAs results in more robust feature vectors as compared to other deterministic feature selection techniques, in problems related to liver tissue classification from CT images.

In the last decade, the use of multiple classifier systems has been proposed in order to optimize the performance of CAD systems. A set of classifiers whose individual predictions are fused through a combining strategy, usually a voting scheme, to classify new examples constitutes an ensemble of classifiers (EC) [14]. The attraction that this topic exerts on machine learning and diagnostic decision support research is based on the premise that ECs are often much more accurate than any individual classifier of the set [15,16]. Early diagnosis of melanoma has been facilitated by combining three types of classifiers, namely linear discriminant analysis (LDA), k-nearest neighbor (k-NN), and a decision tree, with a voting scheme [17]. A multiple classification system based on a committee of NNs, trained by the Levenberg—Marquardt algorithm, along with a voting scheme across the NN outputs has been used by Jerebko et al. [18] for the detection of colonic polyps in CT colonography data. Furthermore, a novel system for diabetes diagnosis

bootstrap re-sampling was utilized in order to estimate the generalization ability of the CAD architectures based on the available relatively small-sized data set.

Results: The best mean classification accuracy (84.96%) is achieved by the second EC using a fused feature set, and the weighted voting scheme. The fused feature set was obtained after appropriate feature selection applied to specific subsets of the original feature set.

Conclusions: The comparative assessment of the various CAD architectures shows that combining three types of classifiers with a voting scheme, fed with identical feature sets obtained after appropriate feature selection and fusion, may result in an accurate system able to assist differential diagnosis of focal liver lesions from non-enhanced CT images.

© 2007 Elsevier B.V. All rights reserved.
has been proposed [19], which is based on retinal images fractal characteristics and applies a voting scheme across the outputs of an EC consisting of a back-propagation trained NN, a radial basis function NN, and a GA-based classifier. The use of texture features and shape parameters along with a multi-classifier modular architecture composed from a self-organizing map (SOM) and/or k-NN classifiers has been recently proposed by Christodoulou et al. [15], aiming at the characterization of individuals with asymptomatic carotid stenosis at risk of stroke. Although previous studies [20,21] have shown that CAD systems based on various texture features and ECs can enhance the diagnosis efficiency of CT focal liver lesions, the evaluation of the proposed methods on small-sized samples constitutes a significant drawback for these studies. To address this drawback, re-sampling methods like cross-validation, jack-knife, and bootstrap can be applied [22–24]. In Ref. [18], cross-validation has been applied for sensitivity estimation of a colonic polyp detection system, while in Ref. [25] the bootstrap method has been used for the development of a diagnosis system able to differentiate benign and malignant tumors from breast ultrasound images. The bootstrap method, which was introduced by Efron [22] as an approach to calculate confidence intervals for parameters where standard methods cannot be applied, is based on re-sampling with replacement. The comparative assessment of various re-sampling methods has shown that the bootstrap method provides less biased and more consistent results than the jack-knife method [26].

The principal aim of the present paper is to assess the potential of ECs in the development of a CAD system able to discriminate four hepatic tissue types (normal liver, hepatic cyst, hemangioma, and hepatocellular carcinoma) from non-enhanced CT images. Furthermore, the use of a variety of texture features as input to the CAD system is examined while the application of feature selection based on a GA is investigated aiming at improving the resulting classification performance. In order to overcome problems with small data sets and biased classification performances, the bootstrap method is applied. In this framework, five different CAD architectures based on the above design concepts are comparatively assessed.

The rest of the paper is organized as follows: in Section 2, the generic system design concepts are presented, including description of the data used, the methodology of feature extraction and selection, the ECs and the applied voting schemes, as well as the five alternative architectures of the CAD system. In Section 3, the experimental results of the five CAD system architectures are presented and compared, followed by conclusions presented in Section 4.

2. Methodology

The generic design of a CAD system aiming at the classification of CT liver tissue into one of the four classes: normal liver (C1), hepatic cyst (C2), hemangioma (C3), and hepatocellular carcinoma (C4), is presented in Fig. 1. Firstly, regions of interest (ROIs) drawn by an experienced radiologist on CT images were driven to a feature extraction module, where five different texture feature sets were obtained using first order statistics (FOS), spatial gray level dependence matrices (SGLDM), gray level difference matrix (GLDM), Laws’ texture energy measures (TEM), and fractal dimension measurements (FDM). Then, the full feature sets or their reduced versions obtained after proper feature selection in the feature selection module, were fed to an EC. Two alternative approaches for constructing ECs were studied. The primary classifiers of the first ensemble (EC1) were generated by applying a single learning algorithm to different feature subsets, while the classifiers of the second ensemble (EC2) were generated by using different learning algorithms on the entire feature set. The predictions obtained from the primary classifiers of each ensemble were combined using appropriate voting schemes.

In the following, the modules of the CAD systems are presented, and the five alternative CAD system architectures are described.
2.1. Image acquisition

Abdominal non-enhanced CT images with a spatial resolution of 512 × 512 pixels and 8-bit gray level at the W150 + 60 window taken from both patients and healthy controls were acquired using a conventional single detector CT scanner (Philips LX, Philips Medical Systems). The slice thickness ranged from 5 mm to 8 mm, depending on the lesion size. A total of 38 healthy controls were identified by the radiologist, while the diagnosed hepatic lesions from patients with C2 (15 patients), C3 (24 patients), and C4 (20 patients), were validated by needle biopsies, denoted C1 (5 patients), were acquired using a conventional single detector CT scanner (Philips LX, Philips Medical Systems). The slice thickness ranged from 5 mm to 8 mm, depending on the lesion size. A total of 38 healthy controls were identified by the radiologist, while the diagnosed hepatic lesions from patients with C2 (15 patients), C3 (24 patients), and C4 (20 patients), were validated by needle biopsies, denoted C1 (5 patients), were validated by needle biopsies, denoted C2 (15 patients), C3 (24 patients), and C4 (20 patients), were validated by needle biopsies, denoted C1 (5 patients), C2 (15 patients), C3 (24 patients), and C4 (20 patients).

2.2. Feature extraction

Five sets of texture features were calculated for each ROI corresponding to FOS, SGLDM, GLDM, TEM, and FDM.

2.2.1. First order statistics

Features from FOS [27] are easily computed from the intensity function of the image. In our experiments, six features were calculated for each ROI from the following equations:

average gray level : \( \text{avg}_{\text{FOS}} = \sum g H(g) \) (1)

standard deviation : \( \text{std}_{\text{FOS}} = \sqrt{\sum (g - \text{avg}_{\text{FOS}})^2 H(g)} \) (2)

entropy : \( \text{ent}_{\text{FOS}} = -\sum g H(g) \ln(H(g)) \) (3)

coefficient of variation : \( \text{cv}_{\text{FOS}} = \frac{\text{std}_{\text{FOS}}}{\text{avg}_{\text{FOS}}} \) (4)

\[ \text{sk}_{\text{FOS}} = \frac{1}{\text{std}_{\text{FOS}}^2} \sum g (g - \text{avg}_{\text{FOS}})^3 H(g) \] (5)

\[ \text{kur}_{\text{FOS}} = \frac{1}{\text{std}_{\text{FOS}}^4} \sum g (g - \text{avg}_{\text{FOS}})^4 H(g) - 3 \] (6)

where \( g \) are the possible values of the gray level of the image and \( H(g) \) is the percentage of the pixels in the image with gray level equal to \( g \).

2.2.2. Spatial gray level dependence matrices

The texture characteristics of each ROI, which actually correspond to second order statistics, can be derived from the SGLDM of the ROI [27,28]. The elements of the SGLDM represent the values of the probability function \( P_{ij} \), which measure the normalized frequency in which all pixel pairs consisting of pixels with interpixel distance \( d \) along a direction \( \theta \) and with gray level values \( i \) and \( j \), respectively, appear in the image. The features calculated in our experiments from SGLDM are given by the following equations:

angular second moment :

\[ \text{asm}_{\text{SGLDM}} = \sum_i \sum_j p_{ij}^2 \] (7)

contrast :

\[ \text{con}_{\text{SGLDM}} = \frac{N_8}{n^2} \left\{ \sum_{i,j} p_{ij} \mid i-j = n \right\} \] (8)

correlation :

\[ \text{cor}_{\text{SGLDM}} = \frac{\sum_i \sum_j (ij p_{ij}) - \mu^2}{\sigma^2} \] (9)

variance :

\[ \text{var}_{\text{SGLDM}} = \sum_i \sum_j (ij p_{ij}) - \mu^2 \] (10)

inverse difference moment :

\[ \text{idm}_{\text{SGLDM}} = \sum_i \sum_j P_{ij} \frac{P_{ij}}{1+(i-j)^2} \] (11)

entropy :

\[ \text{ent}_{\text{SGLDM}} = -\sum_i \sum_j p_{ij} \ln(p_{ij}) \] (12)

homogeneity :

\[ \text{hg}_{\text{SGLDM}} = \sum_i \sum_j P_{ij} \frac{P_{ij}}{1+|i-j|} \] (13)

cluster tendency :

\[ \text{clt}_{\text{SGLDM}} = \sum_i \sum_j (i+j-2\mu) p_{ij} \] (14)

where \( P_{ij} \) is the \((i, j)\) element of SGLDM, \( \mu \) and \( \sigma \) are the mean value and the standard deviation of the elements of the matrix, respectively, and \( N_8 \) is the number of gray levels in the image. In the present study, four values for the angle \( \theta \) were used: \( \theta = 0^\circ, 45^\circ, 90^\circ, \) and \( 135^\circ \), and the features

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Sex (male/female)</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td>35</td>
<td>70</td>
<td>55.5</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>44</td>
<td>70</td>
<td>58.3</td>
</tr>
<tr>
<td>C3</td>
<td></td>
<td>45</td>
<td>63</td>
<td>52.7</td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td>61</td>
<td>78</td>
<td>70.5</td>
</tr>
</tbody>
</table>
were calculated for distances of 1, 2, 4, 6, 8, and 12 pixels. For each distance, the final feature value was computed by averaging over the feature values corresponding to the four angular directions. Thus, a total of 48 texture characteristics was obtained for each ROI.

2.2.3. Gray level difference matrix

We consider here a class of local properties [29] based on absolute differences between the gray levels of pixels in the ROI. Let \( I(x, y) \) be the image intensity function. For any given displacement \( \delta = (\Delta x, \Delta y) \) let \( I_\delta(x, y) = |I(x, y) - I(x + \Delta x, y + \Delta y)| \), and \( f(\delta) \) be the probability density of \( I_\delta(x, y) \). The value of \( f(\delta) \) is obtained from the number of times \( I_\delta(x, y) \) occurs for a given \( \delta \), i.e.

\[
E(\delta) = \frac{1}{N} \sum_{x,y} |I(x, y) - I(x + \Delta x, y + \Delta y)|
\]

where \( N \) is the number of gray levels in the image and \( \delta \) is considered from the set of possible forms of the vector \( \delta \), which are obtained after transposing the four Laws’ zero-sum masks were used: \( L5 \equiv L3 \times L3 = (1, 4, 6, 4, 1), \) \( S5 \equiv L3 \times S3 = -E3 \times E3 = (-1, 0, 2, 0, -1), \) \( R5 \equiv S3 \times S3 = (1, -4, 6, -4, 1), \) \( E5 \equiv L3 \times E3 = (-1, -2, 0, 2, 1), \) and \( W5 \equiv -E3 \times S3 = (-1, 2, 0, -2, 1). \) Vector \( L5 \) performs local averaging, \( S5 \) and \( E5 \) are spot and edge detectors, respectively, while \( R5 \) and \( W5 \) can be regarded as “ripple” and “wave” detectors, respectively. If we multiply the column vectors of length five, which are obtained after transposing \( L5, S5, R5, E5, \) and \( W5, \) by the row vectors \( L5, S5, R5, E5, \) and \( W5, \) we obtain Laws’ \( 5 \times 5 \) masks. In our study, the following four Laws’ zero-sum masks were used: \( L55 = L5^5, E55 = -E5^5, L5S5 = L5^3S5, R5S5 = R5^3S5. \)

After convolving each ROI image with each of the four masks, the following were calculated:

\[
\begin{align*}
\text{sum of absolute values} & : \quad a_{\text{TEM}} = \frac{\sum_x \sum_y |I(x, y)|}{\# \text{ of pixels of the original image}} \\
\text{sum of squares} & : \quad s_{\text{TEM}} = \frac{\sum_x \sum_y I^2(x, y)}{\# \text{ of pixels of the original image}} \\
\text{entropy} & : \quad e_{\text{TEM}} = -\sum_g H_L(g) \ln H_L(g)
\end{align*}
\]

where \( I(x, y) \) is the intensity function of the image after its convolution with the Law’s mask, \( g \) are the possible values of the gray level of the convolved image, and \( H_L(g) \) is the percentage of the pixels of the convolved image with gray level equal to \( g \). Thus, twelve Laws’ energy measures \( 4 \times 3 \) statistics per mask) were calculated for each ROI.

2.2.4. Laws’ texture energy measures

Laws’ TEM [30] are derived from three simple vectors of length three: \( L3 \equiv (1, 2, 1), E3 \equiv (-1, 0, 1), \) and \( S3 \equiv (-1, 2, -1), \) which represent the one-dimensional operations of center-weighted local averaging, the symmetric first differenting for edge detection, and the second differenting for spot detection, respectively. If these vectors are convolved with themselves or with each other, five vectors of length five are obtained: \( L5 \equiv L3 \times L3 = (1, 4, 6, 4, 1), \) \( S5 \equiv L3 \times S3 = -E3 \times E3 = (-1, 0, 2, 0, -1), \) \( R5 \equiv S3 \times S3 = (1, -4, 6, -4, 1), \) \( E5 \equiv L3 \times E3 = (-1, -2, 0, 2, 1), \) and \( W5 \equiv -E3 \times S3 = (-1, 2, 0, -2, 1). \) Vector \( L5 \) performs local averaging, \( S5 \) and \( E5 \) are spot and edge detectors, respectively, while \( R5 \) and \( W5 \) can be regarded as “ripple” and “wave” detectors, respectively. If we multiply the column vectors of length five, which are obtained after transposing \( L5, S5, R5, E5, \) and \( W5, \) by the row vectors \( L5, S5, R5, E5, \) and \( W5, \) we obtain Laws’ \( 5 \times 5 \) masks. In our study, the following four Laws’ zero-sum masks were used: \( L55 = L5^5, E55 = -E5^5, L5S5 = L5^3S5, R5S5 = R5^3S5. \)

After convolving each ROI image with each of the four masks, the following were calculated:

\[
\begin{align*}
\text{sum of absolute values} & : \quad a_{\text{TEM}} = \frac{\sum_x \sum_y |I(x, y)|}{\# \text{ of pixels of the original image}} \\
\text{sum of squares} & : \quad s_{\text{TEM}} = \frac{\sum_x \sum_y I^2(x, y)}{\# \text{ of pixels of the original image}} \\
\text{entropy} & : \quad e_{\text{TEM}} = -\sum_g H_L(g) \ln H_L(g)
\end{align*}
\]

where \( I(x, y) \) is the intensity function of the image after its convolution with the Law’s mask, \( g \) are the possible values of the gray level of the convolved image, and \( H_L(g) \) is the percentage of the pixels of the convolved image with gray level equal to \( g \). Thus, twelve Laws’ energy measures \( 4 \times 3 \) statistics per mask) were calculated for each ROI.

2.2.5. Fractal dimension measurements

In order to extract more information from the liver CT images, a feature extraction method based on the concepts of the fractional Brownian motion (FBM) model and multiple resolution imagery was employed [5]. According to the FBM model developed by Mandelbrot [31], the roughness of a surface, naturally occurring as the end result of random walks, can be characterized by its fractal dimension
(FD). For a CT liver image, which can be considered as such a surface, FD is estimated from the equation 
\[ E(\Delta I^2) = c(\Delta r)^{6-2FD} \]
where \( E(\cdot) \) denotes the expectation operator, \( \Delta l \equiv (l(x_2, y_2) - l(x_1, y_1)) \) is the intensity variation of two pixels, \( c \) is some constant, and \( \Delta r \equiv ||(x_2, y_2)-(x_1, y_1)|| \) is the spatial distance of pixels. A simpler method is to estimate the \( H \) parameter, which is \( H = 3 - FD \), from the equation 
\[ E(\Delta I^2) = c(\Delta r)^{2H} \]
If we apply the log function to both sides of the last equation we obtain:

\[ \log[E(\Delta I^2)] = \log(c) + 2H \log(\Delta r) \]  \hspace{1cm} (23)

Given an image \( I \), the intensity difference vector is defined as \( \text{IDV} \equiv [\text{id}(1), \text{id}(2), \ldots, \text{id}(s)] \) where \( s \) is the maximum inter-pixel distance depending on the size of the image, and \( \text{id}(k) \) is the average of the absolute intensity difference of all pixel pairs with horizontal, vertical or diagonal distance \( k \). According to Eq. (23), the value of \( H \) can be obtained by using least-squares linear regression to estimate the slope of the curve of \( \text{id}(k) \) versus \( k \) in log-log scales. Obviously, the value of \( H \) is equal to the half of the slope of the curve.

According to Ref. [5], we can obtain more texture information, i.e. for the lacunarity or the regularity of an image, if we calculate the values \( H(k) \), which correspond to the \( H \) value of the image computed for an image resolution \( k \). The original image corresponds to a \( k = 1 \) resolution, the original image reduced to a half-sized image corresponds to a \( k = 2 \) resolution, etc. Using the pyramidal approach, described in [5], for multiresolutional analysis we obtained three features, \( H_{FDM}^{1}, H_{FDM}^{2}, H_{FDM}^{3} \), for each ROI, which correspond to the \( H \) value of the ROI at image resolutions \( k = 1, 2, 3 \).

### 2.3. Feature selection

For the purpose of feature selection, a GA based on Ref. [32] was used in the present paper. The algorithm makes use of a randomly created initial population of \( N \) chromosomes. Each chromosome is a binary mask, with 1 indicating that the feature is selected, and 0 that the corresponding feature is omitted. The initial population is updated through the following procedure:

For \( N_b \) generations

\begin{itemize}
  \item Calculate the fitness function of the \( N \) chromosomes
  \item Select \( N/2 \) pairs from population using the elitist selection method
  \item Mate selected chromosomes using two-point crossover, with crossover probability \( P_c \)
  \item Switch value of chromosome bits, with mutation probability \( P_m \)
  \item Update population
\end{itemize}

Assign fitness values to new population, store best results

The maximum squared Mahalanobis distance between classes computed for the samples belonging to the training set was used as fitness function [8]. Since the number of selected features is not taken into account in computing the fitness function, a “penalty” function for feature sets exceeding a given dimensionality threshold was applied. Thus, the corresponding individuals were assigned a fitness value equal to 50% of the average population fitness. The GA was run for a dimensionality threshold equal to 10. This threshold has been successfully applied in the selection of texture features using GAs in the problem of focal liver lesion characterization [8]. The GA run parameters, manually tuned after a series of trials, were \( N = 200 \), \( N_b = 250 \), \( P_c = 0.8 \) and \( P_m = 0.008 \).

### 2.4. Classification

The estimated texture feature sets were applied to either of two different ensembles of classifiers (EC1 and EC2). EC1 was constructed by combining five multilayer perceptron NNs (MLP-NN), each fed with one out of the five distinct texture feature sets (MLP-NN1 with FOS, MLP-NN2 with SGLDM, MLP-NN3 with GLDM, MLP-NN4 with TEM, MLP-NN5 with FDM). EC2 was constructed by combining one MLP-NN, one probabilistic NN (P-NN), one 1-nearest neighbor (1-NN) classifier, and two modified \( k \)-NNs (\( mk \)-NN) (\( k > 1 \)) classifiers, each fed with the combination of the five computed texture feature sets. For each EC, the final decision was generated by combining the diagnostic outputs of the corresponding primary classifiers through appropriate voting schemes. In order to obtain reliable results, in terms of classification accuracy, the development of both EC1 and EC2 was based on the bootstrap method. To this end, a training set consisting of 147 ROIs was sampled with replacement from the available 147 ROIs. The ROIs not appearing in the training set were randomly allocated into two equally sized sets (validation and testing sets). The procedure was repeated for 50 times, resulting in 50 groups of training, validation and testing sets. For each group of sets, EC1 and EC2 were developed and tested. In order to train the primary classifiers of each ensemble the training and validation sets were used. The generalization ability of both the primary classifiers and the ECs was tested using the testing set. The various CAD architectures were assessed by estimating the mean classification accuracy and the corresponding standard deviation for all bootstrap sets.
2.4.1. Multilayer perceptron neural network
The MLP-NN classifier [33] used in this study is based on a feed-forward NN consisting of one input layer with a number of input neurons equal to the number of features fed into the NN, one hidden layer with variable number of neurons, and one output layer consisting of two output neurons, encoding the different types of liver tissue. Thus, the outputs (00), (01), (10), and (11) correspond to the four liver tissue classes C1, C2, C3, and C4, respectively. The MLP-NN was trained, using the training set, by the batched back-propagation (BP) algorithm with adaptive learning rate and momentum [33]. Moreover, the optimal MLP-NN was trained, using the training set, by the batched back-propagation (BP) algorithm with adaptive learning rate and momentum [33]. The use of the tan-sigmoid function in the hidden layer, the values of the input vectors were normalized from −1.0 to 1.0. The use of the tan-sigmoid function in the output neurons resulted in output values in the range [0.0, 1.0]: a value of less than 0.5 corresponded to 0, while a value greater or equal to 0.5 corresponded to 1. The MLP-NN was trained, using the training set, by the batched back-propagation (BP) algorithm with adaptive learning rate and momentum [33]. Moreover, the optimal MLP-NN classifier in terms of hidden neurons and internal NN parameters (appropriate values of momentum and initial learning rate) was determined using a trial-and-error process, until no further improvement of classification accuracy in the validation set could be obtained.

2.4.2. Probabilistic neural network
The P-NN performs interpolation in multidimensional space [2]. The P-NN consists of one input layer with number of neurons equal to the number of used features, a hidden (or training pattern unit) layer, a summation unit layer, and an output layer. In order to classify a ROI, the corresponding feature set of used features, a hidden (or training pattern unit) layer with number of neurons equal to the number of hepatic tissue, consists of two output neurons, encoding the class of the feature vector , one hidden layer with number of neurons equal to the number of selected in the range −1.0, 1.0]. The use of the tan-sigmoid function in the hidden layer, the values of the input vectors were normalized from −1.0 to 1.0. The use of the tan-sigmoid function in the output neurons resulted in output values in the range [0.0, 1.0]: a value of less than 0.5 corresponded to 0, while a value greater or equal to 0.5 corresponded to 1. The MLP-NN was trained, using the training set, by the batched back-propagation (BP) algorithm with adaptive learning rate and momentum [33]. Moreover, the optimal MLP-NN classifier in terms of hidden neurons and internal NN parameters (appropriate values of momentum and initial learning rate) was determined using a trial-and-error process, until no further improvement of classification accuracy in the validation set could be obtained.

Finally, the neuron in the output layer classifies the ROI into the class with the highest probabilistic density function. The applied training procedure, in order to determine the value of the smoothing parameter and achieve high classification performance, is the same as in the case of the MLP-NN classifier.

2.4.3. k-Nearest neighbor classifier
The k-NN classifier is a very popular statistical classifier because of its simplicity and its easiness of development [11,34]. In the k-NN (k ≥ 1) classifiers, in order to classify a new feature vector, its k nearest neighbors from the training set are identified. The feature vector is classified to the most frequent class among its neighbors based on a similarity measure, like the Euclidian distance, which is used in the present study. In this paper, a 1-NN classifier along with two mk-NN (modified k-NN, k > 1, classifiers) mk1-NN and mk2-NN, have been developed. The mk-NN classifiers are based on the use of a distance-based weight. When a feature vector (X) is to be classified, the classifier finds the k nearest to X patterns (y1, y2, ..., yk), which belong to the training set, based on the Euclidian distance D. Each class Ci, i = 1, ..., 4, corresponding to the four types (C1, C2, C3, and C4) of hepatic tissue, contests the feature vector with a voting power wi:

\[
w_i = \sum_{j=1}^{k} D^{-1}(X, y_j) f(y_j, C_i), \quad \text{with } i = 1, ..., 4
\]

where function f is defined as

\[
f(y_j, C_i) = \begin{cases} 1, & \text{if } C_i = \text{class of } y_j \\ 0, & \text{else} \end{cases}
\]

The feature vector is finally declared into the class that contributes maximum weights in the neighbors.

The structures of mk1-NN and mk2-NN, which participate in the EC2, were determined by using the validation set. That means that the chosen value of k1, predefined in our study to be in the range 2 ≤ k1 ≤ 5, was the value of k1 for which the classification accuracy of the mk1-NN classifier was maximized in the validation set. Based on the same concept, the optimal value of k2, predefined to be within the range 6 ≤ k2 ≤ 9, was chosen.

2.4.4. Voting scheme
Once the primary classifiers of an EC have been generated, the issue of how to combine their predictions arises. Recently, the combination of
primary classifier outputs has become an important research field, since in many cases the ECs provide better performance than any of the component primary classifiers [35]. Among the techniques for combining the predictions obtained from the primary classifiers, voting is the most popular one [36]. A variety of voting schemes have been proposed, in order to improve the classification accuracy [37]. The approaches used in this study include a plurality, and a weighted voting scheme [18].

Figure 2  Architecture of (a) CAD1, (b) CAD2, (c) CAD3, (d) CAD4, and (e) CAD5.
2.5. CAD system architectures

Five alternative architectures (CAD1, ..., CAD5) were developed based on the generic design of the CAD system presented in Fig. 1. CAD1 and CAD2 were constructed using EC1, while CAD3, CAD4, and CAD5 were based on EC2. In CAD1 (Fig. 2(a)), each of the full-dimensional FOS, SGLDM, GLDM, TEM, and FDM feature sets, estimated in the feature extraction module, is fed into one of the five primary classifiers of EC1 (MLP-NN1, ..., MLP-NN5). CAD2 (Fig. 2(b)) differs from CAD1 in that feature selection is applied to the high-dimensionality feature vectors estimated from SGLDM, GLDM, and TEM. The selected features from the SGLDM, GLDM, and TEM sets are presented in Table 2. In CAD3 (Fig. 2(c)), each primary classifier uses as input the 89-dimensional feature set, which results from the combination of the full-dimensional FOS, SGLDM, GLDM, TEM, and FDM features sets. CAD4 (Fig. 2(d)) differs from CAD3 in that feature selection is applied to SGLDM, GLDM, and TEM feature sets prior to the combination with the full-dimensional FOS and FDM feature sets. The selected SGLDM, GLDM and TEM features are identical with the ones applied in CAD2 (Table 2). Thus, each primary classifier of EC2 in CAD4 uses as input a 30-dimensional feature set. CAD5 (Fig. 2(e)) differs from CAD4 in that further feature selection is applied to the 30-dimensional feature set used by CAD4. The resulting 12-dimensional feature set (Table 3) provides the input to each primary classifier of EC2 in CAD5. The final decision of each CAD system (CAD1, ..., CAD5) is generated by combining the individual predictions of the primary classifiers of EC1 and EC2 through either plurality or weighted voting schemes.

3. Results and discussion

The classification accuracies achieved by CAD1, ..., CAD5 using the 50 groups of sets (training, validation and testing set), obtained through bootstrap resampling, were estimated. For each architecture, the results are presented in terms of mean values and standard deviations of the classification accuracy of the primary classifiers, as well as the total CAD system performance with use of either plurality or weighted voting scheme.

The results for CAD1 and CAD2 are presented in Table 4. It can be observed that the use of FOS and TEM feature sets yielded to the best individual classification accuracies. In particular, the highest classification accuracy (78.43% / 8.23%) was obtained with the use of the FOS feature vector (MLP-NN1), followed by a 70.96% / 8.53% classification accuracy obtained with the use of the full TEM feature vector (MLP-NN4 of CAD1), and a 70.68% / 8.64% classification accuracy achieved by using the reduced TEM feature vector (MLP-NN4 of CAD2). FDM and full or reduced SGLDM feature sets yielded in general to worse individual classification accuracies in the bootstrap testing sets. The feature selection module in CAD2 did not alter significantly the performance of the primary classifiers. Thus, equally performing classifiers were obtained with less computational cost. Regarding the voting schemes, it can be observed that the use of FOS and TEM feature sets yielded to the best individual classification accuracies. In particular, the highest classification accuracy (78.43% ± 8.23) was obtained with the use of the FOS feature vector (MLP-NN1), followed by a 70.96% ± 8.53 classification accuracy obtained with the use of the full TEM feature vector (MLP-NN4 of CAD1), and a 70.68% ± 8.64% classification accuracy achieved by using the reduced TEM feature vector (MLP-NN4 of CAD2). FDM and full or reduced SGLDM feature sets yielded in general to worse individual classification accuracies in the bootstrap testing sets. The feature selection module in CAD2 did not alter significantly the performance of the primary classifiers. Thus, equally performing classifiers were obtained with less computational cost. Regarding the voting schemes, it can be observed that the weighted voting scheme outperformed the plurality voting scheme in the testing set. Similar observations were reported in the experiments done in [15]. However, it is notable that the best of the primary classifiers of CAD1 and CAD2 (MLP-NN1) performed slightly better than the ensemble using the weighted voting scheme. That is inline with many experimental results showing that the use of fixed combining rules (e.g., voting scheme) along with strong primary classifiers (i.e., classifiers which perform well by their own) results to ECs which do not alter significantly the performance of the primary classifiers. Thus, equally performing classifiers were obtained with less computational cost. Regarding the voting schemes, it can be observed that the weighted voting scheme outperformed the plurality voting scheme in the testing set. Similar observations were reported in the experiments done in [15]. However, it is notable that the best of the primary classifiers of CAD1 and CAD2 (MLP-NN1) performed slightly better than the ensemble using the weighted voting scheme. That is inline with many experimental results showing that the use of fixed combining rules (e.g., voting scheme) along with strong primary classifiers (i.e., classifiers which perform well by their own) results.
not outperform the best primary classifier of the ensemble [35,38]. The classification accuracies of CAD1 and CAD2 were assessed on the basis of generalization ability. Thus, CAD2 (77.32% ± 8.17%) using the weighted voting scheme was considered to be the optimal architecture.

The results for CAD3—CAD5 are presented in Table 5. Regarding these architectures, all primary classifiers resulted, in general, to high classification accuracies, due to the fusion of the input feature sets. More specifically, the best individual classification accuracy (84.21% ± 6.68%) was achieved by the mk1-NN classifier using the reduced feature set after a one-step feature selection (CAD4). The next best individual classification accuracy, 82.90% ± 6.27%, was achieved by the mk2-NN of CAD4, followed by the slightly lower performances of the P-NN and 1-NN of CAD4. The one-step feature selection improved the classification performance of the primary classifiers of CAD4, while further feature selection in the primary classifiers of CAD5 reduced the individual classification accuracies in the testing set. Thus, the 30-dimensional feature set used in CAD4 is the most robust among the feature sets used in the three architectures. As for the use of the voting schemes in CAD3, CAD4, CAD5 it can be observed that the use of either the plurality or weighted voting scheme in the testing set had almost identical effect, because most of the times the plurality of the outputs of the individual classifiers were identical, thus making the use of weights when combining these outputs of little importance. It can be assumed that the classifiers of EC2 which use identical input feature sets are more correlated in comparison with the classifiers of EC1, where the weighted voting scheme outperformed well some of the primary classifiers. Thus either the plurality or the weighted voting scheme of CAD3, CAD4 and CAD5 performed equally well or outperformed slightly the best of the primary classifiers of EC2.

Among the three systems, CAD4, using the weighted voting scheme is considered as the best one leading to a classification accuracy of 84.96% ± 5.67%.

<table>
<thead>
<tr>
<th>Classifier combination</th>
<th>Training set</th>
<th>Validation set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAD1</td>
<td>CAD2</td>
<td>CAD1</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td></td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>set</td>
<td></td>
<td>set</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP-NN1 (FOS)</td>
<td>97.44 ± 3.14</td>
<td>97.25 ± 3.14</td>
<td>86.38 ± 6.85</td>
</tr>
<tr>
<td>MLP-NN2 (SGLDM)</td>
<td>95.95 ± 15.91</td>
<td>96.33 ± 5.11</td>
<td>66.53 ± 7.5</td>
</tr>
<tr>
<td>MLP-NN3 (GLDM)</td>
<td>89.11 ± 5.31</td>
<td>71.57 ± 9.89</td>
<td>64.63 ± 7.14</td>
</tr>
<tr>
<td>MLP-NN4 (TEM)</td>
<td>97.56 ± 9.60</td>
<td>95.74 ± 14.60</td>
<td>83.09 ± 6.26</td>
</tr>
<tr>
<td>MLP-NN5 (FDM)</td>
<td>70.86 ± 9.47</td>
<td>70.86 ± 9.47</td>
<td>64.29 ± 6.51</td>
</tr>
</tbody>
</table>

Table 4 Mean and standard deviation values for individual and total classification accuracies of EC1 of CAD1, CAD2 in training, validation, and testing bootstrap sets

Plurality voting scheme

Weighted voting scheme

Weighted voting scheme in the testing set had a slight the best of the primary classifiers of EC2. Among the three systems, CAD4, using the weighted voting scheme is considered as the best one leading to a classification accuracy of 84.96% ± 5.67%.

Based on Tables 4 and 5, it is observed that the primary classifiers of EC2 classify better the liver regions, as compared to those of EC1. This is due to the greater dimension of the feature set which is fed into the classifiers of EC2, and that this set consists of different feature subsets, which interact with each other. The comparative assessment of CAD2 and CAD4, both using the weighted voting scheme, shows that the latter is the best performing architecture for discriminating the four liver tissue types.

The final choice of CAD4 using the weighted voting scheme is confirmed after its statistical comparison in terms of accuracy in bootstrap testing sets with all other CAD architectures. The results show that CAD4 using the weighted voting scheme performs better than all other CAD systems with either high statistical significance (p < 0.01), or medium statistical significance (p = 0.32) in case of comparison with CAD4 with plurality voting scheme.

Furthermore, the performance of CAD4 using the weighted voting scheme has been evaluated in the testing sets using the one-versus-all comparisons [39]. More specifically, the sensitivity and specificity when discriminating one type of liver tissue from the remaining three were estimated using each bootstrap testing set. The resulted sensitivities and specificities, in terms of mean values and corresponding standard deviations, are presented in Table 6. It can be observed that CAD4 using the weighted voting scheme is characterized by high sensitivity and specificity in the diagnosis of C1, C2, and C4, while a significantly lower sensitivity is achieved in the case of C3 due to misclassification of C3 into C1 (about 70% of misclassification cases) and C4. This can be explained by the similar texture values of the ROIs belonging to classes C1, C3 and C4, thus making the discrimination difficult for the ensemble of classifiers.
According to the above CAD4 using the weighted voting scheme is the most appropriate CAD for discriminating the four liver tissue types. CAD4 combines five different learning algorithms on the most robust feature set which resulted from the fusion of texture features obtained by five texture extraction methods.

It is important to note that the computational complexity during the development of the CAD systems refers mainly to the NN classifiers (MLP-NN, P-NN) training. More specifically, training the optimal NN classifier (on a Pentium III processor running at 1 GHz) using all bootstrap sets takes approximately 7—10 h for the MLP-NN, and 4—5 h for the P-NN, depending on the dimensionality of the input set. Once the NN classifiers have been trained and a new ROI has been defined, the extraction of the 30 textural features and the classification procedure demand very low computational power and are finished in short time (~5 s).

Currently, CAD4 using the weighted voting scheme is under clinical evaluation as a part of an integrated telematics-enabled system aiming at medical image archiving, management and diagnosis assistance [40]. The proposed CAD system can be extended using (i) more types of features (from shape analysis, etc.), (ii) other types of liver lesions, (iii) results of laboratory examinations and patient’s history, and (iv) liver images from other imaging modalities like ultrasound. Furthermore, in order to achieve a final decision characterized by higher classification accuracy, the use of more advanced combination strategies based on genetic programming will be investigated. It has to be noted that several other CAD architectures could be developed based on all possible combinations of texture features sets, application or not of feature selection, and classification methods. The exhaustive development and evaluation of the aforementioned combinations is in progress aiming to the identification of the impact of different features and classifiers in the diagnosis of focal liver lesions, and consequently the development of a CAD system with minimum number of features and maximum diagnosis performance.

### Table 5
Mean and standard deviation values for individual and total classification accuracies of EC2 of CAD3, CAD4, CAD5 in training, validation, and testing bootstrap sets

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Training set</th>
<th>Validation set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP-NN</td>
<td>99.74 ± 0.096</td>
<td>98.97 ± 4.73</td>
<td>98.75 ± 4.34</td>
</tr>
<tr>
<td>P-NN</td>
<td>98.11 ± 4.89</td>
<td>96.97 ± 3.11</td>
<td>96.68 ± 5.82</td>
</tr>
<tr>
<td>1-NN</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
</tr>
<tr>
<td>mk1-NN</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
</tr>
<tr>
<td>mk2-NN</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
<td>100.00 ± 0.00</td>
</tr>
</tbody>
</table>

### Table 6
Mean and standard deviation values of sensitivity and specificity measurements using the one-vs.-all approach for CAD4 with weighted voting scheme in the bootstrap testing sets

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-vs.-all</td>
<td>95.83 ± 5.45</td>
<td>81.93 ± 10.66</td>
</tr>
<tr>
<td>C2-vs.-all</td>
<td>98.60 ± 5.63</td>
<td>100.00 ± 0.00</td>
</tr>
<tr>
<td>C3-vs.-all</td>
<td>44.44 ± 26.60</td>
<td>97.14 ± 3.22</td>
</tr>
<tr>
<td>C4-vs.-all</td>
<td>81.57 ± 16.43</td>
<td>96.19 ± 3.82</td>
</tr>
</tbody>
</table>
4. Conclusion

The aim of the present paper was to define a CAD system architecture able to accurately classify hepatic tissue from non-enhanced CT images as normal liver, hepatic cyst, hemangioma, and hepatocellular carcinoma.

The system design was based on the use of texture features, feature selection techniques, and ECs. For each CT liver ROI, five types of texture feature sets, based on first order statistics, spatial gray level dependence matrices, gray level difference matrices, Laws’ texture energy measures, and fractal dimension measurements, were extracted resulting in a total of 89 features. A GA-based feature selection method was applied to feature sets, while dimensionality reduction was desired, while two alternative ECs were investigated consisting solely of NN classifiers, or of a combination of NN and statistical classifiers. A plurality and a weighted voting scheme were used to combine the outputs of the primary classifiers of the examined CAD systems. The optimal architecture of the CAD system was chosen based on the mean classification accuracies in 50 testing sets obtained through the bootstrap method. The best performing architecture achieved a mean classification accuracy equal to 84.96% in the bootstrap testing sets and used (a) a fused 30-dimensional feature set obtained after appropriate feature selection applied in specific subsets of the original 89-dimensional feature set, (b) an EC consisting of two NNs and three statistical classifiers, and (c) a weighted voting scheme.

The results indicate that CAD systems can provide a valuable "second opinion" tool for the radiologists in the differential diagnosis of focal liver lesions from non-enhanced CT images, thus improving the diagnosis accuracy, and reducing the needed validation time.

References

Differential diagnosis of CT focal liver lesions


