ORIGINAL ARTICLE

Computer-assisted liver graft steatosis assessment via learning-based texture analysis

Sara Moccia[1,2] · Leonardo S. Mattos[1] · Ilaria Patrini[2] · Michela Ruperti[2] · Nicolas Poté[3,4] · Federica Dondero[5] · François Cauchy[5] · Ailton Sepulveda[5] · Olivier Soubrane[5] · Elena De Momi[2] · Alberto Diaspro[6] · Manuela Cesaretti[5,6]

Abstract

Purpose Fast and accurate graft hepatic steatosis (HS) assessment is of primary importance for lowering liver dysfunction risks after transplantation. Histopathological analysis of biopsied liver is the gold standard for assessing HS, despite being invasive and time consuming. Due to the short time availability between liver procurement and transplantation, surgeons perform HS assessment through clinical evaluation (medical history, blood tests) and liver texture visual analysis. Despite visual analysis being recognized as challenging in the clinical literature, few efforts have been invested to develop computer-assisted solutions for HS assessment. The objective of this paper is to investigate the automatic analysis of liver texture with machine learning algorithms to automate the HS assessment process and offer support for the surgeon decision process.

Methods Forty RGB images of forty different donors were analyzed. The images were captured with an RGB smartphone camera in the operating room (OR). Twenty images refer to livers that were accepted and 20 to discarded livers. Fifteen randomly selected liver patches were extracted from each image. Patch size was 100×100. This way, a balanced dataset of 600 patches was obtained. Intensity-based features (INT), histogram of local binary pattern ($H_{\text{LBP}_{riu2}}$), and gray-level co-occurrence matrix (F_{GLCM}) were investigated. Blood-sample features (Blo) were included in the analysis, too. Supervised and semisupervised learning approaches were investigated for feature classification. The leave-one-patient-out cross-validation was performed to estimate the classification performance.

Results With the best-performing feature set ($H_{\text{LBP}_{riu2}} + \text{INT} + \text{Blo}$) and semisupervised learning, the achieved classification sensitivity, specificity, and accuracy were 95, 81, and 88%, respectively.

Conclusions This research represents the first attempt to use machine learning and automatic texture analysis of RGB images from ubiquitous smartphone cameras for the task of graft HS assessment. The results suggest that is a promising strategy to develop a fully automatic solution to assist surgeons in HS assessment inside the OR.

Keywords Liver · Transplantation · Texture analysis · Machine learning · Surgical data science

✉ Sara Moccia
sara.moccia@iit.it; sara.moccia@polimi.it

1 Department of Advanced Robotics (ADVR), Istituto Italiano di Tecnologia, Via Morego 30, 16136 Genoa, GE, Italy

2 Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milan, MI, Italy

3 Department of Pathology, Hôpital Beaujon, DHU UNITY, AP-HP, Clichy, France

4 INSERM UMR1149, Paris, France

5 Department of HPB Surgery and Liver Transplantation, Hôpital Beaujon, AP-HP, Clichy, France

6 Department of Nanophysics, Istituto Italiano di Tecnologia, Via Morego 30, 16136 Genoa, GE, Italy

Introduction

Liver transplantation (LT) is the treatment of choice for patients with end–stage liver disease for which no alternative therapies are available [4]. Due to increasing demand and shortage in organ supply, expanded donor selection criteria are applied to increase the number of grafts for LT. Since extended criteria donors generates augmented morbidity and mortality in recipient population, liver graft quality assessment is crucial.

Hepatic steatosis (HS) is one of the most important donor characteristic that can influence graft function and so LT outcome, mostly because of severe ischemia reperfusion injury [13]. Defined as the intracellular accumulation of triglycerides resulting in the formation of lipid vesicles in the hepatocytes, HS is commonly assessed by histopathological examination of liver tissue samples extracted with biopsy. Through visually analyzing the quantities of large sized lipid droplets in the sample, an HS score is assigned to the sample in a semiquantitave fashion. Livers classified as with 5–30% fatty infiltration are associated with decreased patient and graft survival, but are still considered suitable for transplantation due to the limited donor availability [20]. Severe HS ($\geq 60\%$) is instead associated with primary graft dysfunction or non-function and is not compatible with transplantation [6,29].

Despite histopathological analysis of biopsied liver tissue being currently the gold reference standard for diagnosis and grading of HS in liver grafts, it is invasive, time-consuming, and expensive. Due to the short time availability between liver procurement and transplantation [24], the surgeon usually performs HS assessment through clinical evaluation (medical history, blood tests) and qualitative visual graft assessment [31]. In this context, visual liver texture analysis is recognized as crucial in grading HS [31]: livers that cannot be transplanted due to high HS (Fig. 1, right) are usually characterized by inhomogeneous texture and are more yellowish than the transplantable ones (Fig. 1, left). It is nonetheless recognized that the precise estimation of HS remains challenging even in experienced hands [31].

On this background, the development of a robust, quantitative, practical, cost–effective and rapid method to support the surgeon in deciding whether to accept or discard liver grafts is mandatory. Considering challenges in diagnosis, preliminary efforts to the automated or semiautomated HS assessment have been proposed and a complete review can be found in [9]. Examples include [28], which reported a sensitivity (Se) of 79% in recognizing the HS level from computer-tomography (CT) images, and [19] which reported an area under the receiving operating characteristic curve of 75% by exploiting fibroscanning. Liver-bioelectrical-impedance analysis and Raman spectroscopy were used in [1,11], respectively. A semiautomatic HS-grading approach that exploits magnetic resonance (MR) spectroscopy has been proposed in [26], achieving a Spearman correlation coefficient of 0.90.

It is worth noting that all the proposed methodologies require additional imaging instrumentation, which may be not always available in the remote graft procurement hospitals. Moreover, at most the methods concluded that there is a correlation between liver physical characteristics (e.g., liver stiffness and impedance) and HS grade, without providing a solution for liver graft quality assessment.

Despite visual liver texture analysis being crucial for clinical HS assessment [31], to the best of authors' knowledge no efforts have been done to develop a computer-assisted diagnostic tool that exploits automatic texture analysis to assess graft steatosis. Moreover, liver texture analysis has the advantage of being performed on standard RGB optical imaging, without requiring additional instrumentations. It is worth noting that modern cellphone cameras provide decent quality images for liver assessment and are ubiquitous. Therefore, they could be the solution for automatic HS assessment not only in remote hospitals, but also in low-income countries where other imaging equipment may not be available. Indeed, the use of RGB cameras for tissue classification is becoming quite popular in different fields, such as skin cancer diagnosis [8].

The emerging and rich literature on surgical data science for tissue classification in optical images outside the field of HS assessment is focusing more and more on using machine learning algorithms to classify tissues according to texture-based information [18]. The histogram of local binary patterns (LBP) is exploited for tissue classification in several anatomical districts, such as abdomen, larynx, gastro-intestinal tract (e.g., [15,16,22,23,32]). Gray-level co-occurrence matrix (GLCM)-based features [10] have also been exploited for tissue classification. Examples include [30] for gastroscopy and [21] for colorectal images.

Inspired by these recent and promising studies, and in particular by our previous research focused on the laryngeal district [22], in this paper we aim at investigating whether liver texture analysis from RGB images acquired with smartphones in the operating room (OR) coupled with machine learning can provide reliable results, to be used as support for LT decision.

This paper is organized as follows: "Methods" section explains the proposed approach to textural feature extraction and classification. The results are presented in "Results" section and discussed in "Discussion" section, reporting the main strengths and drawback of the proposed approach and suggesting future work to overcome the drawbacks. To conclude, "Conclusion" section summarizes the contribution of this paper.

Methods

In this section, the feature extraction strategy is explained ("Feature extraction" section) as well as the classification model training ("Model training" section). We will explore both supervised ("Supervised approaches for patch classification" section) and semisupervised ("Semisupervised approach for image classification" section) classification approaches. The evaluation protocol, which includes materials, parameter setting, and performance measure definition,
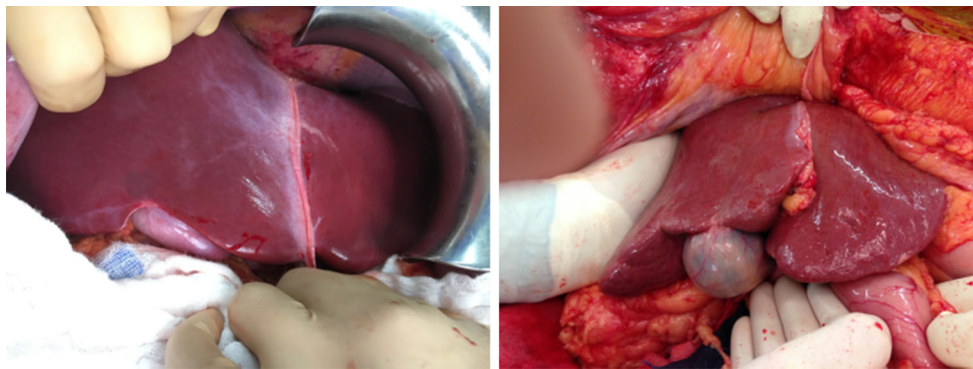
Fig. 1 Sample RGB liver images acquired in the operating room. Images are captured with different lighting conditions and different tissue-camera pose. Images refer to a transplanted-liver graft (left) and a discarded one (right)
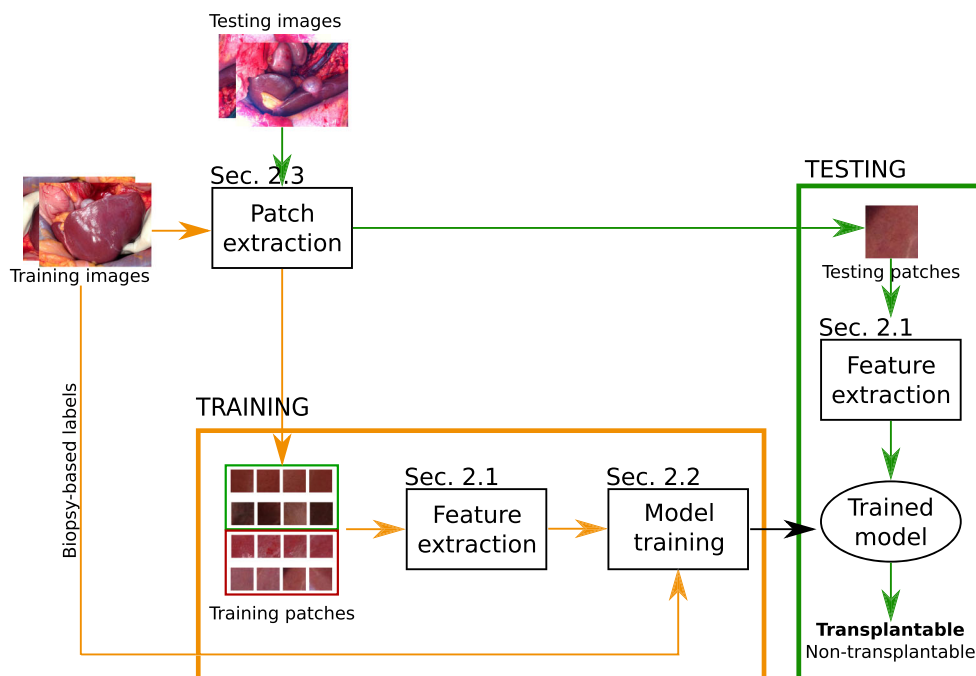
Fig. 2 Proposed workflow for graft hepatic steatosis assessment. From 40 RGB liver images of 40 different donors, a dataset of patches with size 100×100 is extracted. From each patch, a set of textural features is computed. The dataset is divided in training and testing patches. The features from the training patches are used to train a classifier model. The trained model is used to assess HS from the testing patches

is explained in "Evaluation" section. The workflow of the proposed method for LT assessment is shown in Fig. 2.

Feature extraction

When choosing the classification features, it is necessary to consider that liver images may be captured under various illumination conditions and from different viewpoints. As a consequence, the features should be robust to the camera pose as well as to the lighting conditions. Furthermore, with a view of a real-time computer-aided application, they should be computationally cheap. LBPs fully meet these requirements [25].

A rather popular LBP formulation is the uniform rotation-invariant one ($LBP_{riu2}^{R,P}$). The $LBP_{riu2}^{R,P}$ formulation requires to define, for a pixel $\mathbf{c} = (c_x, c_y)$, a spatial circular neighborhood of radius R with P equally spaced neighbor points ($\{\mathbf{p}_p\}_{p \in (0, P-1)}$):

$$
LBP_{riu2}^{R,P}(\mathbf{c})
= \begin{cases} \sum_{p=0}^{P-1} s(g_{\mathbf{p}_p} - g_{\mathbf{c}}), & \text{if } U(LBP^{R,P}) \leq 2 \\ P + 1, & \text{otherwise} \end{cases} \tag{1}
$$

where $g_{\mathbf{c}}$ and $g_{\mathbf{p}_p}$ denote the gray values of the pixel $\mathbf{c}$ and of its pth neighbor $\mathbf{p}_p$, respectively. $s(g_{\mathbf{p}_p} - g_{\mathbf{c}})$ is defined as:

$$s(g_{\mathbf{p}_p} - g_{\mathbf{c}}) = \begin{cases} 1, & g_{\mathbf{p}_p} \geq g_{\mathbf{c}} \\ 0, & g_{\mathbf{p}_p} < g_{\mathbf{c}} \end{cases} \tag{2}$$

and $U(\mathrm{LBP}^{R,P})$ is defined as:

$$U(\mathrm{LBP}^{R,P}) = |s(g_{\mathbf{p}_{P-1}} - g_{\mathbf{c}}) - s(g_{\mathbf{p}_0} - g_{\mathbf{c}})| \\ + \sum_{p=1}^{P-1} |s(g_{\mathbf{p}_p} - g_{\mathbf{c}}) - s(g_{\mathbf{p}_{p-1}} - g_{\mathbf{c}})| \tag{3}$$

Here, the $H_{\mathrm{LBP}_{riu2}}$, which counts the occurrences of $\mathrm{LBP}_{riu2}^{R,P}$, was used as textural feature and normalized to the unit length.

To include image intensity information, which has been reported as related to the HS level from the clinical community [31], we also included intensity-based features (INT), which consisted of image mean and standard deviation, computed for each RGB channel in the image.

For comparison, we also extracted the GLCM matrix-based textural features. The GLCM computes how often pair of pixels $(\mathbf{c}, \mathbf{q})$ with specific values and in a specified spatial relationship (defined by θ and d, which are the angle and distance between $\mathbf{c}$ and $\mathbf{q}$) occur in an image. In the GLCM formulation, the GLCM width (W) is equal to the GLCM height (H) and corresponds to the number of quantized image intensity gray levels. For the $w = h$ intensity gray level, the GLCM computed with θ and d is defined as:

$$\mathrm{GLCM}_{\theta,d}(h, w) \\ = \begin{cases} 1, & I(\mathbf{c}) = h \text{ and} \\ & I(c_x + d \cdot \cos(\theta), \ c_y + d \cdot \sin(\theta)) = w \\ 1, & I(\mathbf{c}) = h \text{ and} \\ & I(c_x - d \cdot \cos(\theta), \ c_y - d \cdot \sin(\theta)) = w \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We extracted GLCM-based features from the normalized $\mathrm{GLCM}_{\theta,d}$, which expresses the probability of gray-level occurrences, is obtained by dividing each entry of the $\mathrm{GLCM}_{\theta,d}$ by the sum of all entries, as suggested in 10. The GLCM feature set (F_{GLCM}) consisted of GLCM contrast, correlation, energy and homogeneity.

As in [22], and since texture is a local image-property, we decided to compute textural features from image patches, which were extracted as explained in "Evaluation" section.

Model training

In this section, we will first describe our approach for supervised patch classification ("Supervised approaches for patch classification" section). In "Semisupervised approach for image classification" section we will deal with the semisupervised approach for image classification.

Supervised approaches for patch classification

To perform patch classification, support vector machines (SVM) with the Gaussian kernel (Ψ) were used [3]. Indeed, SVM allowed overcoming the *curse-of-dimensionality* that arises analyzing our high-dimensional feature space [5,17]. The *kernel-trick* prevented parameter proliferation, lowering computational complexity and limiting over-fitting. Moreover, as the SVM decisions are only determined by the support vectors, SVM are robust to noise in training data. For our binary classification problem, given a training set $\mathrm{T} = \{y_t, \mathbf{x_t}\}_{t \in \mathrm{T}}$, where $\mathbf{x_t}$ is the tth input feature vector and y_t is the tth output label, the SVM decision function (f), according to the "dual" SVM formulation, takes the form of:

$$f(\mathbf{x}) = \mathrm{sign}\left[\sum_{t \in \mathrm{T}} a_t^* y_t \Psi(\mathbf{x}, \mathbf{x_t}) + b \right] \tag{5}$$

where

$$\Psi(\mathbf{x}, \mathbf{x_t}) = \exp\left\{ -\gamma \|\mathbf{x} - \mathbf{x_t}\|_2^2 / \sigma^2 \right\}, \quad \gamma > 0 \tag{6}$$

b is a real constant and a_t^* is computed as follow:

$$a_t^* = \max\left\{ -\frac{1}{2} \sum_{k,l=1}^{N} y_t y_l \Psi(\mathbf{x_t}, \mathbf{x_l}) a_t a_l + \sum_{t \in \mathrm{T}} a_t \right\} \tag{7}$$

with:

$$\sum_{t \in \mathrm{T}} a_t y_t = 0, \quad 0 \leq a_t \leq C, \quad t \in \mathrm{T} \tag{8}$$

In this paper, γ and C were retrieved with grid search and cross-validation on the training set, as explained in "Evaluation" section.

For the sake of completeness, the performance of random forest (RF) [2] in classifying image patches was also investigated.

Semisupervised approach for image classification

After performing the patch classification with SVM and RF, we identified the best-performing feature set as the one that guarantees the highest Se. With the best-performing feature set, we further investigated the use of multiple instance learning (MIL), a semisupervised machine learning technique, for performing full image classification (instead of patch classification) from patch-based features. In fact, it is worth noting that the pathologist gold-standard biopsy-based classification

is associated to the whole image, and not to the single patch. Thus, considering all patches from an image of a graft with high HS as pathological may influence the classification outcome, as HS is commonly not homogeneous in the hepatic tissue [31]. Therefore, we decided to investigate whether MIL can support HS diagnosis from (unlabeled) patches extracted from (labeled) RGB images.

Among MIL algorithms, we investigated the use of single instance learning (SIL) [27], which has the strong advantage of allowing the fusion of patch-wise information (such as textural features) with image-wise information (such as blood-sample features) [27], thus providing further information for the classification process. Here, we decided to investigate the popular SVM-SIL formulation, which showed good classification performance in several fields outside the proposed one [27].

For our semisupervised binary classification problem, let $T_p \subseteq T$ be the set of positive images (rejected grafts), and $T_n \subseteq T$ the set of negative images (accepted grafts). Let $\widetilde{T}_p = \{t \mid t \in T \in T_p\}$ and $\widetilde{T}_n = \{t \mid t \in T \in T_n\}$ be the patches from positive and negative images, respectively. Let $L = L_p + L_n = \mid \widetilde{T}_p \mid + \mid \widetilde{T}_n \mid$ be the total number of patches. For any patch $t \in T$ from an image $T \in$ T, let $\mathbf{x}_t$ be the feature vector representation of t. Thus, $\mathbf{x_T} = \sum_{t \in T} \mathbf{x_t}$ is the feature vector representation of image T. The SVM-SIL optimization, here written in the "primal" SVM formulation for better readability, aims at minimizing $\mathbf{J}$:

$$\mathbf{J}(\mathbf{w}, b, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{L} \sum_{X \in \chi} \sum_{x \in X} \xi_x \tag{9}$$

subject to:

$$\mathbf{w}\boldsymbol{x_t} + b \le -1 + \xi_t, \qquad \forall t \in \widetilde{T}_n \tag{10}$$

$$\mathbf{w}\boldsymbol{x_t} + b \ge +1 - \xi_t, \qquad \forall t \in \widetilde{T}_p \tag{11}$$

$$\xi_t \ge 0 \tag{12}$$

where ξ_t is the relaxing term introduced for the soft-margin SVM formulation, b is a real value, $\mathbf{w}$ the SVM weight vector. Also in this case, C was retrieved with grid search and cross-validation on the training set, as explained in "Evaluation" section.

As SIL allows fusing patch-wise (i.e., texture features) and image-wise (i.e., blood features) features, in addition to the best-performing feature set, features from blood tests (Blo) were used, too. In particular, alanine aminotransferase, aspartate aminotransferase, bilirubin, liver Hounsfield unit (HU), difference between the liver and the spleen HU, and gamma glutamyl transferase were considered. Further, patient' age, weight and height were also considered. Thus, Blo feature size was 9. The Blo features are commonly used for HS assessment by surgeons [31], as introduced in "Introduction"

section. Thus, their use would not alter the actual clinical practice.

Evaluation

In this study, we analyzed 40 RGB images, which refer to 40 different potential liver donors. HS was assessed with histopatological analysis performed after liver biopsy.

Biopsy was performed during procurement, taking surgical triangular hepatic samples up to 2 cm. One pathologist analyzed the histological sections. Steatosis was visually assessed based on the percentage of hepatocytes with intracellular large lipid droplets by using a semicontinuous scale [0:5:100%].

From the dataset, 20 livers referred to discarded grafts, as with a HS $\ge$ 60%. The remaining 20 livers had a HS $\le$ 20% and were transplanted. Images were acquired with a smartphone RGB camera. Image size was 1920×1072 pixels. All the images were acquired with open-surgery view, as no laparoscopic procurement is performed for cadaveric donors [14]. Challenges associated with the dataset included:

- Wide range of illumination
- Varying camera pose
- Presence of specular reflections
- Different organ position

Visual samples of liver images are shown in Fig. 3.

From each image, liver manual segmentation was performed to separate the hepatic tissue from the background (Fig. 4). The manual segmentation of the liver images was performed with the help of the software MATLAB ®. The liver contour in each image was manually drawn by marking seed points along the lived edges, which were then connected with straight lines by the software.

The whole image was then divided in non-overlapping patches of size 100×100 pixels starting from the top-left image corner. We chose such patch size as image-patch size is usually of the order of magnitude of $10^2 \times 10^2$ pixels (e.g., [32]). The most right part of the image, for which it was not possible to select full patches, was discarded. This did not represent a problem since the liver was always displayed at the center of the image. A patch was considered valid for our analysis if it overlapped with at least 90% of the liver mask.

To have the same number of patches from each patient, we first computed the minimum number of image patches that we could obtain among all images, which was 15. Then, we randomly extracted 15 patches from all the other images. As result, our patch dataset was composed of 300 patches extracted from transplanted liver and 300 from non-transplanted ones. Sample patches for transplanted and non-transplanted livers are shown in Fig. 5.
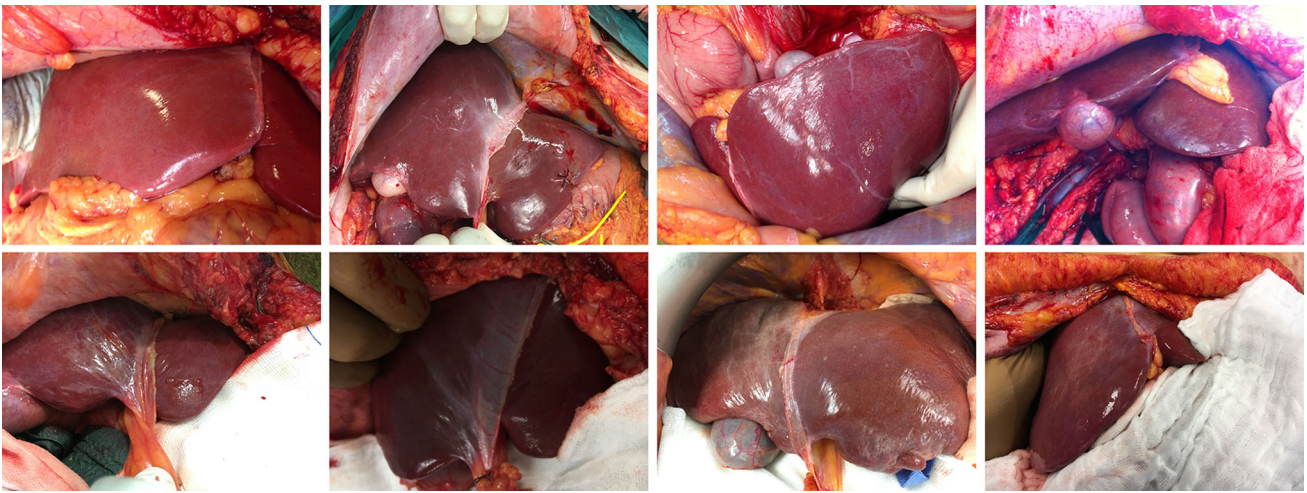
Fig. 3 Dataset sample images. The images refer to (first row) accepted and (second row) rejected liver grafts. Images were acquired at different distance and orientation with respect to the liver. Images present different illumination levels. Specular reflections are present due to the smooth and wet liver surface
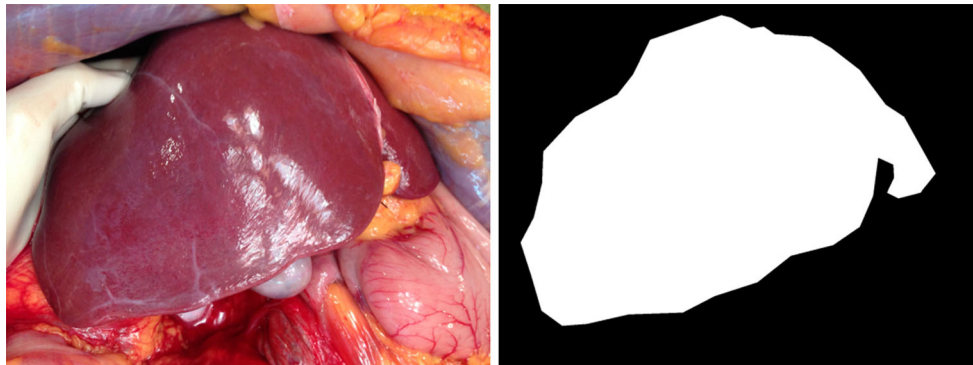
Fig. 4 Liver and liver mask obtained through manual segmentation

Fig. 5 Dataset sample patches. The green and red boxes refer to patches extracted from transplanted and non-transplanted livers. Each row in a box refers to patches extracted from the same liver image
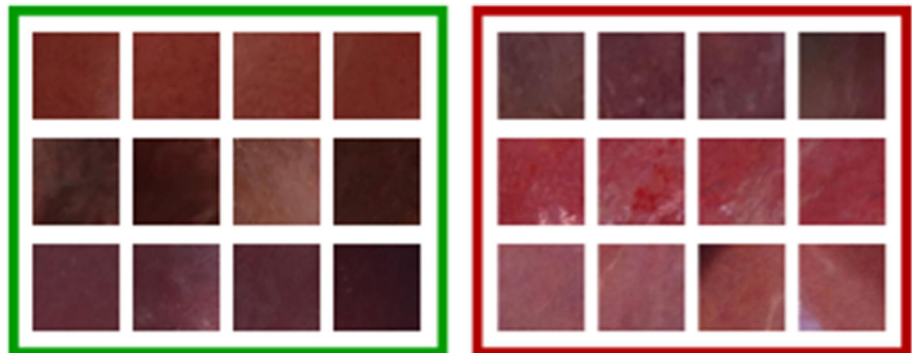
For the feature extraction described in "Feature extraction" Section, the $LBP_{R,P}^{riu2}$ were computed with the following $(R; P)$ combinations: $(1; 8)$, $(2; 16)$, $(3; 24)$, and the corresponding $H_{LBP_{riu2}}$ were concatenated. Such choice allows a multi-scale, and therefore a more accurate description of the texture, as suggested in 22. The $LBP_{R,P}^{riu2}$ were computed for each RGB image channel.

Nine $GLCM_{\theta,d}$ were computed for each RGB channel using all the possible combinations of (θ, d), with $\theta \in \{0°, 45°, 90°\}$ and $d \in \{1, 2, 3\}$, and the corresponding F_{GLCM} sets were concatenated. The chosen interval of θ allows to approximate rotation invariance, as suggested in 10. The values of d were chosen to be consistent with the scale used to compute $LBP_{R,P}^{riu2}$.

Prior to classification, we also investigated feature reduction by means of principal component analysis (PCA). In particular, we applied PCA on our best-performing (highest Se) feature set. We then retained the first principal components with explained variance equal to 98% and performed the classification described in "Supervised approaches for patch classification" section.

For performing the classification presented in "Model training" section, the SVM hyper-parameters, for both the supervised and semisupervised approaches, were retrieved via grid search and cross-validation on the training set. The grid search space for γ and C was set to $[10^{-10}, 10^{-1}]$ and $[10^{0}, 10^{10}]$, respectively, with 10 values spaced evenly on $\log_{10}$ scale in both cases. The number of trees for RF training was retrieved with a grid search space set to $[40,100]$ with six values spaced evenly.

The feature extraction and classification were implemented with scikit-image[1] and scikit-learn.[2]

We evaluated the classification performance of SVM, RF and SVM-SIL using leave-one-patient-out cross-validation. Each time, the patches extracted from one patient were used for testing the performance of the classification model (SVM, RF or SVM-SIL) trained with (only) the images of all the remaining patients. The separation at patient level was necessary to prevent data leakage.

To evaluate the classification performance, we computed the classification (Se), specificity (Sp), and accuracy (Acc), where:

$$Se = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

$$Acc = \frac{TP + TN}{TN + FP + TP + FN} \tag{15}$$

being TP, TN, FP and FN the number of true positive, true negative, false positive and false negative, respectively.

We used the Wilcoxon signed-rank test (significance level $\alpha = 0.05$) for paired sample to assess whether the classification achieved with our best-performing (highest Se) feature vector significantly differs from the ones achieved with the other feature sets in Table 1.

Results

Table 2 shows the area under the ROC for the SVM classification obtained with the feature vectors in Table 1. The higher

---

Table 1 Tested feature vectors for supervised patch classification with support vector machines and random forest and corresponding number of features

Feature vector	INT	F_{GLCM}	$H_{LBP_{riu2}}$	$H_{LBP_{riu2}} + INT$
Number of features	6	108	162	168

INT Intensity mean and variance, F_{GLCM} gray-level co-occurrence matrix-based descriptors, $H_{LBP_{riu2}}$ normalized histogram of rotation-invariant uniform local binary patterns

Table 2 Area under the receiving operating characteristic (ROC) curve obtained with support vector machines (SVM) and different feature vectors

	Area under the ROC
INT	0.65
F_{GLCM}	0.71
$H_{LBP_{riu2}} + INT$	0.76
$H_{LBP_{riu2}}$	0.77

INT intensity mean and variance, F_{GLCM} Gray-level co-occurrence matrix-based descriptors, $H_{LBP_{riu2}}$ Normalized histogram of rotation-invariant uniform local binary patterns
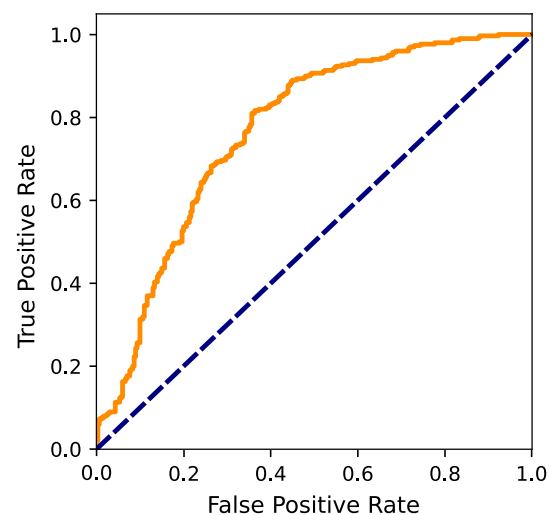
Fig. 6 Receiving operating characteristic (ROC) curve for classification with the normalized histogram of rotation-invariant uniform local binary patterns and support vector machines

area under the ROC (0.77) was obtained with $H_{LBP_{riu2}}$. The relative ROC curve is shown in Fig. 6.

The classification performance obtained with SVM and INT, F_{GLCM}, $H_{LBP_{riu2}}$ and $H_{LBP_{riu2}} + INT$ is shown in Table 3. The best performance was obtained with $H_{LBP_{riu2}}$, with Se = 0.82, Sp = 0.64 and Acc = 0.73. Using only INT features led to the worst classification performance for rejected grafts with Se = 0.58. Significant differences were found when comparing our best-performing feature ($H_{LBP_{riu2}}$) with INT and F_{GLCM}. The confusion matrices for feature comparison are reported in Fig. 7.

Table 3 Patch classification performance measure obtained with support vector machines (SVM) and different feature vectors

	Se	Sp	Acc
INT	0.58	0.80	0.69
F_{GLCM}	0.73	0.60	0.66
$H_{LBP_{riu2}} + INT$	0.80	0.64	0.72
$H_{LBP_{riu2}}$	0.82	0.64	0.73

INT Intensity mean and variance, F_{GLCM} Gray-level co-occurrence matrix-based descriptors, $H_{LBP_{riu2}}$ Normalized histogram of rotation-invariant uniform local binary patterns, *Se* sensitivity, *Sp* specificity, *Acc* accuracy

Table 4 Patch classification performance measure obtained using the histogram of rotation-invariant local binary patterns as feature with different classifiers

	Se	Sp	Acc
SVM + PCA	0.83	0.62	0.73
RF	0.72	0.61	0.67
SVM	0.82	0.64	0.73

Classification with feature reduction is reported, too. *Se* Sensitivity, *Sp* specificity, *Acc* accuracy, *PCA* principal component analysis, *SVM* support vector machines, *RF* random forest

When exploiting PCA-based feature reduction for $H_{LBP_{riu2}}$, Se = 0.83, Sp = 0.62, and Acc = 0.73 were obtained (Table 4). No significant differences with respect to the case without feature selection were found, and therefore we decided to avoid using PCA to keep the overall algorithm computational cost low. Similar results were achieved also with $H_{LBP_{riu2}} + INT$.

When classifying $H_{LBP_{riu2}}$ with RF, Se = 0.72, Sp = 0.61, and Acc = 0.67 were obtained (Table 4). Significant differences with respect to SVM performance were not found.
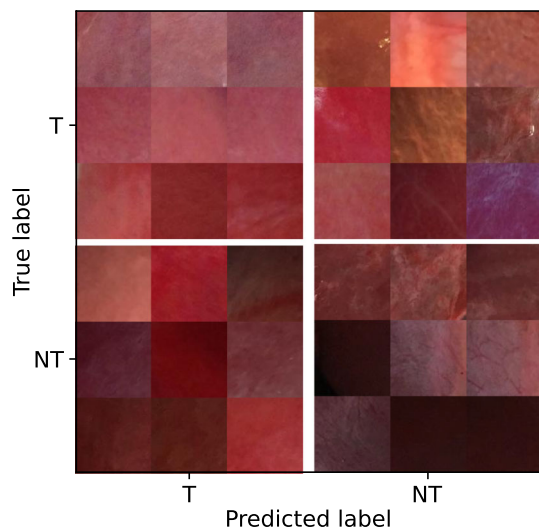
The visual confusion matrix for the patch classification performed with SVM and $H_{LBP_{riu2}}$ is shown in Fig. 8.

From our patch-based experimental analysis, among all the tested feature sets, $H_{LBP} + INT$ and H_{LBP} were the best-performing feature sets. Thus, we decided to test SVM-SIL with these two feature vectors, including also the Blo features as introduced in "Semisupervised approach for image classification" section. The features investigated for SVM-SIL classification and the correspondent number of features are reported in Table 5. With SVM-SIL, $H_{LBP} + INT + Blo$ showed the best classification performance, with Se = 0.95, Sp = 0.81, and Acc = 0.88. When using $H_{LBP} + Blo$, Acc = 0.82 was achieved. The confusion matrix for the

Fig. 7 Confusion matrices (CMs) for the classification of image patches from transplanted (T) and non-transplanted (NT) liver graft images. **a** CM for gray-level co-occurrence-based features (F_{GLCM}). **b** CM for intensity-based features (INT). **c** CM for intensity-based and local binary patter features ($H_{LBP_{riu2}} + INT$). **d** CM local binary patter-based features ($H_{LBP_{riu2}}$). CMs were obtained with support vector machines
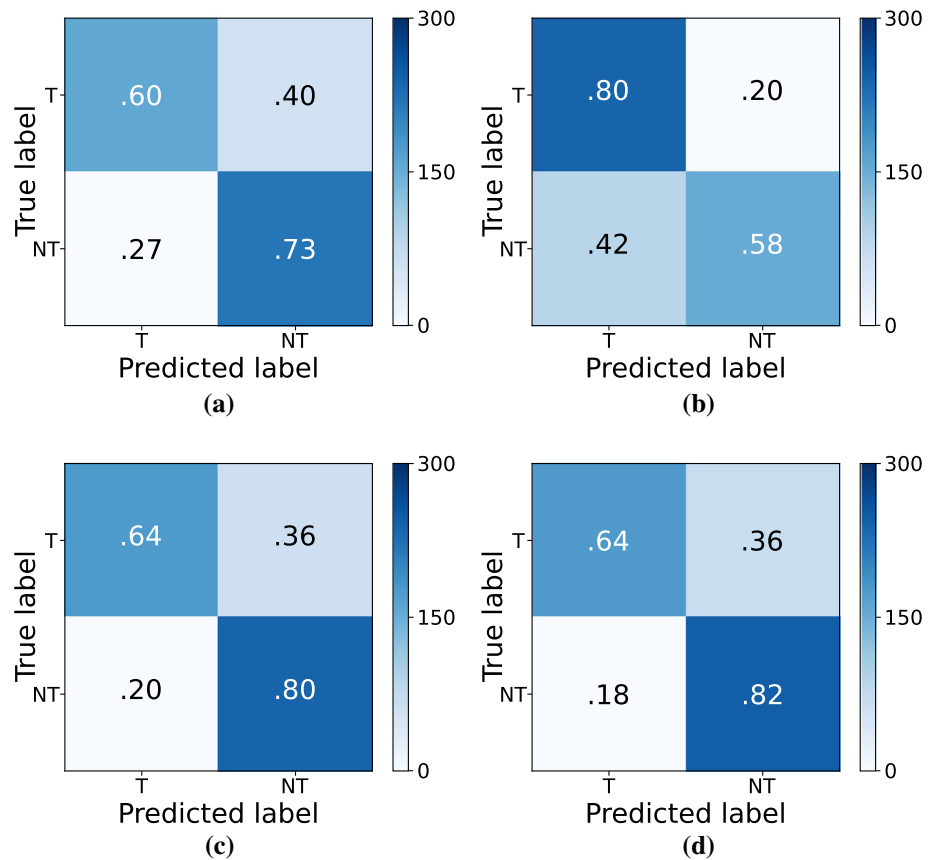
Fig. 8 Visual confusion matrices for supervised patch classification obtained with the histogram of rotation-invariant uniform local binary patterns and support vector machines. *NT* Non-transplanted patches, *T* transplanted patches

Table 5 Tested feature vectors for semisupervised image classification with single instance learning-support vector machines and corresponding number of features

| Feature vector	$H_{\text{LBP}_{riu2}} + \text{INT}$	$H_{\text{LBP}_{riu2}} + \text{INT} + \text{Blo}$
Number of features	168	177

INT Intensity mean and variance, $H_{\text{LBP}_{riu2}}$ Normalized histogram of rotation-invariant uniform local binary patterns, *Blo* blood features

SVM-SIL classification for $H_{\text{LBP}} + \text{INT} + \text{Blo}$ is reported in Fig. 9. Visual samples of liver classification outcomes with $H_{\text{LBP}_{riu2}} + \text{INT} + \text{Blo}$ for SVM-SIL classification are shown in Fig. 10. It is worth noting that SVM-SIL failed in classifying images of rejected liver grafts only once.

As for the algorithm computational cost, the liver manual segmentation took $\sim$ 3 s on average per image. The classification process (both for SVM, RF and SVM-SIL) took $\sim 10^{-5}$ s. The time for the computation of $H_{\text{LBP}} + \text{INT}$ for one patch was ~ 0.02 s. Experiments were performed on a CPU Intel® Core™i7-3537U @ 2.0 GHz x 4 with 7.6 GB of available RAM; Linux operative system, kernel 4.4.0-98-generic (x86_64) Ubuntu 16.04.3 LTS distribution.

Discussion

In this paper, we presented and fully evaluated an innovative approach to the computer-aided assessment of HS in RGB images acquired with smartphones in the OR, which exploits liver texture analysis coupled with machine learning. With respect to the approaches in the literature, our method only requires RGB images and blood-sample tests. More-
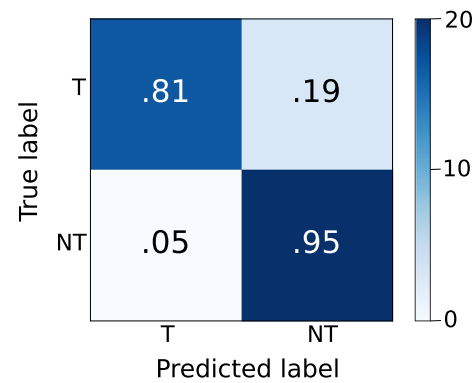
Fig. 9 Confusion matrix (CM) for the classification of transplanted (T) and non-transplanted (NT) liver graft images. CM are obtained with $H_{\text{LBP}} + \text{INT} + \text{Blo}$. The classification is performed with support vector machines (SVM)-single instance learning (SVM-SIL). The colorbar indicates the number of correctly classified images

over, it provides the surgeons with a classification outcome on whether to accept or discard a liver graft.

For our experimental analysis, the highest (supervised) patch classification performance was obtained with $H_{\text{LBP}_{riu2}}$ and $H_{\text{LBP}_{riu2}} + \text{INT}$, which performed equally. F_{GLCM} performed worse and this is probably due to the GLCM lack of robustness to illumination condition changes. In fact, when acquiring liver images, no assumption on keeping the illumination constant was done, resulting in different levels of illumination in the images. Similarly, also INT features were not able to face such variability in the illumination.

Classification performance with and without PCA did not differ significantly. Therefore, we decided to avoid performing PCA feature reduction for lowering the algorithm computational cost with a view to real-time applications.

Significant differences were not found when comparing RF and SVM performance. This is something expected, if one compares our results with the literature (e.g., [7,22]).

By visually inspecting the wrongly classified patches (Fig. 8), it emerged that misclassification occurred for patches that are challenging to classify also for the human eye. In fact, images were acquired without a controlled acquisition protocol, making the classification not trivial.

SVM-SIL provided a more reliable and robust classification with respect to supervised approaches, both in terms of Se and Sp. In fact, SVM-SIL misclassified a rejected liver image only once. This can be attributed to the fact that SIL did not make any assumption on ground-truth patch labels, but only exploited the ground-truth classification of the images obtained through histopathology for training purposes. The inclusion of blood-test features helped increasing the classification accuracy with respect to using only textural features. Nonetheless, it is worth noting that Blo alone was not sufficient to achieve accurate HS diagnosis. Indeed, during our preliminary analysis, we achieved an Acc = 0.75 with super-
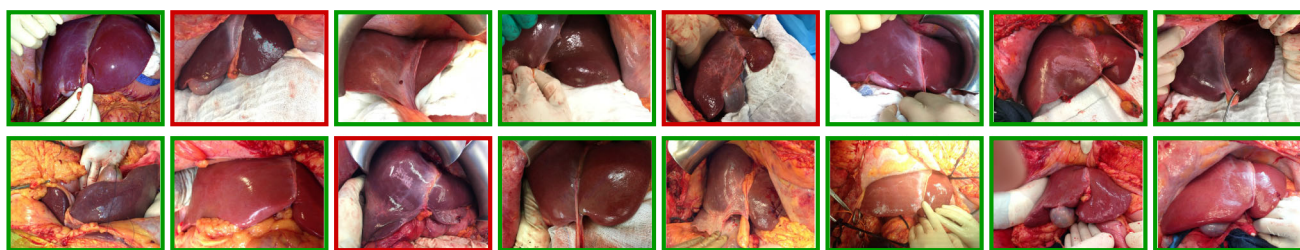
Fig. 10 Samples of classification outcomes for transplanted (first row) and non-transplanted (second row) liver grafts. Classification refers to support vector machines (SVM)—single instance learning (SIL) with $H_{\mathrm{LBP}_{riu2}} + \mathrm{INT} + \mathrm{Blo}$. The green and red boxes refer to correct and wrong classification outcomes, respectively. SVM-SIL wrongly classified a rejected liver only once

vised SVM-based classification when considering Blo alone. This supports the hypothesis that textural-information inclusion is a valid support for HS diagnosis.

The computational time required by the proposed method was not compatible with real-time application due to the computational cost associated with the liver manual segmentation. To reduce the computational cost and make the process more automatic, color thresholding could be used to segment the liver. Nonetheless, in this paper, we decided to perform manual liver masking to keep the experimental setup as controlled as possible with the goal of investigating the potentiality of machine learning in the analyzed context.

A direct comparison with the state of the art results was not possible, due to the lack of benchmark datasets. Moreover, as reported in "Introduction" section, the methods in the literature only correlated hepatic physical characteristics with the HS level, not providing a method for graft quality assessment.

A limitation of the proposed study could be seen in its patch-based nature, even though this is something commonly done in the literature for hepatic tissue assessment (e.g., [26]). We decided to work with patches to have a controlled and representative dataset to fairly evaluate different features.

Moreover, due to the size of our dataset, we decided to perform leave-one-patient-out cross-validation for the algorithm evaluation. Despite leave-one-patient-out cross-validation being a well-established method for performance evaluation based on a small set of samples, it could provide classification performance overestimation [12]. Therefore, to have a robust estimation of the classification performance, it would be necessary to evaluate the classification method with different sets of images never used for training. Nonetheless, as to do so, a bigger dataset would be required. Thus, as future work, we aim at enlarging the training dataset, exploiting also different RGB camera devices, to validate the experimental analysis presented here. We also aim at investigating if including a measure of confidence on classification, such as in [22,23], could help further improving classification reliability.

Conclusion

In conclusion, the most significant contribution of this work is showing that LBP-based features and SVM-SIL, along with blood-sample tests, can be used as support for HS assessment. This is highly beneficial for practical uses as the method can be potentially developed to run in real-time, being compatible with the short time available between the time of liver procurement and the LT. Moreover, the only required imaging source is a standard RGB camera, which can be easily used in the OR without requiring additional imaging sources such as MRI or Raman spectrometer.

It is acknowledged that further research is required to further ameliorate the algorithm as to offer all possible support for diagnosis and achieve classification performance comparable with those obtained with biopsy. However, the results presented here are surely a promising step toward a helpful processing system to support the decision process for HS assessment in liver procurement setting.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval All studies have been approved and performed in accordance with ethical standards.

Informed consent Informed consent was obtained from all individuals for whom identifying information is included in this article.

References

1. Bhati C, Silva M, Wigmore S, Bramhall S, Mayer D, Buckels J, Neil D, Murphy N, Mirza D (2009) Use of bioelectrical impedance analysis to assess liver steatosis. In: Transplantation proceedings, vol 41. Elsevier, Amsterdam, pp 1677–1681
2. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
3. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167

4. Chen CL, Fan ST, Lee SG, Makuuchi M, Tanaka K (2003) Living-donor liver transplantation: 12 years of experience in Asia. Transplantation 75(3):S6–S11

5. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, Prague, vol 1, pp 1–2

6. D'alessandro AM, Kalayoglu M, Sollinger HW, Hoffmann RM, Reed A, Knechtle SJ, Pirsch JD, Hafez GR, Lorentzen D, Belzer FO (1991) The predictive value of donor liver biopsies for the development of primary nonfunction after orthotopic liver transplantation. Transplantation 51(1):157–163

7. Duro DC, Franklin SE, Dubé MG (2012) A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens Environ 118:259–272

8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118

9. Goceri E, Shah ZK, Layman R, Jiang X, Gurcan MN (2016) Quantification of liver fat: a comprehensive review. Comput Biol Med 71:174–189

10. Haralick RM, Shanmugam K (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 3(6):610–621

11. Hewitt KC, Rad JG, McGregor HC, Brouwers E, Sapp H, Short MA, Fashir SB, Zeng H, Alwayn IP (2015) Accurate assessment of liver steatosis in animal models using a high throughput Raman fiber optic probe. Analyst 140(19):6602–6609

12. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, Montreal, Canada, vol 14, pp 1137–1145

13. Koneru B, Dikdan G (2002) Hepatic steatosis and liver transplantation current clinical and experimental perspectives. Transplantation 73(3):325–330

14. Lechaux D, Dupont-Bierre E, Karam G, Corbineau H, Compagnon P, Noury D, Boudjema K (2004) Technique du prélèvement multiorganes: cœur-foie-reins. In: Annales de Chirurgie, vol 129. Elsevier, Amsterdam, pp 103–113

15. Li B, Meng MQH (2009) Texture analysis for ulcer detection in capsule endoscopy images. Image Vis Comput 27(9):1336–1342

16. Liang P, Cong Y, Guan M (2012) A computer-aided lesion diagnose method based on gastroscopeimage. In: 2012 International conference on information and automation. IEEE, pp 871–875

17. Lin Y, Lv F, Zhu S, Yang M, Cour T, Yu K, Cao L, Huang T (2011) Large-scale image classification: fast feature extraction and SVM training. In: 2011 IEEE conference on computer vision and pattern recognition. IEEE, pp 1689–1696

18. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, Hashizume M, Katic D, Kenngott H, Kranzfelder M, Malpani A, Marz K, Neumuth T, Padoy N, Pugh C, Schoch N, Stoyanov D, Taylor R, Wagner M, Hager GD, Jannin P (2017) Surgical data science for next-generation interventions. Nat Biomed Eng 1(9):691

19. Mancia C, Loustaud-Ratti V, Carrier P, Naudet F, Bellissant E, Labrousse F, Pichon N (2015) Controlled attenuation parameter and liver stiffness measurements for steatosis assessment in the liver transplant of brain dead donors. Transplantation 99(8):1619–1624

20. Marsman WA, Wiesner RH, Rodriguez L, Batts KP, Porayko MK, Hay JE, Gores GJ, Krom RA (1996) Use of fatty donor liver is associated with diminished early patient and graft survival. Transplantation 62(9):1246–1251

21. Misawa M, Se Kudo, Mori Y, Takeda K, Maeda Y, Kataoka S, Nakamura H, Kudo T, Wakamura K, Hayashi T, Katagiri A, Baba T, Ishida F, Inoue H, Nimura Y, Oda M, Mori K (2017) Accuracy of computer-aided diagnosis based on narrow-band imaging endocytoscopy for diagnosing colorectal lesions: comparison with experts. Int J Comput Assist Radiol Surg 12:1–10

22. Moccia S, De Momi E, Guarnaschelli M, Savazzi M, Laborai A, Guastini L, Peretti G, Mattos LS (2017) Confident texture-based laryngeal tissue classification for early stage diagnosis support. J Med Imaging 4(3):034502

23. Moccia S, Wirkert SJ, Kenngott H, Vemuri AS, Apitz M, Mayer B, De Momi E, Mattos LS, Maier-Hein L (2018) Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Trans Biomed Eng. https://doi.org/10.1109/TBME.2018.2813015

24. Mor E, Klintmalm GB, Gonwa TA, Solomon H, Holman MJ, Gibbs JF, Watemberg I, Goldstein RM, Husberg BS (1992) The use of marginal donors for liver transplantation. A retrospective study of 365 liver donors. Transplantation 53(2):383–386

25. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987

26. Qayyum A, Nystrom M, Noworolski SM, Chu P, Mohanty A, Merriman R (2012) MRI steatosis grading: development and initial validation of a color mapping system. Am J Roentgenol 198(3):582–588

27. Quellec G, Cazuguel G, Cochener B, Lamard M (2017) Multiple-instance learning for medical image and video analysis. IEEE Rev Biomed Eng 10:213–234

28. Rogier J, Roullet S, Cornélis F, Biais M, Quinart A, Revel P, Bioulac-Sage P, Le Bail B (2015) Noninvasive assessment of macrovesicular liver steatosis in cadaveric donors based on computed tomography liver-to-spleen attenuation ratio. Liver Transplant 21(5):690–695

29. Selzner M, Clavien PA (2001) Fatty liver in liver transplantation and surgery. In: Seminars in liver disease, Copyright© 2001 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662, vol 21, pp 105–114

30. Shen X, Sun K, Zhang S, Cheng S (2012) Lesion detection of electronic gastroscope images based on multiscale texture feature. In: 2012 IEEE international conference on signal processing, communication and computing (ICSPCC). IEEE, pp 756–759

31. Yersiz H, Lee C, Kaldas FM, Hong JC, Rana A, Schnickel GT, Wertheim JA, Zarrinpar A, Agopian VG, Gornbein J, Naini BV, Lassman CR, Busuttil RW, Petrowsky H (2013) Assessment of hepatic steatosis by transplant surgeon and expert pathologist: a prospective, double-blind evaluation of 201 donor livers. Liver Transplant 19(4):437–449

32. Zhang Y, Wirkert SJ, Iszatt J, Kenngott H, Wagner M, Mayer B, Stock C, Clancy NT, Elson DS, Maier-Hein L (2017) Tissue classification for laparoscopic image understanding based on multispectral texture analysis. J Med Imaging 4(1):015001