

# An evolutionary algorithm-based optimization method for the classification and quantification of steatosis prevalence in liver biopsy images

Alexandros Arjmand<sup>a, \*\*</sup>, Vasileios Christou<sup>a</sup>, Ioannis G. Tsoulos<sup>a</sup>, Markos G. Tsipouras<sup>b</sup>, Alexandros T. Tzallas<sup>a, \*</sup>, Christos Gogos<sup>a</sup>, Euripidis Glavas<sup>a</sup>, Nikolaos Giannakeas<sup>a, \*\*\*</sup>

<sup>a</sup> School of Informatics and Telecommunications, Department of Informatics and Telecommunications, University of Ioannina, Kostakioi, GR, 47100, Arta, Greece

<sup>b</sup> Department of Electrical and Computer Engineering, University of Western Macedonia, GR, 50100, Kozani, Greece

## ARTICLE INFO

### Keywords:

Liver biopsy  
Steatohepatitis  
Fatty liver  
Image analysis  
Machine learning  
Evolutionary algorithms  
Grammatical evolution

## ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) covers a range of chronic medical conditions varying from hepatocellular inflammation which characterizes nonalcoholic steatohepatitis (NASH) to steatosis, as the key element of a nonalcoholic fatty liver (NAFL). It is globally linked to the increasing prevalence of obesity and other components of metabolic syndrome and is expected to be the main indication for the existence of the liver disease in the coming years. Its eradication has become a major challenge due to the difficulties in clinical diagnosis, complex pathogenesis and the lack of approved therapies. In this study, an automated image analysis method is presented providing quantitative assessments of fat deposition in steatotic liver biopsy specimens. The methodology applies image processing, machine learning and evolutionary algorithm optimization techniques, producing a 1.93% mean classification error compared to the semiquantitative interpretations of specialized hepatologists.

## 1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a growing cause of chronic liver ailment worldwide [1]. Its main characteristic is the excessive fat deposition in the liver tissue which is not due to alcohol consumption but to risk factors such as obesity, insulin resistance and features of the metabolic syndrome. While other conditions such as genetic disorders of lipid metabolism can also evoke hepatic fat deposition, they tend to occur much less frequently in contrast to the above risk factors. NAFLD can also develop into non-alcoholic steatohepatitis (NASH). NASH has a significantly higher risk of developing cirrhosis than NAFLD, which causes further decompensation leading to liver transplantation in some individuals or even death. In general, NAFLD is closely related to hepatocellular carcinoma (HCC) [2], which is commonly thought to appear in the setting of cirrhosis or after a long-term chronic hepatitis B (CHB) infection. Estimates based on cross-sectional data show that 20–30% of Western World adults suffer from NAFLD, of which 2–5% also have a severe liver injury and 1–2%

may reach the end-stage of NASH cirrhosis.

While the staging of fibrosis is important for monitoring the clinical risk of cirrhosis progression, the valid assessment of fat cells' quantitative accumulation is of utmost importance and necessary for the early onset of pharmacological treatment. Although many non-invasive imaging techniques have been developed, liver tissue imaging via biopsy sample digitization has become the gold standard in clinical diagnoses. Needle biopsy specimens are usually embedded in 4 µm-thickness sections and stained with several histological dyes depending on the pathological condition, including Hematoxylin-Eosin (H&E), Masson's trichrome and Sirius red. Subsequently, their imaging through digital microscopy can provide an accurate prediction of several critical liver findings, including collagen proportionate area (CPA) indicating fibrosis, tissue inflammation, ballooning degeneration and steatosis.

In the early years, NAFLD examinations included a subjective microscopic interpretation of liver biopsies, usually leading to a clinical barrier called inter or intra-observer variability, referring to a diagnostic disagreement even among expert physicians. To overcome this issue,

\* Corresponding author.

\*\* Corresponding author.

\*\*\* Corresponding author.

E-mail addresses: [k.arjmand@uoi.gr](mailto:k.arjmand@uoi.gr) (A. Arjmand), [b.christou@uoi.gr](mailto:b.christou@uoi.gr) (V. Christou), [itsoulos@uoi.gr](mailto:itsoulos@uoi.gr) (I.G. Tsoulos), [mtsipouras@uowm.gr](mailto:mtsipouras@uowm.gr) (M.G. Tsipouras), [tzallas@uoi.gr](mailto:tzallas@uoi.gr) (A.T. Tzallas), [cgogos@uoi.gr](mailto:cgogos@uoi.gr) (C. Gogos), [eglavas@uoi.gr](mailto:eglavas@uoi.gr) (E. Glavas), [giannakeas@uoi.gr](mailto:giannakeas@uoi.gr) (N. Giannakeas).

several studies over the years have shown the importance of automated fat quantification through computer software for the accurate prediction and exclusion of false-positive results. Initially, Zaitoun et al. [3] concluded that Hematoxylin and Eosin (H&E) is the ideal histological stain for stereological and morphometric steatosis quantification in chronic hepatitis C (CHC) patients. Marsman et al. [4] demonstrated that image analysis systems can successfully automate the determination of total fat content in liver biopsy specimens, as diagnostic outcomes have been shown to correlate well with a pathologist's visual interpretation. The Roullier et al. [5] study relied on pixel intensity values thresholding for the segmentation of tissue sample areas and lipid droplet regions. Similarly, later studies [6,7] have revealed that color homogeneity can be a determining factor in distinguishing steatosis and other non-fat findings. The works from Liquori et al. [8] and Batool [9] suggested a more effective approach, that of employing morphological operators for non-circular objects (not referring to fat cells) elimination. The modern automated approaches presented in the works of Giannakes et al. [10] and Nativ et al. [11] have emphasized the significance of combining unsupervised machine learning (clustering) and active contour modeling (ACM) techniques, aiming at the segmentation of fat droplets. Sciarabba et al. [12] similarly proposed a clustering-based tool for hepatic steatosis assessment in liver biopsies. The proposed automated tool is then able to recognize fat droplets through 2-level thresholding and three shape parameters (solidity, elongation and roughness). Homeyer et al. [13] used adjacency statistics as shape features for the identification of fat droplets in histological images. Roy et al. [14] proposed an object detection method that combines the segmentation and classification of circular fat cells, as well as the segregation of overlapped steatosis components. A comprehensive work has been provided by Vanderbeck et al. [15], where expert pathologists' annotations were analyzed by supervised machine learning algorithms and in which the separation of various liver structures, including portal arteries, bile ducts and lipid droplets were attempted. In a subsequent study [16], the same group made a significant contribution to the detection of hepatocellular ballooning. This condition is currently at the center of doctors' clinical interest, alongside lobular inflammation and fibrosis assessment, since they represent important factors in the progression of chronic liver disease (CLD). Ishikawa et al. [17] proposed a method for analyzing stained liver images that can separate fat droplets and sinusoids from the stroma. Liu et al. [18] proposed the qFIBS automated tool, aimed at quantifying the spread of nonalcoholic steatohepatitis (NASH) patterns through histological evaluation. The proposed tool is based on the analysis of specific parameters regarding the following cardinal histopathological features: a) fibrosis, b) inflammation, c) hepatocyte ballooning and d) steatosis. Yip et al. [19] used four machine learning methods (Logistic Regression, Ridge Regression, AdaBoost and Decision Tree) for detecting NAFLD. Their experimental findings showed that the models which utilized Logistic Regression and Ridge Regression showed better results compared to the other two. Teramoto et al. [20] used a topological data analysis methodology combined with machine learning methods to classify tissue images into NAFLD sub-types. The proposed algorithm utilized the Matteoni classification system in biopsies taken from liver tissue. Vicas et al. [21] relied on CNN image analysis methods, with U-Net emerging as the optimal deep model for lipid droplet segmentation purposes. Yang et al. [22] created and trained a deep CNN using a large dataset of 50,000 images for the quantification of hepatic steatosis in liver samples. Guo et al. [23] proposed the Mask-RCNN method which is based on deep learning for segmenting liver steatosis droplets in clumps. Mask-RCNN utilizes bounding box detection and is able to predict object masks. Han et al. [24] created two one-dimensional convolutional neural networks (CNNs) utilizing raw radio-frequency ultrasound data taken from the liver. The first network was used as a binary classifier for the detection of NAFLD while the latter as a fat fraction estimator.

For the reason that few multiple liver disease detection systems have been implemented, the current study aims at the separation of four

anatomical structures to avoid false-positive identification results, so that a reliable tool can be provided for estimating the total fat ratio in the liver tissue. More technically, the developed methodology relies on image processing techniques for determining the liver tissue area and detecting circular objects of interest. The method then focuses on a 4-class identification problem through supervised machine learning, intending to distinguish fat cells and exclude other histopathological alterations from steatosis prevalence computations. The four classes refer to 1) fat droplets, 2) balloon cells, 3) sinusoids and 4) veins (Fig. 1). As a result, the accurate characterization of fat structures could lead to clinical trials in the objective quantification of fat accumulation in NAFLD patients.

## 2. Methods

In total, a three-step method is proposed, which ultimately leads to the calculation of the steatosis ratio in digitized liver biopsy samples.

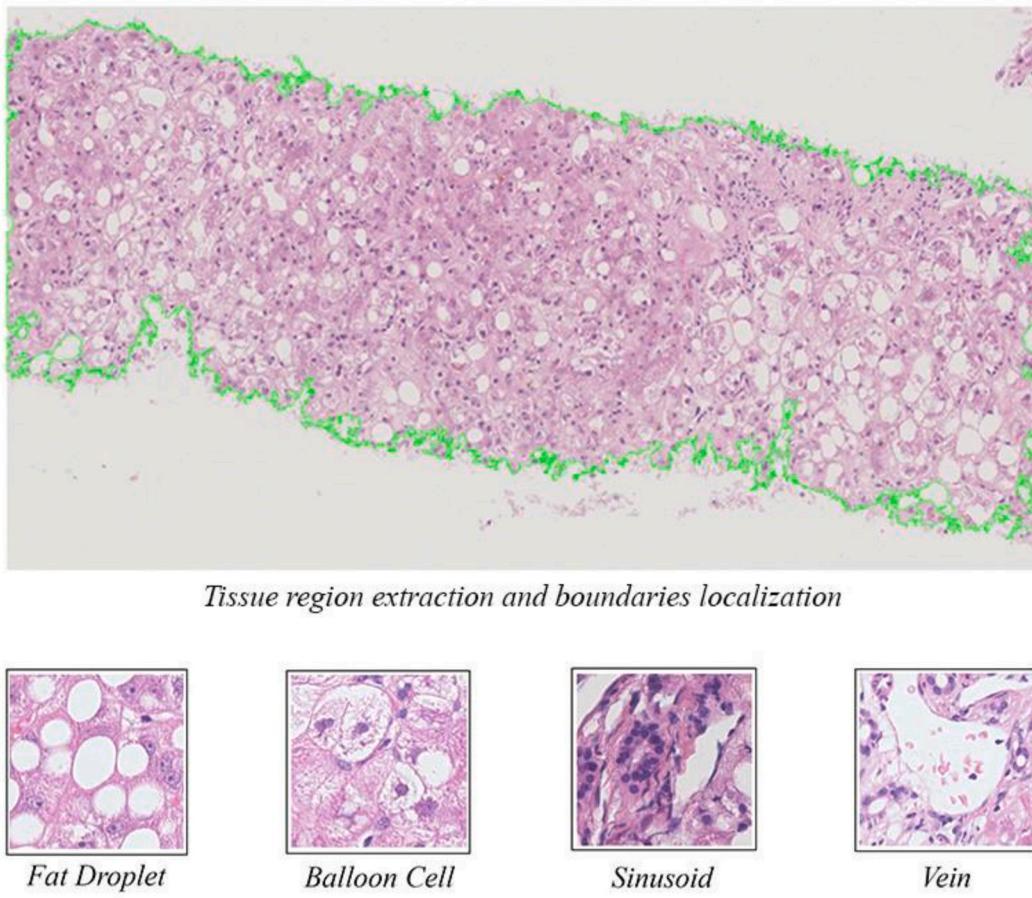
In the first step, a global thresholding approach is applied to separate background and tissue pixels in a set of  $\times 20$  testing images, followed by a recursive morphological operation for the elimination of non-circular white objects. The second step focuses on the annotation of histological structures in a set of  $\times 40$  training images. Finally, these annotations train a supervised classifier that distinguishes lipid droplets from other circular tissue anatomies, so that the steatosis prevalence can be computed. This degree of prevalence refers to the percentage of fat in the entire tissue sample for each  $\times 20$  testing image. Fig. 2 illustrates the flowchart of the developed method.

### 2.1. Tissue region and objects of interest segmentation

The first stage aims at identifying the entire tissue region in the  $\times 20$  testing images. For this purpose, an efficient tissue extraction method is applied. In more detail, the RGB histological image is initially converted to grayscale, to which its contrast is then increased by normalizing all 8-bit variance values and setting the contrast limits within a closed [0,1] interval. The Otsu method is then called upon to compute the global threshold leading eventually to the separation of the histological tissue from its background in an exported binary image (white pixels  $\rightarrow$  background, black pixels  $\rightarrow$  tissue). However, this step is not enough, as in most cases each microscopic specimen includes very small histological areas that are excised during the removal of the biopsy sample from the patient.

These do not belong to the main body of the retrieved specimen and should be excluded as part of the liver tissue region. According to this observation, the segmentation method first calculates the area for each 8-connected pixel region (adjacent pixels of the same object that are connected along the horizontal, vertical or diagonal direction). Then, connected pixel regions with an area less than 3% of the largest detected region are discarded as part of the tissue sample. As a result, an improved binary image is produced through which an active contour is set to determine the actual boundaries of the segmented tissue area in the original RGB biopsy image. Fig. 1 illustrates the final tissue segmentation result in a color source image with a green contour line.

The next step attempts to detect white areas of interest in the segmented binary tissue region, by applying morphological filtering with a circular structural element consisting of a 5-pixel radius value. More emphatically, a morphological opening process within a loop increases the radius of the circular mask by 2 pixels in order not only to smooth the boundaries of each filtered object but also to eliminate unnecessary artifacts with a radius of fewer than 5 pixels, which are considered as noise. Thus, only pure circular white structures that characterize fat cells are included. Finally, active contours are once again called upon to detect the boundaries for each filtered circular structure, as well as to calculate a series of histological features for them in the green channel of the initial RGB biopsy image. This extracted feature vector at a later stage leads to the separation of these detected



**Fig. 1.** Extraction of tissue component through Otsu thresholding in a grayscale intensity image, obtained from the initial RGB input values and derived from the H&E histological stain. The next step focuses on the localization of its regional boundaries with an active contour model. The segmented tissue area can then lead to the identification of 4 histological alterations, namely: a) fat droplets, b) balloon cells, c) sinusoids and d) veins.

objects as lipid cells from other microscopic anatomies with a trained classifier.

## 2.2. Selected features of training samples

Taking into account the training of a classification algorithm, a feature set is now calculated for histological structures of interest annotated by specialist hepatologists in the  $\times 40$  magnification images. This set composes the same feature vector as the filtered  $\times 20$  circular objects in the previously segmented testing images. More specifically, this feature set refers to four main categories so that the differences between the four aforementioned hepatic classes can be distinguished. These include the object of interest shape, texture, position and magnitude (Table 1), which are later fed as input to the proposed study classification algorithm.

Based on the Table 1, a supervised classifier  $D = (x_i, y_i)_{i=1}^N$  is trained, where  $x = \{\text{Eccentricity, Extent, Major axis length, Minor axis length, Mean intensity, Mean pixel value, StD pixel value, Solidity, Centroid-x, Centroid-y, Area, Diameter, Perimeter}\}$  denotes the input feature vector,  $y \in \{\text{fat droplet, ballooned cell, sinusoid, vein}\}$  the output labels, whereas  $i = 1, \dots, N$  the number of training samples and classes, respectively.

## 2.3. Classification method

This paper utilizes the GenClass method proposed by Tsoulos et al. [25], which is a tool based on grammatical evolution (GE).

### 2.3.1. Grammatical evolution

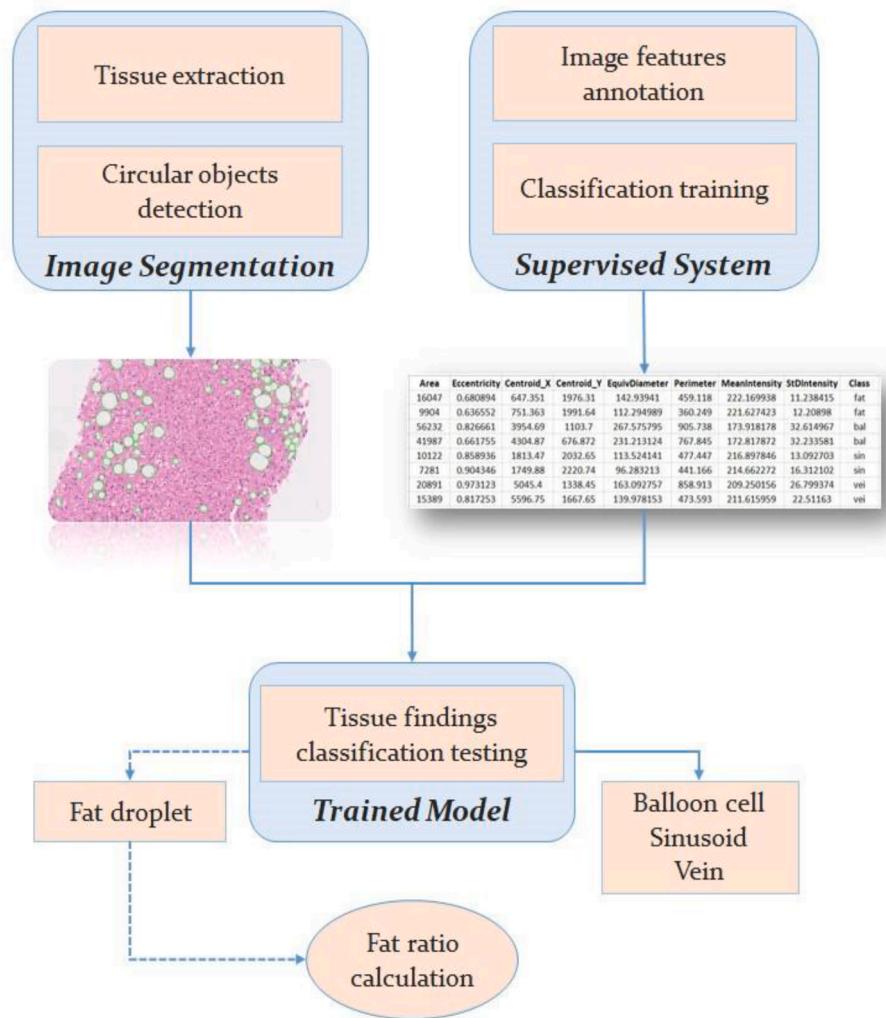
GE is a subset of evolutionary algorithms (EAs), which have been utilized with notable success in the automatic creation of programs. Specifically, genetic programming (GP) is amongst the most well-known EAs. GP has utilized the Lisp programming language as its target language, although many researchers have created a home-grown language that is specialized to the problem they are trying to solve [26–28].

GE was proposed by O'Neill and Ryan [29] and unlike GP, it doesn't apply the evolutionary process on the programs but on variable-length binary strings. GE is able to generate programs in any language by utilizing these binary strings in order to select production rules. These rules are written in a Backus–Naur Form (BNF)<sup>1</sup> grammar and the result is the creation of a program having a syntactically valid syntax that undergoes evaluation using a fitness function.

GE is inspired by the biological process of creating a protein using genetic material. The evolutionary process regarding the creation of a protein begins with the genome which is comprised of DNA as a string of sub-components which in turn are transcribed to RNA codons. These RNA codons are translated into amino-acid sequences which form the building blocks of the protein. The protein created using this evolutionary process in its environment forms the phenotype [31].

In GE, the phenotype is the resulting computer program created using a genome comprised of binary strings. The genome is converted

<sup>1</sup> BNF, is a meta-language (a language that is able to syntactically describe a programming language) which is able to specify the sequences of symbols which can create a syntactically valid program [30].



**Fig. 2.** Flowchart of the automated diagnostic method, consisting of image segmentation and supervised classification techniques to identify multiple liver tissue objects. The final classification stage leads to the fat ratio calculation in whole biopsy slide images.

**Table 1**  
Extracted features from annotations.

Feature Category	Description
1	Shape-based
2	Texture-based
3	Position
4	Magnitude

into a series of integers which are mapped onto predefined rules and form the computer program. The mapping procedure from genotype to phenotype is one of many processes mimicking the biological procedure where the same genetic material is utilized in different genes. The goal of GE is to adapt a computer program to an objective function for a specific problem using an iterative evolutionary process. This process is inspired by biology and includes procedures like descent with variation, mutation, crossover, genetic transcription and gene expression. The algorithm begins with a set of computer programs named population. The population's members are encoded as binary strings in a sub-symbolic form and they are evolved using an iterative process. The binary strings have different lengths and they are mapped to a well-structured symbolic form as executable Context-Free Grammar (CFG). This grammar is defined in BNF which is expressed as several production rules containing terminals and non-terminals. Non-terminals are abstract symbols that can be replaced by groups of terminal symbols

according to the production rules, while terminal symbols are specific lexical elements. The optimization process involves reading 8-bit data blocks (named Condon) from the candidate solutions genome which is then decoded to an integer. When the reading process reaches the string's end, it loops back to the string's beginning creating this way a circular genome. The integers are mapped to BNF expressions until a fully syntactically correct expression is created. This process may not use the entire genome of a solution or it might use the decoded genome many times due to its circular nature [31].

### 2.3.2. GenClass

Tsoulos et al. [25] proposed the GE-based GenClass portable tool for classification problems. Unlike recent methods for classification tasks including multilayer feed-forward neural networks [32] and Support Vector Machines (SVMs) [33], GenClass can create classification programs using GE [29]. The GE process requires the CFG of the target language (in GenClass, the target language is expressed in BNF format) and an associated fitness function. The CFG grammar is defined as  $G = (N, T, S, P)$ . In this formula,  $N$  defines a set of non-terminal symbols while  $T$  defines a finite set of terminal symbols having the constraint  $N \cap T = \emptyset$ . The last two symbols are  $S$  and  $P$  with the terminal symbol  $S$  named as the start symbol of the grammar and finally the  $P$  symbol defining a finite set of production rules having a form  $A \rightarrow a$  or  $A \rightarrow aB$ ,  $A \in N$ ,  $a \in T$ .

T.

---

Algorithm 1: GenClass

- 1: **Begin Initialization**
- 2: Load training data.
- 3: Set the maximum number of generations ( $N_G$ ).
- 4: Set the number of individuals forming the population ( $N_C$ ).
- 5: Set the selection rate ( $P_S$ ).
- 6: Set the mutation rate ( $P_M$ ).
- 7: Initialize the population's members ( $P_M$ ).
- 8: **End**
- 9: **Begin Evolution**
- 10: **For**  $i = 1, \dots, N_G$  **do**
- 11: Create a classification program for each individual
- 12: Calculate the fitness for all individuals.
- 13: Apply the selection operator.
- 14: Apply the mutation operator.
- 15: **End**
- 16: **End**
- 17: **Begin Evaluation**
- 18: Create a classification program for the selected best chromosome of the evolution process.
- 19: Apply the program to the test set and return the outcome.
- 20: **End**

---

In GE, the chromosomes are expressed as vectors containing integers where each vector's element represents a production rule from the BNF grammar. The GE process begins from the start symbol of the grammar, following an iterative process that produces the program string. This procedure replaces non-terminal symbols with the right hand of the selected production rule. The selection process uses a two-step approach where the succeeding vector element (depicted with the symbol  $V$ ) is replaced with the selected production rule using the formula ( $\text{Rule} = V \bmod R$ ). In this formula,  $R$  represents the number of production rules for the current non-terminal symbol. This two-step approach is repeated until it produces a valid expression or the individual's end is reached.

Regarding the latter, the individual is defined as invalid which results in assigning a large value to its corresponding fitness. The GenClass method is depicted in Algorithm 1 above.

### 3. Results

#### 3.1. Materials and methods

All biopsy specimens used in the current work are colored with the gold standard H&E histological staining and derived from human subjects, who gave their signed consent for the inclusion of their samples in experimental studies, following the ethical principles of the Helsinki Declaration. Part of the population refers to NAFLD patients suffering from high steatosis prevalence, while in many cases high rates of hepatocyte ballooning are also observed in individuals with NASH progression. The dataset comes from cases collected at St. Mary Hospital (Imperial College Healthcare NHS Trust of London, UK) and captured with a Hamamatsu (Hamamatsu Photonics, Hamamatsu, Japan) microscope. Initially, the digitized biopsies consisted of sizes exceeding  $10,000 \times 10,000$  pixels. Thus, a necessary step involved the compression of these images for the efficient use of computational resources.

Particularly, the current methodology examines two subsets of liver biopsy images. The first includes 13 training images carrying a  $\times 40$  magnification, whereas, the second subset involves 28 testing samples with a smaller  $\times 20$  magnification. These two options reflect real clinical diagnostic procedures, as digital scanners are called upon to recognize various tissue findings on different scales. For this purpose, the normalization of feature values in a closed  $[0,1]$  interval was a prerequisite for both magnification factors.

Emphasis is now placed on the annotation of histological objects in the  $\times 40$  training images. Fig. 3 shows the NDP. View 2 software environment, whereby using a freehand tool the hepatologists have



**Fig. 3.** Presentation of the NDP. View 2 pathology software, which includes a series of image visualization and object of interest annotation tools. For the current annotation task, a freehand tool with different edge line colors is employed for the selection of training samples. Details of the boundary coordinates of the selected areas are recorded in an XML file.

manually annotated white areas of fat content (red contour lines), ballooned hepatocytes (yellow contour lines), sinusoids (green contour lines) and veins (blue contour lines). Each annotated region is automatically exported to an XML file in the form of a two-dimensional image (Cartesian) coordinates in the original  $\times 40$  color image, while being assigned a unique “id” and a class label specified by the clinicians.

At the end of the annotation process, 7,305 in total training samples are provided, consisting of a) 4,023 fat cells, b) 3,064 ballooning areas, c) 165 sinusoids and d) 53 veins, with fat droplets and ballooned hepatocytes making up the majority of the training samples. Even though these numbers indicate an adequate training set, they form an unbalanced dataset, a common obstacle that leads to reduced performance in various classification tasks. As a solution, the training set is reformed, by including 390 random steatosis class objects and 390 random ballooning cases, while keeping all 165 sinusoids and 53 veins.

Based on [Table 1](#), the feature vector for each of the four class objects is also calculated in the green channel for each  $\times 40$  training image. It should be noted here that for both  $\times 20$  and  $\times 40$  image subsets the feature values are calculated in the green color channel, as differences in the pink hue of the H&E histological stain and texture of the four examined anatomical structures are determined more by the 8-bit intensities of the green channel in the RGB color space. To verify the correct termination of the feature values calculation process, [Fig. 4](#) gives an illustration of the final annotation result, where a gray-scale image is exported having the same size as the original RGB tissue sample. It is observed that each annotated class region is displayed with different gray-scale values, while the color dots indicate the calculated eccentricity point.

### 3.2. Image segmentation results

For the evaluation of the methodology, the testing subset of  $\times 20$  magnification samples is taken into consideration for diagnostic purposes. First, an important prerequisite involves examining the visualization results of the initial image segmentation stage.

As shown in [Fig. 5](#), green active contour lines have converged on the

boundaries of all circular structures of interest, which signal the detection of individual fat droplets. The success of the proposed image segmentation stage lies also in the fact that most cases of hepatocyte ballooning, as well as large and noncircular white areas, suggesting liver veins and sinusoids, have been successfully excluded from the total number of fat pixels calculation.

For each  $\times 20$  image of the testing subset ( $n = 28$ ), the fat ratio is calculated as the total area of unknown circular class objects, divided by the initially extracted tissue region ( $F_{segm}$ ). This prevalence ratio is also provided from semi-quantitative estimates by St. Mary’s pathologists ( $F_{doc}$ ), before the four findings classification stage. Hence, the difference between these percentages represents the absolute error for each segmented biopsy image, as follows:

$$Serr = |F_{doc} - F_{segm}|100\% \quad (1)$$

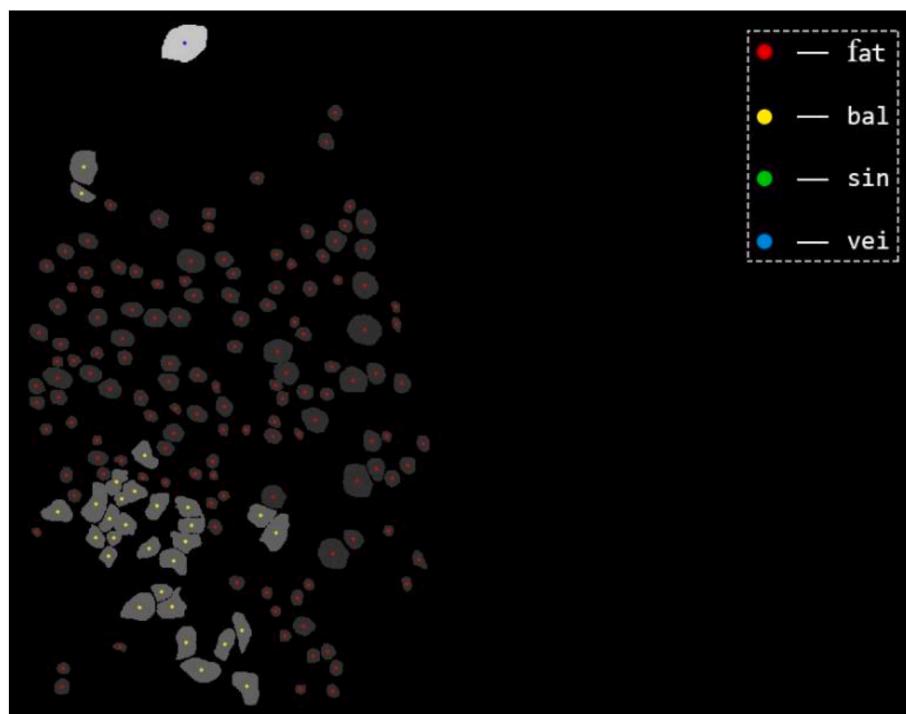
### 3.3. Classification results

Thanks to the methodology expansion, the absolute error between the doctors’ steatosis ratio ( $F_{doc}$ ) evaluations and the proposed ( $F_{GE}$ ) classification stage, which separates the segmented circular objects as fat structures from other histological findings (balloon cells, sinusoids, veins), is estimated by:

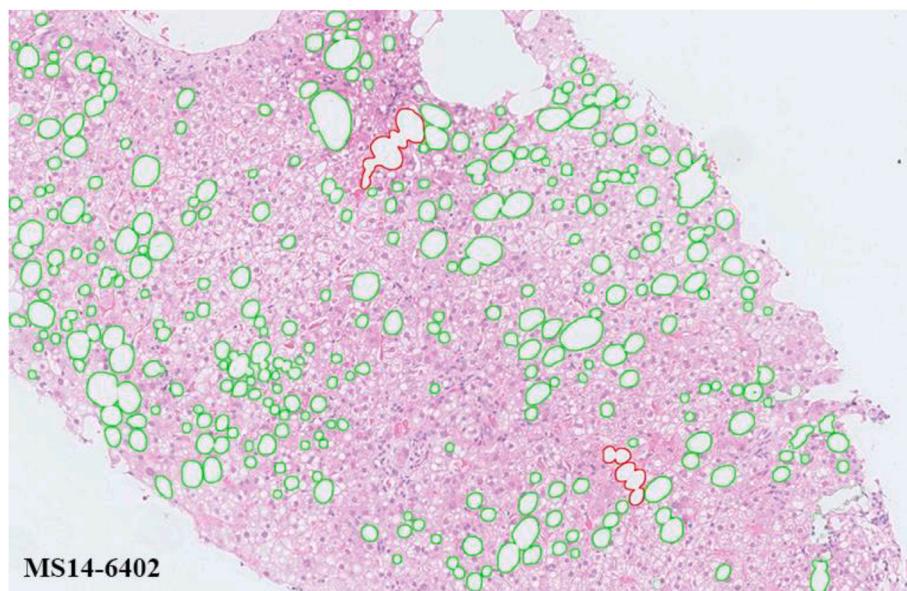
$$GEerr = |F_{doc} - F_{GE}|100\% \quad (2)$$

Based on the integrated methodology, the focus is on the final diagnostic tables ([Table 2a](#) and [Table 2b](#)). More particularly, the third column in [Table 2a](#) shows that in all 28 biopsy specimens the classification stage utilizing the GenClass method based on GE ( $F_{GE}$ ) yields lower steatosis percentages than that of the  $F_{segm}$  image segmentation step (in the second column).

It is stressed that the proposed GE-based GenClass method has reduced the mean fat ratio up to 0.27% (18.21% compared to 18.48%). The decrease in mean value comes from the fact that tissue findings relating to balloon cells, sinusoids and veins, which in the image



**Fig. 4.** Visualization of boundaries and eccentricity values for each previously annotated training sample. The visualization process takes place on a grayscale image exported from the original RGB digitized biopsy, while each eccentricity point is indicated with dots colored according to the annotation edge line color.



**Fig. 5.** Visualization results of the image segmentation stage, where accurate detection of circular fat droplets with green contour lines is performed. However, larger areas of fat agglomeration marked with red contour lines are observed, which point to overlapping lipid droplet regions. This agglomeration phenomenon refers to a limitation of  $\times 20$  magnification which affects the determination of each circular object boundaries, as well as the classification stage performance.

segmentation stage would have been included in the fat ratio quantification, have been removed. Proceeding, the results of the fourth column show all fat ratio percentages according to the semi-quantitative assessment of the St. Mary physicians ( $F_{doc}$ ). These visual interpretations eventually lead to the second and third columns of **Table 2b**, in which the absolute error for the circular structures segmentation and their classification as fat methods, respectively, is calculated for each  $\times 20$  liver biopsy image. As follows, the mean diagnostic error of the classification method ( $GE_{err}$ ) is determined and also compared with that produced by the image segmentation ( $S_{err}$ ) estimates. Based on this, it is concluded that the GenClass algorithm has a mean absolute error of 1.93% and less than 2.08% of the image segmentation stage.

#### 4. Discussion

Non-alcoholic fatty liver disease (NAFLD) and its progression to non-alcoholic steatohepatitis (NASH) are major causes of cirrhosis and hepatocellular carcinoma (HCC) worldwide, eventually leading to end-stage liver disease. The prevalence of NAFLD is determined by the quantitative degree of steatosis (accumulation of fat droplets in the hepatic tissue) without significant necroinflammatory injury, whereas NASH refers to more complex and active lesions ranging from tissue inflammation to hepatocellular ballooning and necrosis [34]. This spectrum of NAFLD and NASH complications prevents the definition of an accurate pharmacological treatment, an obstacle that has led researchers to develop automated quantification and staging tools based on image analysis, with histology being the gold standard in recent clinical trials.

The current work focuses on solving the problem of visual and subjective diagnosis of physicians by building an automated diagnostic tool for the accurate steatosis prevalence quantification in NAFLD biopsy images. In the first stage, the image segmentation method extracts the tissue region from its background and then white circular structures are detected with active contours within the extracted tissue. The procedure involves 28 biopsy images carrying a  $\times 20$  magnification, through which a set of feature values is calculated for classification purposes. In the second stage, expert clinicians form a training dataset with annotated samples from 13 images carrying a larger  $\times 40$  magnification. The annotation process consists of four histological classes, namely a) fat

droplets, b) ballooned hepatocytes, c) sinusoids and d) veins. The final stage utilizes the GenClass grammatical evolution (GE) method, which is trained with the annotated dataset. The trained system is then called upon to classify the segmented  $\times 20$  circular structures, to include only the fat cells in the steatosis quantification in the 28 testing images.

Based on the complete methodology, the classification stage presents a 1.93% mean fat quantification error, which indicates the degree of deviation from semiquantitative estimates of physicians in the 28 testing samples. The classification method presents also a smaller mean absolute error than the circular structures segmentation stage (2.08%) before their characterization. The following sections further analyze the diagnostic results, which are then compared with those derived from other classification algorithms.

##### 4.1. Discussion of diagnostic results

According to **Fig. 5** the circular object segmentation stage presents in some cases an inability to locate the boundaries between adjacent fat droplets. This is because the low  $\times 20$  magnification has resulted in a high compression ratio in the biopsy image, which causes adjacent circular fat structures to overlap (red contour lines). This fat agglomeration pattern produces feature values that do not correspond to those of the annotated fat samples, thus affecting the classification performance of the GenClass method. As a consequence, the classifier is not able to recognize all the fat structures, leading to the steatosis ratio being underestimated. It should be noted that the increased  $\times 40$  magnification is considered the ideal scaling factor among researchers to equalize the quality of biopsy images with computational processing resources. Based on this, it is estimated that the outer boundaries among densely occurring lipid droplets could become more evident, thus causing fewer overlapping fat regions.

In terms of classification performance, based on **Table 2b** the fat droplet overlapping effect resulted in the GE method producing higher  $C_{err}$  fat quantification error than the corresponding  $S_{err}$  image segmentation outputs in 8 testing images: MS14-10783 (0.51% > 0.30%), MS14-12212 (0.29% > 0.06%), MS14-1559 (2.72% > 2.53%), MS14-2382 (2.62% > 2.48%), MS14-2449 (1.89% > 1.73%), MS14-5898 (1.00% > 0.75%), MS14-786 (3.46% > 3.23%), MS148355 (2.31% > 2.22%). However, the trained system produces lower diagnostic error values in 20 images (out of 28 in total), as well as a reduced mean

**Table 2a**

Fat quantification results.

	Testing Image ( $\times 20$ )	Fat Ratio (%) w/o Regions Classification	Fat Ratio (%) with Regions Classification	Fat Ratio (%) from Visual Assessment
		Fsegm	FGE	Fdoc
1	MS12-23945	9.88	8.18	7.65
2	MS13-11663	5.43	5.25	3.47
3	MS13-12414	16.34	16.09	14.06
4	MS13-18536	11.76	11.62	10.64
5	MS13-9453	27.42	27.24	24.13
6	MS13-9925	25.87	25.60	23.99
7	MS14-10783	15.83	15.62	16.13
8	MS14-10801	9.99	9.75	9.82
9	MS14-11933	7.23	7.03	3.77
10	MS14-12212	26.64	26.42	26.71
11	MS14-13040	12.80	12.52	12.63
12	MS14-1559	24.32	24.13	26.85
13	MS14-2382	37.60	37.45	40.08
14	MS14-2449	12.97	12.81	14.70
15	MS14-2451	15.17	14.93	13.98
16	MS14-3215	14.32	14.14	10.84
17	MS14-3686	4.55	4.39	2.27
18	MS14-3854	27.23	26.88	24.32
19	MS14-5561	6.79	6.67	4.52
20	MS14-5898	25.66	25.41	26.41
21	MS14-6402	15.24	14.88	13.42
22	MS14-786	29.28	29.04	32.50
23	MS14-8355	42.27	42.18	44.50
24	MS14-9198	12.23	11.68	7.09
25	MS14-9711	32.20	31.99	29.79
26	MS15-2103	5.14	4.90	2.79
27	MS15-2670	30.99	30.83	29.61
28	MS15-805	12.35	12.10	9.12
<b>Mean Value:</b>	<b>18.48</b>	<b>18.21</b>	<b>17.35</b>	
<b>StD:</b>	<b>10.44</b>	<b>10.50</b>	<b>11.62</b>	

absolute error of up to 0.15%, as opposed to 2.08% of the image segmentation stage. A graphical representation of these outcomes is illustrated in Fig. 6 using gridlines.

On average, the absolute error values coming from the classification stage employing the GE method ( $GE_{err}$ ) (mean = 1.93, StD = 1.13) was found lower in comparison to the image segmentation stage ( $S_{err}$ ) mean = 2.08, StD = 1.18). This difference, 0.15, 95% CI [0.15, 0.29] was statistically significant ( $t(27) = 2.272, p = 0.032$ ).

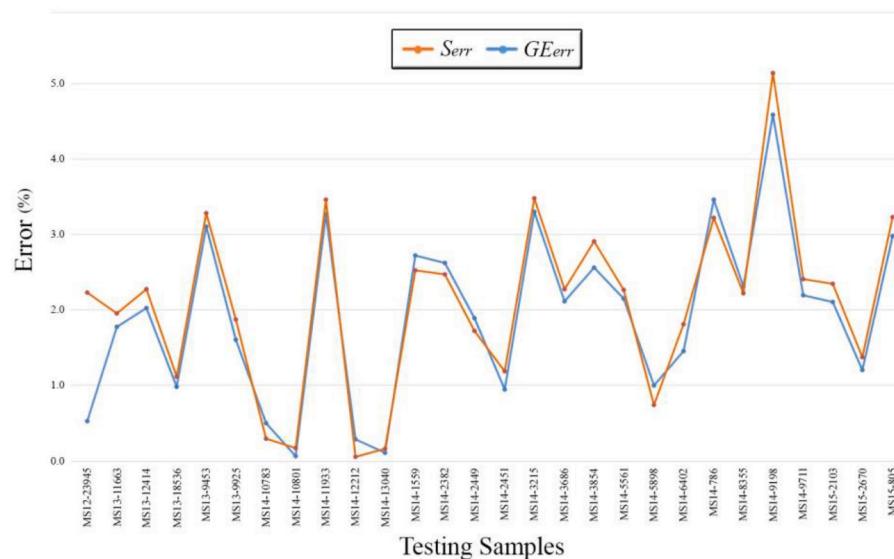
**Table 2b**

Fat quantification error.

	Testing Image ( $\times 20$ )	Image Segmentation Error (%)	Classification Error (%)				
			Serr	NB <sub>err</sub>	k-NN <sub>err</sub>	C4.5 <sub>err</sub>	SVM <sub>err</sub>
1	MS12-23945	2.23	1.69	1.44	1.76	1.65	0.54
2	MS13-11663	1.96	1.54	1.54	1.67	1.54	1.78
3	MS13-12414	2.28	1.32	0.03	1.58	1.25	2.03
4	MS13-18536	1.12	0.63	0.38	0.80	0.65	0.98
5	MS13-9453	3.29	1.63	1.53	1.55	1.82	3.11
6	MS13-9925	1.88	0.99	0.55	1.55	1.10	1.61
7	MS14-10783	0.30	1.37	1.49	1.00	1.19	0.51
8	MS14-10801	0.18	0.82	0.48	0.10	0.78	0.07
9	MS14-11933	3.47	2.96	2.99	3.13	2.91	3.26
10	MS14-12212	0.06	2.89	3.43	1.77	2.77	0.29
11	MS14-13040	0.16	0.37	0.49	0.23	0.32	0.11
12	MS14-1559	2.53	3.51	4.21	3.46	3.45	2.72
13	MS14-2382	2.48	4.66	6.24	4.86	5.06	2.62
14	MS14-2449	1.73	2.65	3.36	2.30	2.53	1.89
15	MS14-2451	1.19	0.93	0.27	0.99	0.93	0.95
16	MS14-3215	3.48	1.74	0.68	2.69	1.58	3.30
17	MS14-3686	2.28	2.11	2.10	2.17	2.16	2.12
18	MS14-3854	2.91	0.71	1.72	0.77	0.28	2.56
19	MS14-5561	2.27	1.42	1.26	1.74	1.52	2.15
20	MS14-5898	0.75	3.16	5.05	2.79	3.32	1.00
21	MS14-6402	1.81	1.16	0.81	1.39	1.28	1.46
22	MS14-786	3.23	4.49	4.93	4.00	4.21	3.46
23	MS14-8355	2.22	3.43	4.97	3.79	3.60	2.31
24	MS14-9198	5.14	4.69	4.66	4.89	4.74	4.59
25	MS14-9711	2.41	0.71	1.53	0.29	0.43	2.20
26	MS15-2103	2.35	2.01	1.63	2.09	1.96	2.11
27	MS15-2670	1.38	0.14	0.91	0.11	0.00	1.21
28	MS15-805	3.23	2.66	2.44	2.78	2.53	2.98
<b>Mean Value:</b>	<b>2.08</b>	<b>2.01</b>	<b>2.18</b>	<b>2.01</b>	<b>1.99</b>	<b>1.93</b>	
<b>StD:</b>	<b>1.18</b>	<b>1.31</b>	<b>1.76</b>	<b>1.34</b>	<b>1.36</b>	<b>1.13</b>	

#### 4.2. Performance comparison with other classification algorithms

In the next step, the performance error of the GenClass method within the classification stage is compared with that of four widely used supervised algorithms. These refer to Naive Bayes, k-Nearest Neighbors (k-NN), C4.5 Decision Tree and Support Vector Machines (SVM), which similarly aim to separate the fat structures from the other three histological classes after the image segmentation stage. According to Table 2b, the GenClass tool has the smallest mean fat quantification



**Fig. 6.** Graphical representation of the methodology's fat quantification error. The formed gridlines indicate 20 (out of 28) lower absolute error values coming from the classification stage employing the GE method ( $GE_{err}$ ) in comparison to the image segmentation stage ( $S_{err}$ ). The absolute error for each stage denotes its deviation from the doctors' semi-quantitative interpretations. It is also observed that due to the  $\times 20$  magnification, fat agglomeration areas lead to increased error levels ( $> 2\%$ ) produced in both  $F_{segm}$  and  $F_{GE}$  fat quantification stages.

error ( $GE_{err} = 1.93\%$ ) compared to the Naive Bayes ( $NB_{err} = 2.01\%$ ), k-NN ( $k\text{-}NN_{err} = 2.18\%$ ), C4.5 Decision Tree ( $C4.5_{err} = 2.01\%$ ) and SVM ( $SVM_{err} = 1.99\%$ ) algorithms. It is pointed out that k-NN is the only classifier that produces a larger mean absolute error compared to the image segmentation method ( $S_{err} = 2.08\%$ ).

To compare the percentage classification error of the  $GE_{err}$  method with the classification error of four aforementioned classification algorithms ( $NB_{err}$ ,  $k\text{-}NN_{err}$ ,  $C4.5_{err}$  and  $SVM_{err}$ ), the paired two-sample  $t$ -tests were used. The results obtained with those statistical tests are shown in Fig. 7 and indicated no statistically significant differences.

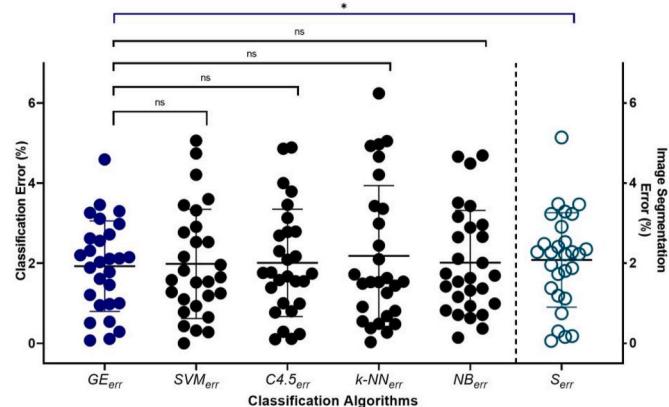
#### 4.3. Machine learning versus Deep Learning in Biopsy Image Analysis

In recent years the medical field has been experiencing a revolution in intelligent diagnostic systems with the advent of deep neural networks. Deep learning architectures, such as Convolutional Neural Networks (CNNs) and Autoencoders, have been successfully applied to medical image analysis tasks, as they filter the most informative pathological structures directly from the image data. As a result, they have also offered innovative solutions in the analysis of NAFLD and NASH pathogens without the necessary inclusion of manual annotations.

The above deep models may have shown high diagnostic performance in liver biopsy specimens, but their training process is time-consuming and requires a large number of image data samples so that they can generalize to the most critical histological characteristics. Therefore, in cases where the diagnostic problem is clear and histological features from human expert annotations are available, traditional machine learning techniques should be employed, so that computational processing and memory resources are reduced during the training of automated analysis systems.

#### 5. Conclusions

In this study, a novel method for the detection and quantification of fat content in liver biopsy samples is presented, combining image processing, machine learning and global optimization techniques. According to the diagnostic results, the proposed computer vision tool produces reliable fat ratio quantification results, as the proposed GenClass algorithm based on grammatical evolution (GE) achieves a mean absolute error of 1.93%, referring to the degree of deviation from the visual assessment of specialist physicians. This is the smallest fat quantification error compared to the Naive Bayes, k-NN, C4.5 decision tree and SVM classifiers (2.01%, 2.18%, 2.01% and 1.99%, respectively) and the



**Fig. 7.** Scatter plot representation and paired two-sample  $t$ -tests results of the comparison for each of the four classification algorithms ( $SVM_{err}$ ,  $C4.5_{err}$ ,  $k\text{-}NN_{err}$ , and  $NB_{err}$ ) with the  $GE_{err}$  method. The asterisk indicates significantly different values \*  $p < 0.05$  and ns indicates no significantly values.

image segmentation stage (2.08%) before the classification of detected histological objects. The main advantage of the overall method lies in the fact that the classification stage imitates the hepatologists' clinical procedure which focuses on the recognition of pathological patterns, by exploiting the characteristics of shape, texture, pixel intensity and magnitude. Based on these conclusions, the automated diagnostic tool can identify multiple hepatic anatomy landmarks in H&E stained biopsy specimens and detect differences between fat droplets and other tissue alterations, a step that can solve the NAFLD prevalence quantification problem in clinical trials.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgements

We acknowledge the support of this work by the project "Immersive Virtual, Augmented and Mixed Reality Center Of Epirus" (MIS 5047221) which is implemented under the Action "Reinforcement of the Research

and Innovation Infrastructure”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

## References

- [1] Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology* 2018;67(1):123–33.
- [2] Neuschwander-Tetri BA. Non-alcoholic fatty liver disease. *BMC Med* 2017;15(1):45.
- [3] Zaitoun A, Al Mardini H, Awad S, Ukbam S, Makadisi S, Record C. Quantitative assessment of fibrosis and steatosis in liver biopsies from patients with chronic hepatitis c. *J Clin Pathol* 2001;54(6):461–5.
- [4] Marsman H, Matsushita T, Dierkhising R, Kremers W, Rosen C, Burgart L, Nyberg SL. Assessment of donor liver steatosis: pathologist or automated software? *Hum Pathol* 2004;35(4):430–5.
- [5] Roullier V, Cavarro-Menard C, Guillaume C, Aube C. Fuzzy algorithms to extract vacuoles of steatosis on liver histological color images. In: 2007 29th annual international conference of the IEEE engineering in medicine and biology society. IEEE; 2007. p. 5575–8.
- [6] El-Badry AM, Breitenstein S, Jochum W, Washington K, Paradis V, Rubbia-Brandt L, Puhan MA, Slankamenac K, Graf R, Clavien P-A. Assessment of hepatic steatosis by expert pathologists: the end of a gold standard. *Ann Surg* 2009;250(5):691–7.
- [7] Arjmand A, Giannakeas N. Fat quantitation in liver biopsies using a pretrained classification based system. *Engineering, Technology & Applied Science Research* 2018;8(6):3550–5.
- [8] Liquori GE, Calamita G, Cascella D, Mastrodonato M, Portincasa P, Ferri D. An innovative methodology for the automated morphometric and quantitative estimation of liver steatosis. *Histol Histopathol* 2009;24(1):49–60.
- [9] Batool N. Detection and spatial analysis of hepatic steatosis in histopathology images using sparse linear models. 2016 sixth international conference on image processing theory, tools and applications (IPTA). IEEE; 2016. p. 1–6.
- [10] Giannakeas N, Tsipouras MG, Tzallas AT, Vavva MG, Tsimplakidou M, Karvounis EC, Forlano R, Manousou P. Measuring steatosis in liver biopsies using machine learning and morphological imaging. In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS). IEEE; 2017. p. 40–4.
- [11] Nativ NI, Chen AI, Yarmush G, Henry SD, Lefkowitch JH, Klein KM, Maguire TJ, Schloss R, Guarner JV, Berthiaume F, et al. Automated image analysis method for detecting and quantifying macrovesicular steatosis in hematoxylin and eosin-stained histology images of human livers. *Liver Transplant* 2014;20(2):228–36.
- [12] Sciarabba M, Vertemati M, Moscheni C, Cossa M, Vizzotto L. Automated lipid droplets recognition in human steatotic liver: some preliminary results. In: Medical image understanding and analysis (miua) conference. London: Kingston University; 2009. p. 234–8.
- [13] Homeyer A, Schenck A, Arlt J, Dahmen U, Dirsch O, Hahn HK. Fast and accurate identification of fat droplets in histological images. *Comput Methods Prog Biomed* 2015;121(2):59–65.
- [14] Roy M, Wang F, Teodoro G, Vos MB, Farris AB, Kong J. Segmentation of overlapped steatosis in whole-slide liver histopathology microscopy images. In: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2018. p. 810–3.
- [15] Vanderbeek S, Bockhorst J, Komorowski R, Kleiner DE, Gawrieh S. Automatic classification of white regions in liver biopsies by supervised machine learning. *Hum Pathol* 2014;45(4):785–92.
- [16] Vanderbeek S, Bockhorst J, Kleiner D, Komorowski R, Chalasani N, Gawrieh S. Automatic quantification of lobular inflammation and hepatocyte ballooning in nonalcoholic fatty liver disease liver biopsies. *Hum Pathol* 2015;46(5):767–75.
- [17] Ishikawa M, Murakami Y, Ahi ST, Yamaguchi M, Kobayashi N, Kiyuna T, Yamashita Y, Saito A, Abe T, Hashiguchi A, et al. Automatic quantification of morphological features for hepatic trabeculae analysis in stained liver specimens. *J Med Imag* 2016;3(2):027502.
- [18] Liu F, Goh GB-B, Tiniakos D, Wee A, Leow W-Q, Zhao J-M, Rao H-Y, Wang X-X, Wang Q, Wan W-K, et al. qfibs: an automated technique for quantitative evaluation of fibrosis, inflammation, ballooning, and steatosis in patients with nonalcoholic steatohepatitis. *Hepatology* 2020;71(6):1953–66.
- [19] Yip T-F, Ma A, Wong V-S, Tse Y-K, Chan H-Y, Yuen P-C, Wong G-H. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (nafld) in the general population. *Aliment Pharmacol Therapeut* 2017;46(4):447–56.
- [20] T. Teramoto, T. Shinohara, A. Takiyama, Computer-aided classification of hepatocellular ballooning in liver biopsies from patients with nash using persistent homology, *Comput Methods Prog Biomed* 195 (105614).
- [21] Vicas C, Rusu I, Al Hajjar N, Lupșor-Platon M. Deep convolutional neural nets for objective steatosis detection from liver samples. In: Proceedings of the 13th IEEE international conference on intelligent computer communication and processing (ICCP). IEEE; 2017. p. 385–90.
- [22] Yang F, Jia X, Lei P, He Y, Xiang Y, Jiao J, Zhou S, Qian W, Duan Q. Quantification of hepatic steatosis in histologic images by deep learning method. *J X Ray Sci Technol* 2019;1:1–13.
- [23] Guo X, Wang F, Teodoro G, Farris AB, Kong J. Liver steatosis segmentation with deep learning methods. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE; 2019. p. 24–7.
- [24] Han A, Byra M, Heba E, Andre MP, Erdman JW, Loomba R, Sirlin CB, O'Brien WD. Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. *Radiology* 2020;295(2):342–50.
- [25] Tsoulos I, Tzallas A, Tsalikakis DNNC. A tool based on Grammatical Evolution for data classification and differential equation solving. *SoftwareX* 2019;10:100297. ISSN 2352-7110.
- [26] Koza JR. Genetic programming: on the programming of computers by means of natural selection, vol. 1. MIT press; 1992.
- [27] Koza JR. Genetic programming II: automatic discovery of reusable programs. Cambridge, MA, USA: MIT Press; 1994.
- [28] Koza JR, Andre D, Keane MA, Bennett FH. Genetic programming III: darwinian invention and problem solving, vol. 3. Morgan Kaufmann; 1999.
- [29] O'Neill M, Ryan C. Grammatical evolution. *IEEE Trans Evol Comput* 2001;5(4):349–58.
- [30] D. D. McCracken, E. D. Reilly, Backus-Naur Form (BNF); 2003.
- [31] Brownlee J. Clever algorithms: nature-inspired programming recipes. Jason Brownlee; 2011.
- [32] Hornik K, Stinchcombe M, White H, et al. Multilayer feedforward networks are universal approximators. *Neural Network* 1989;2(5):359–66.
- [33] Cai C-Z, Wang W-L, Chen Y-Z. Support vector machine classification of physical and biological datasets. *Int J Mod Phys C* 2003;14(5):575–85. 0.
- [34] Bedossa P. Pathology of non-alcoholic fatty liver disease. *Liver Int* 2017;37(Suppl. 1):85–9.