This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

Radiomics-based assessment of Primary Sjögren's Syndrome from salivary gland ultrasonography images

Arso M. Vukicevic[1,2,*], Vera Milic[3], Alen Zabotti[4], Alojzija Hocevar[5], Orazio De Lucia[6], Georgios Filippou[7], Alejandro F. Frangi[8], Athanasios Tzioufas[9], Salvatore De Vita[3], Nenad Filipovic[1,2]

***Abstract*—Salivary gland ultrasonography (SGUS) has shown good potential in the diagnosis of Primary Sjögren's syndrome (pSS). However, a series of international studies have reported needs for improvements of the existing pSS scoring procedures in terms of inter/intra observer reliability before being established as standardized diagnostic tools. The present study aims to solve this problem by employing radiomics features and artificial intelligence (AI) algorithms to make the pSS scoring more objective and faster compared to human expert scoring. The assessment of AI algorithms was performed on a two-centric cohort, which included 600 SGUS images (150 patients) annotated using the original SGUS scoring system proposed in 1992 for pSS. For each image, we extracted 907 histogram-based and descriptive statistics features from segmented salivary glands (SG). Optimal feature subsets were found using the Genetic algorithm-based wrapper approach. Among the considered algorithms (7 classifiers and 5 regressors), the best preforming was the Multilayer perceptron (MLP) classifier (κ = 0.7). The MLP over-performed average score achieved by the clinicians (κ = 0.67) by the considerable margin, while its reliability was on the level of human intra-observer variability (κ = 0.71). The presented findings indicate that the continuously increasing HarmonicSS cohort will enable further advancements in AI-based pSS scoring methods by SGUS. In turn, this may establish SGUS as an effective noninvasive pSS diagnostic tool, with the final goal to supplement current diagnostic tests.**

***Index Terms*—Sjögren's syndrome, salivary glands, ultrasonography, radiomics.**

A. M. Vukicevic and N. Filipovic are at the Faculty of Engineering, University of Kragujevac, Serbia (e-mail: arso_kg@yahoo.com, fica@kg.ac.rs).

V. Milic is at the Institute of Rheumatology, School of Medicine, University of Belgrade, Serbia (e-mail: veramilic1409@gmail.com).

A. Zabotti and S. De Vita are at the Azienda Ospedaliero Universitaria, Santa Maria Della Misericordia di Udine, DAME, Italy (e-mail: zabottialen@gmail.com, salvatore.devita@asuiud.sanita.fvg.it).

A. Hocevar is at the Department of Rheumatology, Ljubljana University Medical Centre, Slovenia (e-mail: alojzija.hocevar@gmail.com).

O. De Lucia is at the Department of Rheumatology, Clinical Rheumatology Unit, ASST Centro Traumatologico Ortopedico G. Pini-CTO, Milano, Italy (e-mail: orazio.delucia@asst-pini-cto.it).

G. Filippouis at the Section of Rheumatology, Department of Medical Sciences, University of Ferrara, Italy (e-mail: gf.filippou@gmail.com).

A. F. Frangi is with the Center for Computational Imaging & Simulation Technologies in Biomedicine, School of Computing, and Leeds Institute of Cardiac and Metabolic Medicine, School of Medicine, University of Leeds, UK (e-mail: a.frangi@leeds.ac.uk).

A. Tzioufas is with the Department of Pathophysiology, Medical School, National and Kapodistrian University of Athens, Greece (e-mail: agtzi@med.uoa.gr).

I. Introduction

Primary Sjögren's Syndrome (pSS) is a chronic autoimmune disease, whose manifesting symptoms are oral and ocular dryness, fatigue, arthralgia and arthritis. The annual incidence of pSS has been estimated at a range from 200 to 3000 per 100.000 people, with highly unbalanced gender ratio (~10 females per 1 male) [1]. Standardization of the pSS classification has been the subject of debate for decades. Chronologically, four standardized guides are: European Classification (PEC) criteria [2], American European Consensus Group (AECG 2002) classification criteria [3] the American College of Rheumatology (ACR 2012) criteria [4] and the more recent ACR-European League Against Rheumatism (EULAR) 2016 criteria [5]. Briefly, these guides are based on the combination of examined clinical symptoms, results of autoantibody tests and salivary gland (SG) biopsy [6]. All these criteria do not incorporate new insights in pSS enabled by noninvasive salivary gland ultrasonography (SGUS) [7]. According to clinical reports, failing to include any imaging modalities (as mentioned in the standardized guides) has been reported as an obstacle in the practice – as patients frequently complain at invasive tests and biopsies, especially during follow-up studies or when presented with negative findings [8].

Up until now, various SGUS-based pSS scoring approaches have been introduced and showed satisfactory results in comparison to both ACR 2012 and AECG 2002 [9]. The proposed approaches are based on the visual observation of parotid and submandibular SGs' characteristics from SGUS. These scores are further subtracted and compared to the cut-off threshold to determine the final pSS score [10-15]. In order to investigate human-dependency, international experts have recently participated in the consensus meetings with the aim to evaluate reliability of SGUS echo structural parameters [16]. Considering the obtained results of inter/intra observer reliability, it is concluded that there is still no gold standard for pSS diagnosis based on the observation of echo structural abnormalities in SGUS images.

In order to resolve these obstacles, leading SS experts (35 partners from 13 countries) have recently started the HarmonicSS (http://harmonicss.eu) initiative. The aim of the joint European research initiative is to envelop independently reported cohorts and metacentric data with the end goal to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

TABLE I
CHARACTERISTICS OF CLINICAL DATA USED IN THIS STUDY.

Characteristics	pSS Patients	SGUS images (four images were acquired for each patient)		
	All (N=150)	Train (N=500)	Test (N=100)	P value
Age (years)	54 ± 15	53.8 ± 12	55 ± 11	0.702
Sex (female), n (%)	140 (93.3)	468 (93.6)	92 (92)	0.558
Disease duration (years)	7 ± 4.5	7.3 ± 4.1	6.4 ± 3.8	0.401
Ocular symptoms, n (%)	131 (87.9)	445 (89)	82 (82)	0.051
Oral symptoms, n (%)	129 (85.9)	425 (85)	90 (90)	0.190
Positive biopsy of MSG*, n/91 (%)	86/91 (94.5)	287/303 (95)	57/61 (93)	0.526
Anti-SSA*, n/146 (%)	102/146 (69.8)	331/486 (68)	77/98 (78)	0.050
Anti-SSB*, n/N=127 (%)	64/127 (50.3)	221/423 (52)	35/85 (41)	0.064
De Vita scores (0, 1, 2, 3) distribution (%)	/	31, 11, 41, 17	25, 25, 25, 25	/

N indicates a number of subjects; n indicates a number of positive findings.
*Tests were not performed on all N subjects. Percentage was computed with respect to the number of examined subjects: n / number of examined subjects (%).

ease further progress in the diagnosis and treatment of pSS. As recently suggested, one of desirable advances in SGUS is the development of dedicated computerized tools that could reduce screening time and dependency on human experts [17]. To the best of our knowledge, there is still no available solution with such ability on the market, nor reported in the scientific literature. By using the growing HarmonicSS cohort, the aim of the present study was to propose a novel radiomics-based approach for the assessment of pSS in SGUS images.

II. MATERIALS

After obtaining the institutional review board approvals, we retrospectively reviewed medical records of 150 patients from two clinical centers in Europe: Belgrade (Serbia) and Udine (Italy). US examinations assumed routine acquisition of parotid and submandibular glands longitudinal scans. Since four images were acquired for each patient, the cohort included 600 SGUS images. Belgrade clinical center contributed with 112 patients (448 images) examined with GE LogiqE9 device with a linear high-frequency transducer (6-15MHz), while the center from Udine provided 38 patients (152 images) examined with the ESAOTE MyLabClassC US machine with a linear high frequency probe (6-18MHz).

All subjects underwent a diagnostic work-up for pSS according to the AECG [3]. The evaluation included the following: 1) questionnaire with six questions to assess ocular and oral symptoms, 2) evidence of dry eye (Rose-Bengal), 3) presence of anti-SS-A/SS-B antibodies, 4) sialoscintography for the evidence of salivary dysfunction and 5) biopsy of minor salivary glands. Characteristics of patients involved in this study are shown in Table I.

In this study, pSS scores of SGUS images were defined using the original scoring system proposed by De Vita et al. [14]. This easy-to-apply score was chosen because adequate discriminant analyses were employed to select the items to build the score itself, and these items were subsequently confirmed to be of value. Since the scoring is expert-based approach, ground truth values were defined by using the Delphi method. The scoring assumed grading images on the 0-3 scale regarding SGs' echo structural characteristics, as proposed in Luciano et al. [14]. SGUS images were randomized and assessed twice by five independent clinicians, whose expertise varied between experienced to leading rheumatologists in the field. After the definite scores were obtained by the experts' consensus, the resulting class distribution in the database was 30%, 13%, 39% and 18%, respectively (class distribution of images used for the development of train and test sets are given in Table I).

A. Reliability of pSS clinical assessment in SGUS images

Intra-observer and inter-observer reliability were assessed using the kappa coefficient. The intra-observer agreement was measured using the Cohen's weighted kappa, showing the substantial agreement: $\kappa = 0.71 \pm 0.11$ ($\kappa_{min} = 0.58$ and $\kappa_{max} = 0.88$). The overall inter-observer agreement (before the expert consensus) was measured using the Scott/Fleiss' kappa ($\kappa = 0.61$).

In order to compare the performances of proposed radiomics-based algorithms with clinicians, the performance indicators given in this paragraph were calculated with respect to scores adapted through the expert consensus. The mean Pearson's correlation of the five observers with ground truth was $R^2 = 0.690 \pm 0.137$ ($R^2_{min}=0.461$ and $R^2_{max}= 0.837$). The percentage agreement was 70.1 ± 10.3% (within the range 55.6 – 82.1%). The agreement of observers with ground truth was measured using the weighted Cohen's kappa $\kappa = 0.66 \pm 0.14$ ($\kappa_{min} = 0.44$ and $\kappa_{max} = 0.83$).

III. METHODS

Characteristics that are specific to the assessment of pSS from SGUS images are: a) high variance of SGs in appearance, shape and size, and b) low relevance of SGs' surrounding tissues for the diagnosis. Considering the cohort size, and these requirements, we propose a novel radiomics-based procedure as a suitable approach for solving the given problem. The procedure's workflow is sketched in Fig. 1 and Fig. 2, while its composing steps are described in the rest of this section.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

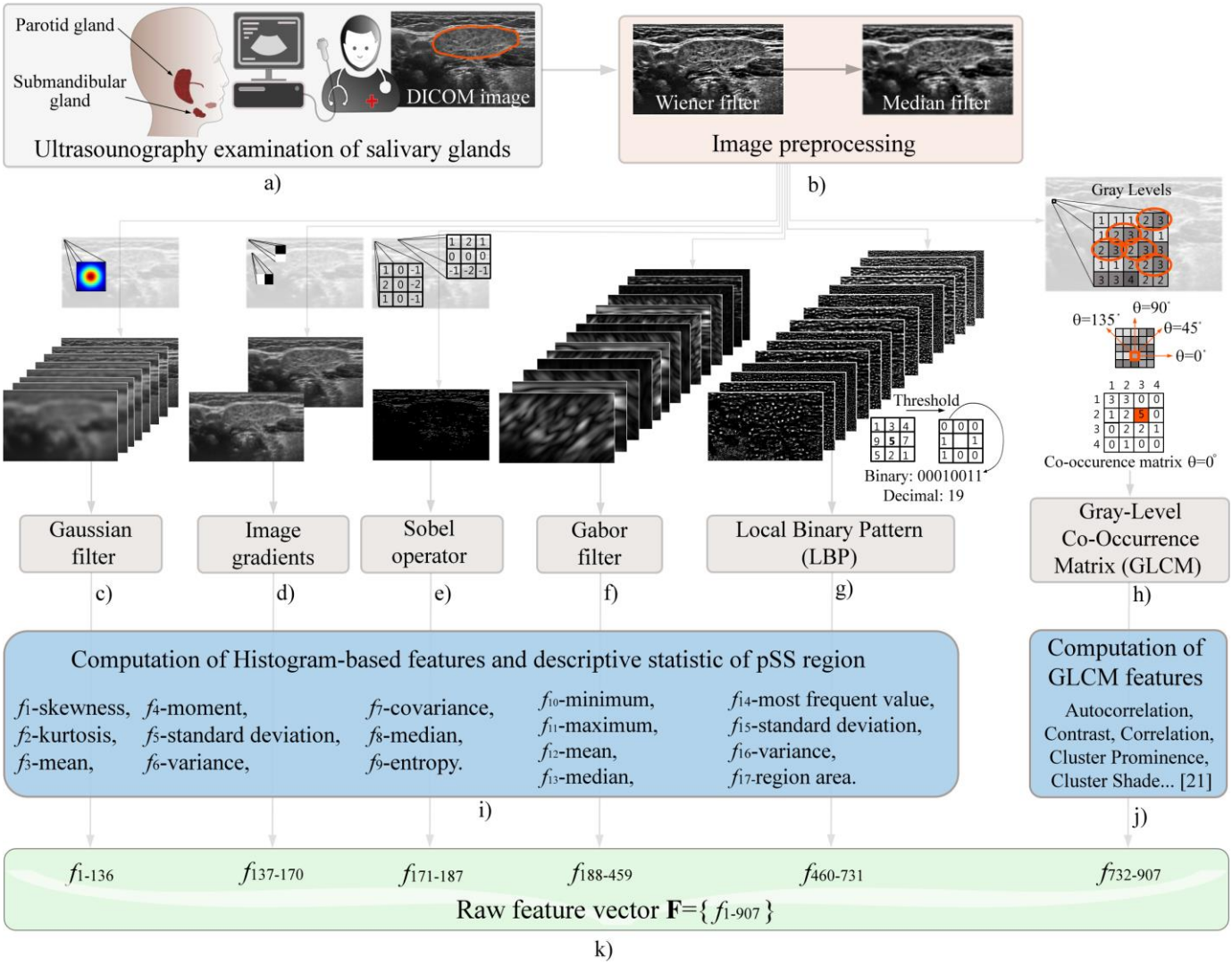

Fig.1. Overview of the feature extraction procedure. From a segmented SG region, radiomics features were extracted by using: Gaussian filter, image gradients, Sobel operator, Gabor filter, Local Binnary Patter and Gray Level Coocurence Matrix. The raw feature vector **F** consisted of a total 907 features, which were obtained by varying parameters of the considered feature extractors.

A. Features extraction

Semi-automatic segmentation of SGs was performed using the Snake algorithm (Fig. 1(a)) [18]. The presence of artifacts and pepper noise were reduced by using the Wiener and Median filters, respectively (Fig. 1(b)). The extraction of radiomics features from the segmented SGUS image was performed using a series of algorithms (Fig 1(c-j)). Invariance of the proposed procedure to both size and shape of the segmented SGs was ensured by computing: a) Histogram-based features $f_{1\text{-}9}$ (after expressing the bin counts in percentage, following Fig. 1(i)); and b) Descriptive statistics features $f_{10\text{-}19}$ that account for pixels inside the segmented SG region (Fig. 1(j)). The considered feature-extractors were:

1) Multi-resolution Image Gaussian Pyramid

The Gaussian filter (GF) is a 2D convolutional smoothing operator, whose kernel was generated using the Gaussian function (Fig. 1 (c)):

$$\mathbf{F}_{\text{Gaussian}}(x,y) = \frac{1}{2\pi\sigma^{e}} e^{-\frac{x^2+y^2}{2\sigma^2}} * \mathbf{I} \quad (1)$$

where $\mathbf{F}_{\text{Gaussian}}$ is the filtered image, * is the convolution operator, **I** is the image, whereas x and y are pixels coordinates in the local Gaussian filter space. By varying the σ parameter in the range ~0:2:16, we obtained eight new filtered images. From each of them we extracted 17 histogram-based features listed in Fig. 1(i) – which represent the feature vector $f_{1\text{-}136}$ in Fig. 1(k).

2) Image gradients and Sobel operator

Each image was convolved with two gradient operators with the 6x6 kernel (Fig. 1(d)); one detecting horizontal gradients ($f_{137\text{-}153}$) and the other detecting vertical gradients ($f_{154\text{-}170}$). Additionally, we processed each image using the Sobel operator (Fig. 1(e)) with the 3x3 kernel ($f_{171\text{-}187}$ in Fig. 1 (k)).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

3) Multi-resolution Gabor representation

The Gabor filter represents a two-dimensional sinusoidal wave (with predefined orientation and wavelength), whose amplitude is multiplied with the Gaussian function [19]:

$$\mathbf{F}_{\text{Gabor}}(x,y,\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(-\frac{x'^2+y^2y'^2}{2\sigma^2}\right)\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \quad (2)$$

where λ is the length of the wave, θ is the wave orientation (so that $x'=x\cos\theta+y\sin\theta$ and $y'=-x\sin\theta+y\cos\theta$), ψ is the phase shift, σ is standard deviation of the Gaussian function and γ is the factor that control elasticity of the filter. We created a bank of 16 Gabor filters (Fig. 1(f)) generated by varying the orientation $\{0,\ \pi/4,\ \pi/2,\ 3\pi/4\}$ and frequency $\{0.12,\ 0.16,\ 0.24,\ 0.32\}$ while Gaussian standard deviation had values $\sigma_x=\{1,\ 2,\ 2,\ 1\}$ and $\sigma_y=\{1,\ 2,\ 4,\ 2\}$. The extracted set of features is marked as $f_{188\text{-}459}$ in Fig. 1(k).

4) Local Binary Pattern (LBP)

The LBP features were computed following the steps in the literature [20]. Briefly, for each pixel LBP compares its value to N pixels along a surrounding circle with a diameter R. If the centre pixel's value is greater than the circle neighbour's value, LPB writes 0, otherwise it writes 1. This gives an N-digit binary number, which is finally converted to decimal for convenience, as sketched in Fig. 1(g). In the present study, we generated the feature set $f_{460\text{-}731}$ in Fig. 1(k) by varying N=8:8:32 and R=4:4:16.

5) Gray-level co-occurrence matrix (GLCM)

The GLCM is a statistical method used for characterization of a texture. GLCM calculates how often pairs of pixels with specific values and in a specified spatial relationships occur in an image. It creates a GLCM matrix (Fig. 1(h)), and then extracts 22 statistical features from the matrix (Fig. 1(j)) [21]. In the present study, we set the number of levels to 10 and offsets to [0 5; -5 5; -5 0; -5 -5; 0 1; -1 1; -1 0; -1-1], resulting in the feature set $f_{732\text{-}907}$ in Fig. 1(k).

B. Data stratification

SGUS images were stratified on the training-learning and independent test sets, to more rigorously assess the generalization ability of the proposed procedure. Both data sets were computed only once and saved, so that they could be loaded on-demand during the development and evaluation of the considered predictive models.

1) Development of balanced independent test set

Since we deal with the development of multiclass predictor using the imbalanced cohort, the size of the independent test-set was determined with the size of the most under-sampled class. In this way, we ensured that the sufficient amount of balanced and real-word samples of each class are available during both learning stage and subsequent independent testing. We created the balanced test-set by randomly sampling 25 images (approximately 30% of the least presented class – grade 1) from each grading category (Fig. 2(a)). The remaining data was further used as the training-set.

2) Development of balanced K-folds for the cross-validation

The training was performed using the k-fold cross-validation. Accordingly, the training set was divided into k=6 folds, ensuring that each fold consisted of the same number of

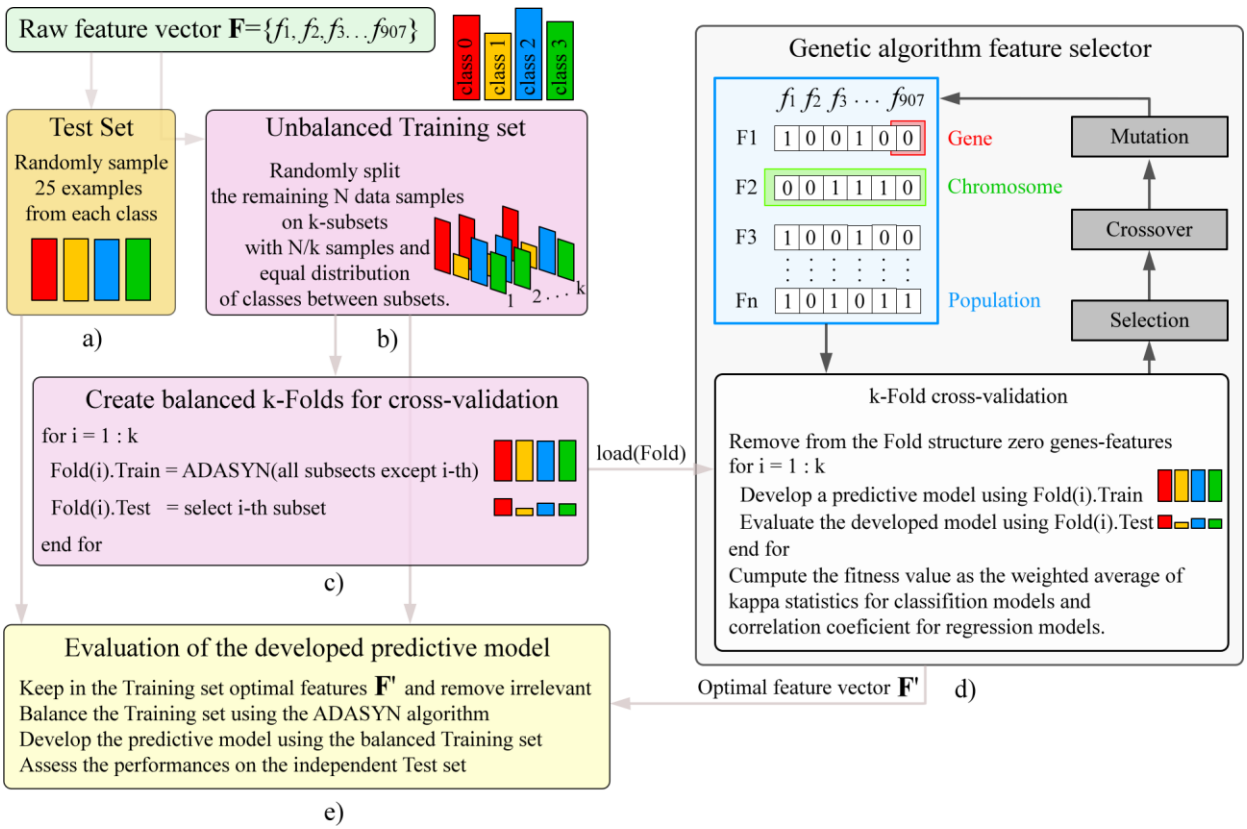

Fig.2. Procedure for the development of predictive models and selection of optimal features subset using Genetic algorithm-based wrapper.

samples with equal distribution of classes (Fig. 2 (b)). Since the training set was unbalanced, we applied the ADASYN algorithm that sharpens boundaries between classes by generating synthetic samples in minority classes (Fig. 2 (c)) [22]. Therefore, the ADASYN improves learning by: 1) reducing the bias introduced by the class imbalance, and 2) adaptively shifting the classification decision boundary toward the difficult-challenging examples [11].

C. Selection of optimal features subset

Since each feature extractor depends on several (hyper) parameters, there is a risk of omitting to use relevant (combination of) features if feature extractors' parameters are not set correctly. This problem was solved by varying the extractors' parameters (section III-A) and using a robust supervised wrapper feature selector to subsequently find an optimal feature subset for a particular predictive model. The workflow of the proposed procedure is given in Fig. 2(d), whereas its key steps are explained in the following paragraphs.

1) Genetic algorithm-based wrapper

Genetic algorithm (GA) is the iterative method for solving optimization problems [23]. The process starts from an initial guess of parameters (called population) subjected for the optimization. At each iteration (called generation), the GA selects some portion of best individuals from the current population and uses them as parents to produce the candidates (called children) for the next generation (crossover and mutation). Over successive generations, this process leads to the evolution of populations of individuals that are better adapted to their environment than the individuals that they originated from (similar to natural adaptation).

In the present study, we employed the GA to select optimal features subset by considering the feature selection as the integer optimization problem within the bounds 0 and 1. In each call of the GA objective function, the predictive model was cross-validated using the previously created k-folds. Parameters of the objective function represented features selector (i.e. F2 chromosome in Fig. 2(d)), so that the parameters with value 1 indicated features that should be selected while parameters with value 0 indicated features to be neglected during the training. The value of the GA objective-function was the kappa-statistics for classification and the Pearson's correlation for the regression predictive models. In this study, population size was set to 600, number of generations was set to 500, while the rest of GA hyper parameters had default values defined within the Matlab *gaoptimset* function (see the Matlab online documentation).

2) Considered predictive models

The pSS scoring could be considered as both classification (the score is the ordinal value: 0,1,2 or 3) and regression (the score is any real number on the interval 0-3) problem. In order to find which one is the most efficient approach for the pSS scoring, we evaluated 7 classifiers and 5 regressors [24]: Decision Table (DT), J48 tree, K-nearest neighbors (KNN), Linear Regression (LinR), Logistic regression (LogR), Multilayer perceptron (MLP), Naive Bayes NET (NB_{NET}), Naive Bayes (NB) and Random forest (RF). The parameter settings for each of the predictive models were set iteratively, while the MLP was configured following the Evolutionary assembling approach [25].

IV. Results

The implementation of the proposed procedure was performed using the Matlab R2010 (MathWorks, Natick, MA) and Java wrapper for the Weka v. 3.8 library (University of Waikato) [13]. The computational time needed to find optimal features and develop predictive models varied among algorithms. In worst case scenarios, it took up to several hours on the Dell PowerEdge server (204 processors, 800GB RAM, 4.5TB SSD). After the learning process had been completed, execution of the developed algorithms for scoring newly supplied SGUS images was done almost in real-time.

Performances of the assessed algorithms are given in Table II. Calculated efficiency indicators were: Pearson's correlation (R^2), Mean absolute error (MAE), Root mean squared error (RMSE) - for regression-based algorithms; and: Accuracy (ACC, %), Area under the receiver operating characteristic curve (AUC), kappa-statistics (κ), MAE and RMSE –

TABLE II
PERFORMANCES OF THE CONSIDERED PREDICTIVE MODELS OBTAINED DURING THE CROSS-VALIDATION AND EVALUATION ON THE INDEPENDENT TEST SET.

	Regression			Classification				
	R^2	MAE	RMSE	ACC	κ	MAE	RMSE	AUC
J48	n /a	n /a	n /a	73.1/57.0	0.64/0.42	0.13/0.22	0.35/0.45	0.84/0.73
LogR	n /a	n /a	n /a	76.3/59.0	0.68/0.45	0.12/0.20	0.34/0.43	0.88/0.80
NB	n /a	n /a	n /a	69.0/57.0	0.58/0.42	0.16/0.21	0.36/0.44	0.88/0.80
NB_{NET}	n /a	n /a	n /a	57.0/55.0	0.42/0.4	0.23/0.24	0.38/0.40	0.83/0.82
RF	0.89/0.83	0.38/0.49	0.51/0.61	84.5/67.0	0.79/0.56	0.18/0.22	0.26/0.32	0.96/0.90
KNN	0.86/0.75	0.22/0.43	0.57/0.75	81.1/67.0	0.74/0.56	0.09/0.16	0.30/0.40	0.87/0.78
MLP	0.87/0.83	0.33/0.49	0.55/0.69	86.0/78.0	0.80/0.70	0.08/0.11	0.24/0.30	0.96/0.93
LinR	0.79/0.85	0.53/0.48	0.68/0.58	n /a	n /a	n /a	n /a	n /a
DT	0.70/0.75	0.58/0.55	0.80/0.73	n /a	n /a	n /a	n /a	n /a

R^2-Pearson's correlation, MAE-Mean absolute error, RMSE-Root mean squared error, ACC-Accuracy (%), κ -Kappa statistics.
Values of the performances' indicators are given as: train / test, AUC–Area under the receiver operating characteristic curve.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

TABLE III
SENSITIVITY ANALYSIS OF THE TOP RANKING PREDICTIVE MODELS ON ERRORS DURING THE SG SEGMENTATION. WE CONSIDERED THREE SCENARIOS: 1-OVERESTIMATED SG (SCALE UP 20%); 2-UNDERESTIMATED SG (SCALE DOWN 20%); AND 3- SEGMENTED SG CONTOUR IS TRANSLATED FOR [20 20], [-20 20], [20 -20] AND [-20 -20] PIXELS.

	RF	MLP	RF	MLP
	classification		regression	
Scenario	ACC / κ / MAE / RMSE /AUC	ACC / κ / MAE / RMSE /AUC	R^2 / MAE / RMSE	R^2 / MAE / RMSE
1	63.0 / 0.52 / 0.22 / 0.34 / 0.86	73.0 / 0.65/ 0.13 / 0.34 / 0.91	0.76 / 0.56 / 0.70	0.78 / 0.55 / 0.70
2	66.0 / 0.55 / 0.21 / 0.33 / 0.88	76.0 / 0.68 / 0.12 / 0.31 / 0.92	0.79 / 0.53 / 0.65	0.81 / 0.52 / 0.64
3	61.0 / 0.46 / 0.23 / 0.37 / 0.85	72.0 / 0.62 / 0.16 / 0.36 / 0.90	0.72 / 0.60 / 0.72	0.72 / 0.58 / 0.72

MAE-Mean absolute error, RMSE-Root mean squared error, ACC-Accuracy (%), **κ**-Kappa statistics.

calculated for classification-based algorithms.

After the identification of the top ranked algorithms, we further analyzed their sensitivity on errors during the SG segmentation (the only user dependent step). For each image in the test set, we automatically created six intra-observers' variability scenarios by: scaling up the segmented contours 20%, scaling down segmented contours 20%, as well as translating segmented contours in four directions for [20 20], [-20 20], [20 -20] and [-20 -20] pixels. The obtained results of the sensitivity analysis are given in Table III.

Histogram of the most frequently used features used for the development of considered predictive models is shown in Fig. 3, while sample results obtained by the top-ranked predictive models are shown in Fig. 4.

A. *Configuration of the top-performing MLP classifier*

The development of the MLP may be intuitively described as scheduling of its hyper parameters (type of activation functions in layers, learning rate, learning momentum, number of neurons per layer, training algorithm, number of learning epochs) with the aim to maximize the classification performances. In order to set these parameters automatically and correctly, we employed the recently proposed Evolutionary assembling approach [25]. The obtained MLP

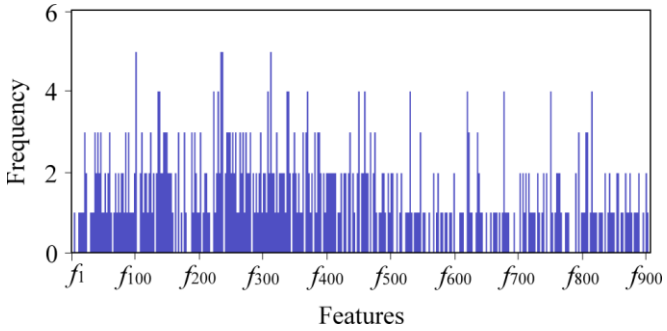

Fig.3. Histogram of features selected for the development of 12 considered predictive models (5 regressors and 7 classifiers).

configuration assumed: 41 neurons in the hidden layer, activation functions in both layers were tansig (hyperbolic tangent sigmoid), training algorithm was set to the trainscg (scaled conjugate gradient backpropagation), learning momentum was set to 0.83 while the maximum number of epochs for training was set to 921. The GA-based wrapper selected the following list of features as optimal subset for the development of MLP: f_{29}, f_{101}, f_{180}, f_{249}, f_{256}, f_{257}, f_{278}, f_{380}, f_{474}, f_{480}, f_{535}, f_{628}, f_{641}, f_{680}, f_{722}, f_{755}, f_{804}, f_{816}, f_{830}, f_{841} and f_{895} (histogram of features selected for the development of 12 considered predictive models are shown in Fig. 3). Robustness of the proposed procedure comes from the fact that hyper parameters of both feature extractors and MLP classifier are set automatically. Thus, we emphasize that it could be applied for solving a wide range of problems in computer aided diagnosis.

V. DISCUSSION

A. *Performances of top-ranked algorithms*

Depending on the type of 12 considered algorithms, number of features selected by the GA wrapper varied from 23 up to 187. Histogram in Fig. 3 indicates that there were no key radiomics features. Instead, the challenge was to find an optimal feature subset that maximizes performance of a particular predictor. Each of the developed predictive models was evaluated twice, with the cross-validation and on the independent test-set, in order to more rigorously assess the generalization performances. Results from Table II indicate that RF, KNN and MLP were top-ranked algorithms in both classification and regression categories. In the following sections, classification and regression approaches for pSS scoring will be discussed separately in order to highlight their benefits.

1) *Benefits from using classification-based algorithms*

The obtained results show that classification algorithms produce a lower mean absolute error and root mean squared error, which is important during the definite pSS classification (when clinicians consider scorings of four SGUS images acquired from a single patient). In such situations, procedures that are able to accurately grade 3 out of 4 images (over 75% accuracy) represent a considerable contribution to the current practice [2]. In our study, RF, KNN and MLP reached above 66% accuracy (guarantee that at least 2 of 4 images will be graded correctly). However, only the MLP classifier surpassed the threshold of 75% accuracy, which we recommend as the most reliable for the pSS diagnosis using the SGUS scoring system developed by De Vita et al. for pSS (14). In terms of the kappa-statistics, which is commonly used in pSS related studies, the MLP showed substantial agreement (κ=0.7) with the ground truth defined via the expert consensus.

2) *Benefits from using regression-based algorithms*

One of the most challenging issues related to the screening of pSS from SGUS is the follow-up of patients, when clinicians have to estimate the disease progress by inspecting two or more SGUS images. Although using scores in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2923773, IEEE Journal of Biomedical and Health Informatics

interval 0-3 is more appropriate for the follow-up (compared to the ordinal scale), it is difficult for clinicians to objectively perform such accurate estimation. In such situation, regression algorithms may appear as useful tools for assisting clinicians. In the present study, the RF and MLP regressors performed the best in terms of both Pearson's correlation (R^2=0.83) and RMSE (Table II) with respect to the ground truth defined via the expert consensus.

B. Sensitivity to errors in SGUS segmentation

Considering RF and MLP as top ranked algorithms, we first used the Stuart-Maxwell's test to prove that predictions of two multi-class classifier are statistically significant (χ^2=9.6 and p =0.022 values for the significance level of α=0.05). Furthermore, we analyzed RF and MLP classifiers sensitivity to errors that may occur during the SG segmentation (the only user dependent step). The obtained results in Table III showed that the proposed predictive models are robust on the SG underestimation (case 2). However, case 1 and case 3 types of inaccurate SG segmentation decreased accuracy of predictive models for 4-8%. Although larger segmentation errors are uncommon for the trained clinicians, the recommendation is to prefer underestimating SG when a user is presented with noisy SGUS images. In summary, we recommend the MLP as the most reliable predictive model for the assessment of De Vita scores from SGUS images.

C. Contribution to the state of the art

1) Computerized analysis of SGUS and assessment of pSS

After literature review, we report that the computerized medical image analysis of SG and pSS remain underestimated problems. Instead, the most of related work is focused on analyzing pSS biopsy images or other diseases present in SGs. Chernomordik et al. proposed a fluorescence scanning imaging system that performs a noninvasive optical biopsy of the Sjögren syndrome (based on the 2D CCD imaging of the lower lip), with the end-goal to replace the traditionally used histological biopsy [26]. Regarding the SGUS-based studies, Chikui et al. suggested using the fractal analyses to characterize SG tumors [27]. The same author afterwards reported that average size of the particles, area ratio of the particles within the region and Hurst-ori were useful predictors for detecting abnormal sialographic stages [28]. Siebers et al. performed multi-feature tissue characterization for differentiating malignant and benign parotid gland lesions using maximum likelihood supervised classifier [29]. Murakami et al. applied 2D wavelet analysis to SGUS images for the diagnosis of SS [30]. A couple of studies investigated the possibility of using the elastography techniques for diagnosing pSS in SGUS. Dejaco et al. used real-time sonoelastography of SGs for the diagnosis and assessment of glandular damage in pSS [31]. Zhang et al. assessed SG stiffness in pSS via acoustic radiation force impulse imaging [32].

Therefore, we found that currently there is a lack of methods for automated analysis and scoring of pSS in SGUS. This may be justified with a few facts: 1) studies that introduced scoring systems were mono-disciplinary (relied mostly on clinicians' experience in image analysis) [10-15]; and 2) the nature (rarity) of pSS makes it difficult for a single institution to collect a larger cohort appropriate for the training of robust AI algorithms. To the best of our knowledge and

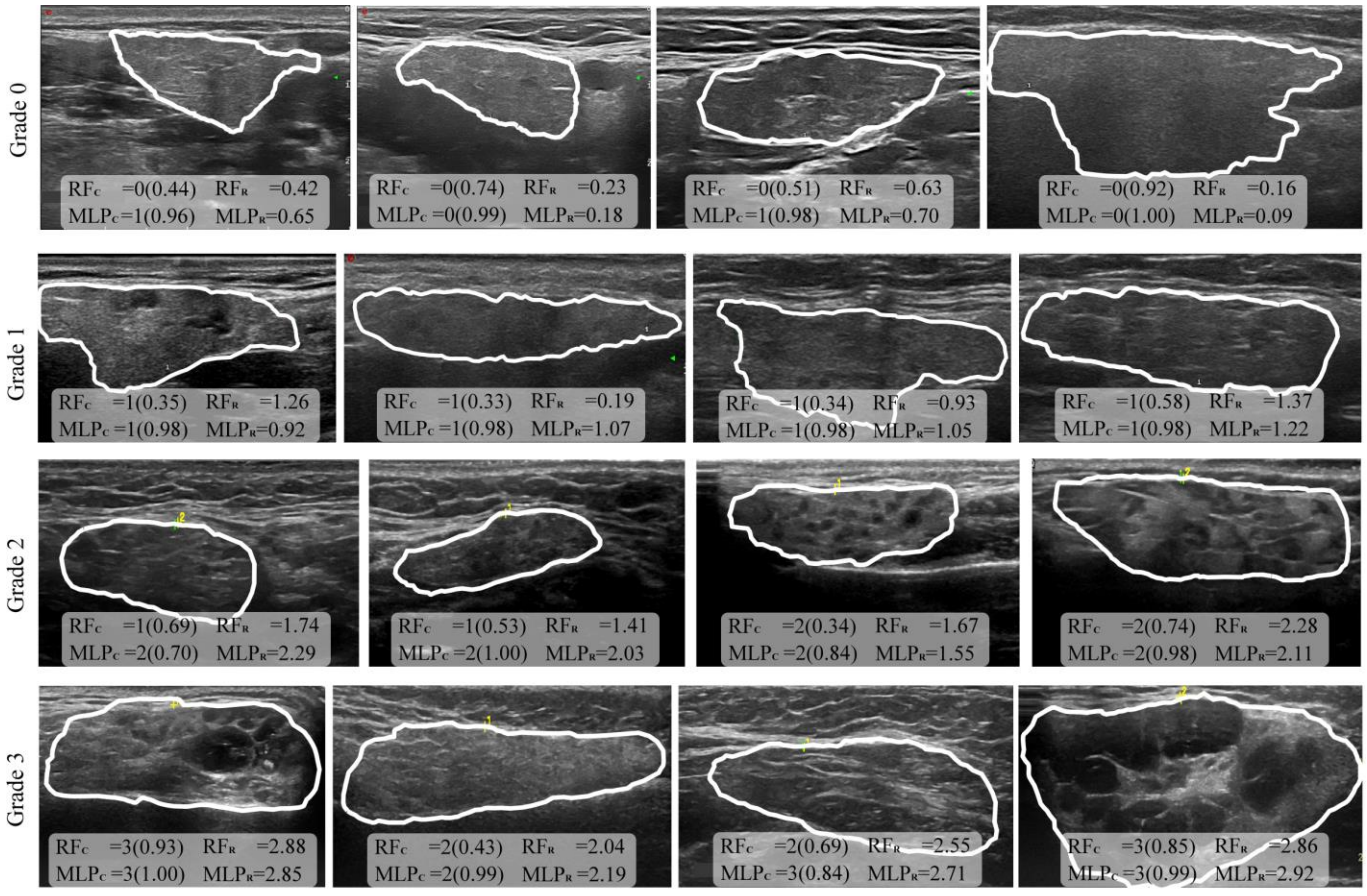

Fig.4. Samples of SGUS images and scores (probabilities) obtained by top-ranked classifiers (MLP_C, RF_C) and regressors (ANN_R, RF_R).

insight into the topic, the present study is the first one that proposes the procedure for computer-aided diagnosis and scoring of pSS from SGUS.

2) Performances of AI with respect to trained clinicians

By comparing the average performances of clinicians (average intra-observer agreement was $\kappa = 0.71$ and average agreement with ground truth was $\kappa = 0.67$) with the performances obtained with the proposed classifier, it may be found that the MLP classifier over-performed ($\kappa = 0.7$) clinicians by the considerable margin while its reliability is on the level of humans' intra-observer variability. Therefore, we confirm the hypothesis that the proposed AI-based procedure represents a potential improvement of healthcare standards present in most clinics worldwide [33]. Also, it is worth emphasizing that it still cannot compete with the most-skilled clinicians who are the leading scientists in the field ($\kappa_{max} = 0.83$ and intra-observer $\kappa_{max} = 0.88$). Regarding the regression-based assessment, the MLP (R^2=0.83) over-performed clinicians (R^2=0.69) by a large margin and it is on the level of leading experts (R^2_{max}= 0.837). In addition, while the score by De Vita was proposed as ordinal values, our findings may be a starting point towards developing a continuous scale that is more appropriate for the follow-up assessment of pSS and for the detection of changes.

3) Reliability of SGUS scoring systems

Incorporation of SGUS scoring systems into standardized diagnosing guides has been prolonged due to their dependency on experts. As a solution to this problem, the score by De Vita et al. was proposed as an easy and practical measure [14]. Our findings support this claim since the average intra observer reliability was quite good $\kappa = 0.71$, while the most experienced clinicians reached excellent results in terms of both intra $\kappa_{max} = 0.88$ and inter reliability $\kappa_{max} = 0.83$. Therefore, we report that the obstacle for wide acceptance of SGUS in pSS screening may be related to the lack of highly skilled sonographers rather than to the need for more suitable scoring systems. The present study confirmed this hypothesis and showed that the problem of clinicians intra/inter observer variability could be solved by employing AI-based algorithms. Particularly, AI algorithms could be trained from data annotated by highly skilled experts and afterwards they could be used to assist and ease the training of less experienced clinicians.

D. Future work on this topic

Further development and improvements of dedicated computerized software tools for the pSS assessment from SGUS may significantly advance the way of treating pSS by reducing the invasiveness, screening time and dependency on experts. We highlight the strong potential of applying such technology for assisting and training of novice clinicians, whose perfecting could improve and equalize the healthcare quality worldwide [33]. Although, at the current stage, we have achieved performance on the edge with trained clinicians [11], we refer to this study as the first milestone of the wider HarmonicSS initiative. Considering the size of the cohort at the moment, we have assessed the radiomics-based algorithms to prove our hypothesis. In the future work, we aim to automate both SG segmentation and pSS scoring by employing other AI methods that benefit from the ongoing cohort growth, like Deep learning methods [34].

VI. Conclusion

Although humans are efficient in high-level cognitive tasks, our limitation in performing lower-level vision tasks such as calculation of textures' statistics or differentiating shades of colors are well studied and depend on many factors (i.e. ageing, genetics, fatigue, environment, diseases and so on) [35]. As an alternative to using descriptive and linguistic nominal attributes to characterize pSS in SGUS images, the present study aimed to assess various radimomics-based AI algorithms for pSS scoring from SGUS using the score proposed by De Vita in pSS [14]. We found that the MLP classifier (κ=0.7) over-performed average score achieved by the clinicians (κ=0.67) by the considerable margin, while its reliability is on the level of humans' intra-observer variability (κ=0.71). We emphasize that the proposed procedures still cannot compete with the leading scientists in the field (κ_{max}=0.83 and intra-observer κ_{max}=0.88). With further increase in the HarmonicSS cohort and improvements, validation and democratization of the AI able to compete leading clinicians in the pSS scoring, SGUS could be established as a reliable assessment procedure supplementing or replacing currently used invasive diagnostic tests.

Acknowledgment

This study was funded by the Serbian government (grant agreements III41007 and ON174028) and EU Horizon 2020 RIA programme (HarmonicSS, grant 731944). The Titan V GPU used for this research was donated by the NVIDIA Corporation to A. M. Vukicevic.

References

[1] M. Ramos-Casals, P. Brito-Zerón, B. Kostov, A. Sisó-Almirall, X. Bosch, D. Buss, A. Trilla, J. Stone, M. A. Khamashta, and Y. Shoenfeld,"Google-driven search for big data in autoimmune geoepidemiology: analysis of 394,827 patients with systemic autoimmune diseases," *Autoimmunity Reviews*, vol. 14(8), pp. 670-9, Aug. 2015.

[2] C. Vitali, S. Bombardieri, H. M. Moutsopoulos, G. Balestrieri, W. Bencivelli, R. M. Bernstein, K. B. Bjerrum, S. Braga, J. Coll, S. De Vita, A. A. Drosos, M. Ehrenfeld, P. Y. Hatron, E. M. Hay, D. A. Isenberg, A. Janin, J. R. Kalden, L. Kater, Y. T. Konttinen, P. J. Maddison, R. N. Maini, R. Manthorpe, O. Meyer, P. Ostuni, Y.Pennec, J. U. Prause, A. Richards, B. Sauvezie, M. Schiødt, M. Sciuto, C. Scully, Y. Shoenfeld, F. N. Skopouli, J. S. Smolen, M. L. Snaith, M. Tishler, S. Todesco, G. Valesini, P. J. W. Venables, M. J. Wattiaux, and P. Youinou, "Preliminary criteria for the classification of Sjögren's syndrome. Results of a prospective concerted action supported by the European Community," *Arthritis Rheum.*, vol. 36(3), pp. 340-7, Mar. 1993.

[3] C. Vitali, S. Bombardieri, R. Jonsson, H. M. Moutsopoulos, E. L. Alexander, S. E. Carsons, T. E. Daniels, P. C. Fox, R. I. Fox, S. S. Kassan, S. R. Pillemer, N. Talal, M. H. Weisman, and European Study Group on Classification Criteria for Sjögren's Syndrome, "Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group," *Ann Rheum Dis.*, vol. 61(6), 554-8, Jun. 2002.

[4] S. C. Shiboski, C. H. Shiboski, L. Criswell, A. Baer, S. Challacombe, H. Lanfranchi, M. Schiødt, H. Umehara, F. Vivino, Y. Zhao, Y. Dong, D. Greenspan, A. M. Heidenreich, P. Helin, B. Kirkham, K. Kitagawa, G. Larkin, M. Li, T. Lietman, J. Lindegaard, N. McNamara, K. Sack, P. Shirlaw, S. Sugai, C. Vollenweider, J. Whitcher, A. Wu, S. Zhang, W. Zhang, J. Greenspan, T. Daniels, and Sjögren's International Collaborative Clinical Alliance (SICCA) Research Groups, "American

College of Rheumatology classification criteria for Sjögren's syndrome: a data-driven, expert consensus approach in the Sjögren's International Collaborative Clinical Alliance cohort," *Arthritis Care Res.*, vol. 64(4), pp. 475-87, Apr. 2012.

[5] C. H. Shiboski, S. C. Shiboski, R. Seror, L. A. Criswell, M. Labetoulle, T. M. Lietman, A. Rasmussen, H. Scofield, C. Vitali, S. J. Bowman, X. Mariette, and International Sjögren's Syndrome Criteria Working Group, "2016 American College of Rheumatology/European League Against Rheumatism Classification Criteria for Primary Sjögren's Syndrome: A Consensus and Data-Driven Methodology Involving Three International Patient Cohorts,"*Ann. Rheum. Dis*., vol.76(1), pp. 9-16, Jan. 2017.

[6] A. Rasmussen, J. A. Ice, H. Li, K. Grundahl, J. A. Kelly, L. Radfar, D. U. Stone, K. S. Hefner, J. M. Anaya, M. Rohrer, R. Gopalakrishnan, G. D. Houston, D. M. Lewis, J. Chodosh, J. B. Harley, P. Hughes, J. S. Maier-Moore, C. G. Montgomery, N. L. Rhodus, A. D. Farris, B. M. Segal, R. Jonsson, C. J. Lessard, R. H. Scofield, and K. L. Sivils, "Comparison of the American-European Consensus Group Sjogren's syndrome classification criteria to newly proposed American College of Rheumatology criteria in a large, carefully characterized sicca cohort," *Ann Rheum Dis.*, vol. 73(1), pp. :31-8, Jan. 2014.

[7] H. Bootsma, F. K. Spijkervet, F. G. Kroese, and A. Vissink, "Toward new classification criteria for Sjögren's syndrome?," *Arthritis & rheumatism*, vol. 65(1), pp. 21–23, Jan. 2013.

[8] C. Vitali, H. Bootsma, S. J. Bowman, T. Dorner, J. E. Gottenberg, X. Mariette, M. Ramos-Casals, P.Ravaud, R.Seror, E.Theander, and A. G. Tzioufas, "Classification criteria for Sjögren's syndrome: we actually need to definitively resolve the long debate on the issue," *Annals of the Rheumatic Diseases*, vol. 72, pp. 476-478, Apr. 2013.

[9] S. Jousse-Joulin, V. Milic, M. V. Jonsson, A. Plagou, E. Theander, N. Luciano, P. Rachele, C. Baldini, H. Bootsma, A. Vissink, A. Hocevar, S. De Vita S, A. G. Tzioufas, Z. Alavi, S. J. Bowman, V. Devauchelle-Pensec V14, and US-pSS Study Group, "Is salivary gland ultrasonography a useful tool in Sjögren's syndrome? A systematic review," Rheumatology, vol. 55(5), pp. 789–800, May 2016.

[10] A. Hocevar, A. Ambrozic, B. Rozman, T. Kveder, and M.Tomsic, "Ultrasonographic changes of major salivary glands in primary Sjogren's syndrome. Diagnostic value of a novel scoring system," *Rheumatology*,vol. 44(6), pp. 768-72, Jun. 2005.

[11] F. Salaffi, M. Carotti, A. Iagnocco, F. Luccioli, R. Ramonda, E. Sabatini, M. De Nicola, M. Maggi, R. Priori, G. Valesini, R. Gerli, L. Punzi, G. M. Giuseppetti, U. Salvolini, and W. Grassi, "Ultrasonography of salivary glands in primary Sjögren's syndrome: a comparison with contrast sialography and scintigraphy," *Rheumatology*, vol. 47(8), pp. 1244-9, Aug. 2008.

[12] V. D. Milic, R. R. Petrovic, I. V.Boricic, J. Marinkovic-Eric, G. L. Radunovic, P. D. Jeremic, N. N. Pejnovic, and N. S.Damjanov, "Diagnostic value of salivary gland ultrasonographic scoring system in primary Sjogren's syndrome: a comparison with scintigraphy and biopsy,"*J Rheumatol.*, vol. 36(7), pp. 1495-500, Jul. 2009.

[13] V. D. Milic, R. R. Petrovic, I. V. Boricic, G. L. Radunovic, N. N. Pejnovic, I. Soldatovic, and N. S. Damjanov, "Major salivary gland sonography in Sjögren's syndrome: diagnostic value of a novel ultrasonography score (0–12) for parenchymal inhomogeneity," *Scandinavian Journal of Rheumatology*, vol. 39(2), pp. 160-166, Mar. 2009.

[14] S. De Vita, G. Lorenzon, G. Rossi, M. Sabella, and V. Fossaluzza,"Salivary gland echography in primary and secondary Sjögren's syndrome," *ClinExpRheumatol.*, vol. 10(4), pp. 351-6, Aug. 1992.

[15] N. Luciano, C. Baldini, G. Tarantini, F. Ferro, F. Sernissi, V. Varanini, V. Donati, D. Martini, M. Mosca, D. Caramella, and S. Bombardieri, "Ultrasonography of major salivary glands: a highly specific tool for distinguishing primary Sjögren's syndrome from undifferentiated connective tissue diseases," *Rheumatology*, vol. 54(12) pp. 2198–2204, Dec. 2015.

[16] S. Jousse-Joulin, E. Nowak, D. Cornec, J. Brown, A. Carr, M. Carotti, B. Fisher, J. Fradin, A. Hocevar, M. V. Jonsson, N. Luciano, V. Milic, J. Rout, E. Theander, A. Stel, H. Bootsma, A. Vissink, C. Baldini, A. Baer, W. F. Ng, S. Bowman, Z. Alavi, A. Saraux, and V. Devauchelle-Pensec, "Salivary gland ultrasound abnormalities in primary Sjögren's syndrome: consensual US-SG core items definition and reliability," *RMD Open.*, vol. 3(1), e000364, Jun. 2017.

[17] C. Baldini, A. Zabotti, N. Filipovic, A. M. Vukicevic, N. Luciano, F. Ferro, M. Lorenzon, and S. De Vita, "Imaging in primary Sjögren's syndrome: the "obsolete and the new," *ClinExpRheumatol*., vol. 36 (Suppl. 112):S215-S221, May-Jun 2018.

[18] C. Xu, and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Trans Image Processing*, vol. 7(3), pp. 359-69, 1998.

[19] D. Gabor, "Theory of Communication. Journal of Institution of Electrical Engineers," vol. 93, pp. 429-457, 1946.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(7), pp. 971-987, Aug. 2002.

[21] R. M. Haralick, K. Shanmugam, and Its'Hak Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics SMC*, vol. 3(6), pp. 610-621, Nov. 1973.

[22] H. He, Y. Bai, E. A. Garcia, and L. Shutao, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." Neural Networks, Publishe2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, June 1-8, 2008, Paper 10365271.

[23] X. S. Yang, *Nature-inspired optimization algorithms*, Elsevier, 2014, ch 4.

[24] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition, 2016.

[25] A. M. Vukicevic, G. R.Jovicic, M. M. Stojadinovic, R. I. Prelevic, N. D. Filipovic, "Evolutionary assembled neural networks for making medical decisions with minimal regret: Application for predicting advanced bladder cancer outcome," *Expert Systems with Applications*, vol. 41(18), pp. 8092-8100, Dec. 2014.

[26] V. Chernomordik, D. Hattery, I. Gannot, and A. H. Gandjbakhche, "Inverse method 3-D reconstruction of localized in vivo fluorescence-application to Sjogren syndrome," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 5(4), pp. 930-935, Jul/Aug. 1999.

[27] T. Chikui,K. Tokumori, K. Yoshiura, K. Oobu, S. Nakamura, and K. Nakamura, "Sonographic texture characterization of salivary gland tumors by fractal analyses," *Ultrasound in Medicine and Biology*, vol. 31(10), pp. 1297-304, Oct. 2005.

[28] T. Chikui,M. Shimizu, T. Kawazu, K. Okamura, T. Shiraishi, and K. Yoshiura, "A quantitative analysis of sonographic images of the salivary gland: a comparison between sonographic and sialographic findings," *Ultrasound in Medicine and Biology*, vol. 35(8), pp.1257-64, Aug. 2009.

[29] S. Siebers,J. Zenk, A. Bozzato, N. Klintworth, H. Iro, and H. Ermert, "Computer aided diagnosis of parotid gland lesions using ultrasonic multi-feature tissue characterization," *Ultrasound in Medicine and Biology*," 36(9), pp. 1525-34, Sep. 2010.

[30] Y. Murakami, A. Shiraishi, T. Sumi, T. Nakamura, and M. Ohki, "SU-E-I-115: Wavelet Analysis of Ultrasound Image for the Diagnosis of Sjögren's Syndrome," *Medical Physics*, vol. 39(6), P5(3651), 2012.

[31] C. Dejaco, T. De Zordo, D. Heber, W. Hartung, R. Lipp, A. Lutfi, M. Magyar, D. Zauner, A. Lackner,C. Duftner, J. Horwath-Winter, W. B. Graninger, and J. Hermann, "Real-time sonoelastography of salivary glands for diagnosis and functional assessment of primary Sjögren's syndrome," *Ultrasound in Medicine and Biology*, vol. 40(12), pp. 2759-67, Dec. 2014.

[32] S. Zhang, J. Zhu, X. Zhang, J. He, and J. Li, "Assessment of the Stiffness of Major Salivary Glands in Primary Sjögren's Syndrome through Quantitative Acoustic Radiation Force Impulse Imaging," *Ultrasound in Medicine and Biology*, vol. 42(3), pp. 645-53, Mar. 2016.

[33] T. C. Ricketts, "The Health Care Workforce: Will It Be Ready as the Boomers Age? A Review of How We Can Know (or Not Know) the Answer," *Annu Rev Public Health*, vol. 32, pp. 417-430, Jan. 2011.

[34] K. Wang, X. Lu, H. Zhou, G. Yongyan, Z. Jian, T. Minghui, W. Changjun, L. Changzhu, H. Liping, J. Tian'an, M. Fankun, L. Yongping, A. Hong, X. Xiao-Yan, Y. Li-ping, L. Ping, T. Jie, and Z. Rongqin, "Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study," *Gut*, vol. 68, pp.729-741, Mar. 2019.

[35] M. H. Pirenne and E. J. Denton, "Accuracy and Sensitivity of the Human Eye," *Nature*, vol. 170, pp. 1039–1042, Dec. 1952.