Confident texture-based laryngeal tissue classification for early-stage diagnosis support

Sara Moccia[a,b,*], Elena De Momi[a], Marco Guarnaschelli[a], Matteo Savazzi[a], Andrea Laborai[c], Luca Guastini[c], Giorgio Peretti[c], Leonardo S. Mattos[b]
[a]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
[b]Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy
[c]Department of Otorhinolaryngology, Head and Neck Surgery, University of Genoa, Genoa, Italy

Abstract. Early-stage diagnosis of laryngeal squamous cell carcinoma (SCC) is of primary importance for lowering patient mortality or after treatment morbidity. Despite the challenges in diagnosis reported in the clinical literature, few efforts have been invested in computer-assisted diagnosis. The objective of this paper is to investigate the use of texture-based machine-learning algorithms for early-stage cancerous laryngeal tissue classification. To estimate the classification reliability, a measure of confidence is also exploited. From the endoscopic videos of 33 patients affected by SCC, a well-balanced dataset of 1320 patches, relative to 4 laryngeal tissue classes, was extracted. With the best performing feature, the achieved median classification recall was 93% (inter-quartile range (IQR) = 6%). When excluding low-confidence patches, the achieved median recall was increased to 98% (IQR = 5%), proving the high reliability of the proposed approach. This research represents an important advancement in the state of art of computer-assisted laryngeal diagnosis and the results are a promising step toward a helpful endoscope-integrated processing system to support the early-stage diagnosis.

Keywords: Laryngeal cancer; tissue classification; texture analysis; surgical data science.

***Sara Moccia, sara.moccia@polimi.it**

1 Introduction

Squamous cell carcinoma (SCC) is the most common cancer of the laryngeal tract, arising from 95% to 98% of all cases of laryngeal cancer.[1] It is well know from medical literature that early-stage SCC diagnosis can lower mortality rate and preserve both laryngeal anatomy and vocal fold function.[2] Histopathological examination of tissue samples extracted with biopsy is currently the gold-standard for diagnosis. However, the relevance of tissue visual analysis for screening purpose has led, in the past few years, to the development of new optical-biopsy techniques, such as narrow-band imaging (NBI) endoscopy,[3] which has become the state of the art for laryngeal tract inspection. The identification of suspicious tissues during the endoscopic examination is, however, challenging due to the late onset of symptoms and to the small modifications of the mucosa, which

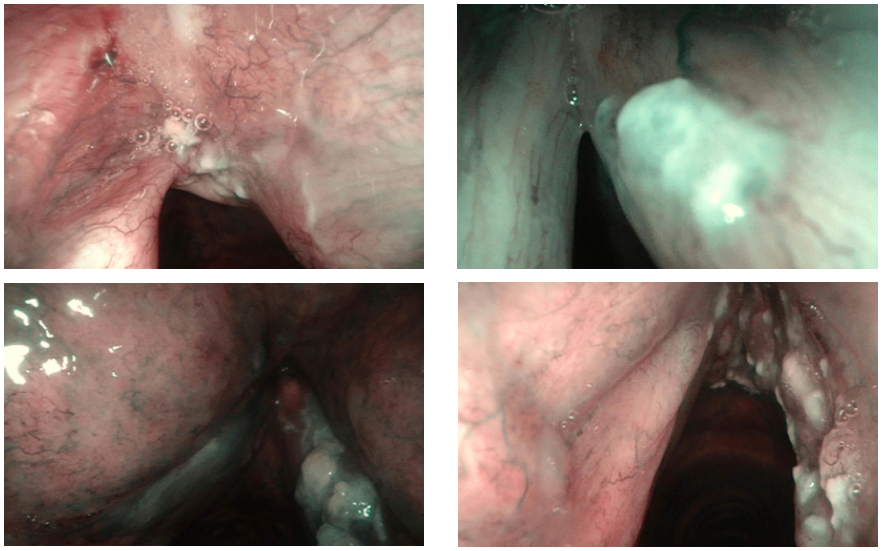


Fig 1: Visual samples of narrow-band imaging laryngeal endoscopic frames of patients affected by squamous cell carcinoma.

can pass unnoticed to the human eye.[4] Main modifications occur to the mucosa vascular tree, with the presence of longitudinal hypertrophic vessels and dot-like vessels, known as intraepithelial papillary capillary loops (IPCL).[3] Changes in the epithelium aspect not related to the vascular tree, such as thickening and whitening of the epithelial layer (leukoplakia), are associated with increased risk of developing SCC, too.[5] Visual samples of laryngeal endoscopic video frames of patients affected by SCC are given in Fig. 1.

Considering the clinical challenges in diagnosis, some preliminary attempts of computer-assisted diagnosis have been presented (Refs. 6,7), despite only Barbalata et al.[6] specifically focus on early-stage diagnosis. The study proposes an algorithm for the classification of early-stage vocal fold cancer based on the segmentation and analysis of blood vessels. Vessel segmentation is performed with matched filtering (MF) coupled with first order derivative of Gaussian. Vessel tortuosity, thickness and density are used as features to discriminate between malignant and benign tissue by means of linear discriminant analysis (LDA). Despite the good results (overall classification accuracy $= 84\%$), the classification proposed in Ref. 6 is strongly sensitive to a-priori set parameters,

e.g., vessel width and orientation. Moreover, focusing on vessels alone does not allow to take into account epithelial modifications that do not affect the vascular tree (i.e., in case of leukoplakia).

The emerging and rich literature on surgical data science for tissue classification outside the field of laryngoscopy has recently focused on more sophisticated techniques, which mainly exploit machine learning algorithms to classify tissues according to texture-based information.[8] In Ref. 9, the histogram of local binary patterns (LBP) is exploited to classify ulcer and healthy regions in capsule endoscopy images using multilayer perceptron. In Ref. 10, the LBP histogram is combined with intensity-based features to classify abdominal tissues in laparoscopic images by means of support vector machines (SVM). Similarly, in Ref. 11, intensity-based features and LBP histogram are used to characterize lesions in gastric images. In Ref. 12, the LBP histogram is combined with gray-level co-occurrence matrix (GLCM)-based features to classify gastroscopy images. AdaBoost is used to perform the classification. In Ref. 13, Gabor filter-based features are used to classify healthy and cancerous tissue in gastroscopy images by means of SVM. A recent work[14] exploits NBI data for colorectal image analysis. Colorectal tissues are classified as neoplastic or healthy by means of GLCM-based features and SVM.

Inspired by these recent and promising studies, in this paper we aim at investigating if texture-based approaches applied to laryngeal tissue classification in NBI images can provide reliable results, to be used as support for early-stage diagnosis. Specifically, we investigate the following two hypotheses:

- Hypothesis 1 (**H1**): Machine-learning techniques can classify laryngeal tissues in NBI images by exploiting textural information;
- Hypothesis 2 (**H2**): By estimating the level of classification confidence and discarding low-

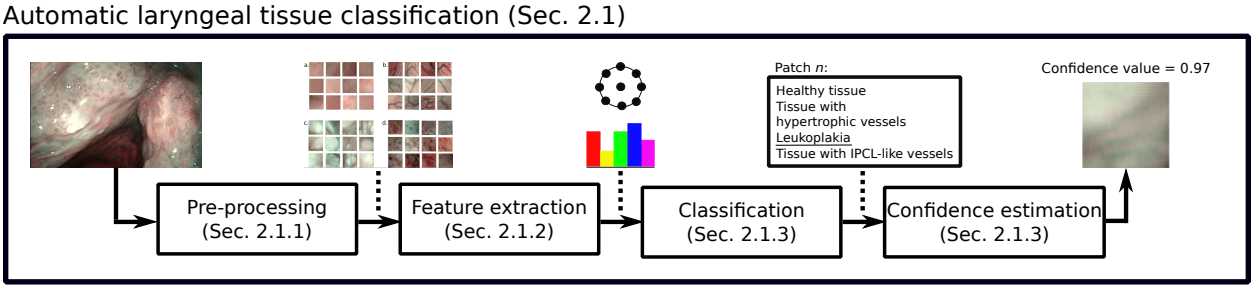


Fig 2: Workflow of the proposed approach to laryngeal tissue classification in narrow-band imaging endoscopic video frames.

confidence samples, the number of incorrectly classified cases can be lowered.

The importance of estimating the level of classification confidence with a view to improving system performance has been widely highlighted in several research fields, such as face recognition,[15] spam-filtering,[16] and glioma and colon cancer recognition.[17] In particular, is has been reported that allowing a system to produce "don't know" results can potentially reduce the number of incorrectly classified cases.[18] In the analyzed scenario, estimating the classification confidence would be beneficial since tissue biopsy would be required only for low-confidence regions in the image.

To the best of our knowledge, we are the first to investigate the use of texture-based classification algorithms for laryngeal tissue analysis.

2 Material and methods

This section explains the proposed approach to automatic laryngeal tissue classification (Sec. 2.1), as well as the evaluation protocol (Sec. 2.2) used to investigate the two hypotheses introduced in Sec. 1.

2.1 Automatic laryngeal tissue classification

The proposed method consists of the following steps: (i) Pre-processing (Sec. 2.1.1), (ii) Feature extraction (Sec. 2.1.2), (iii) Classification (Sec. 2.1.3) and (iv) Confidence estimation (Sec. 2.1.4). The workflow of the approach is shown in Fig. 2.

2.1.1 Pre-processing

Anisotropic diffusion filtering[19] is used to lower noise while preserving sharp edges in NBI images. Specular reflections (SR), usually present due to the wet and smooth laryngeal surface, are automatically identified exploiting their low saturation and high brightness and then masked.[6] After denoising, squared patches are selected from the image, as described in Sec. 2.2.

SR masking is necessary because it may not always be possible selecting patches without SR. This is due to the small extension of early-stage cancerous tissues in the image, which may overlap with SR (especially for the case of intra-papillary capillary loop-like vessels).

2.1.2 Feature extraction

As laryngeal endoscopic images are captured under various illumination conditions and from different viewpoints, the features that encode the tissue texture information should be robust to the pose of the endoscope as well as to the lighting conditions. Furthermore, with a view of a real-time computer-aided application, they should be computationally cheap. In this paper, we investigate the use of the following descriptors to characterize the texture of laryngeal tissues:

Texture-based global descriptors Among classic texture-based global descriptors, LBP are widely considered as the state of the art for medical image texture analysis.[20] LBP are gray-scale

invariant and provide low-complexity, well matching the requisite of this application. The first formulation of LBP ($LBP_{R,P}$) introduced in the literature requires to define, for a pixel $\mathbf{c} = (c_x, c_y)$, a spatial circular neighborhood of radius R with P equally-spaced neighbor points ($\{\mathbf{p}_n\}_{n\in(0,P-1)}$):

$$LBP_{R,P}(\mathbf{c}) = \sum_{n=0}^{P-1} s(g_{\mathbf{p}_n} - g_{\mathbf{c}})2^n \tag{1}$$

where $g_{\mathbf{c}}$ and $g_{\mathbf{p}_n}$ denote the gray values of the pixel $\mathbf{c}$ and of its n^{th} neighbor $\mathbf{p}_n$, respectively, and s is defined as:

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

The most often adopted LBP formulation is the uniform rotation-invariant one ($LBP^{riu2}_{R,P}$).[21] Rotation invariance is suitable for the purpose of this paper since the endoscope pose during the larynx inspection is constantly changing. From $LBP^{riu2}_{R,P}$, the L2-normalized histogram of $LBP^{riu2}_{R,P}$ ($H_{LBP_{riu2}}$) is computed and used as feature vector.

For comparison, the GLCM, a second widely used descriptor, is tested. GLCM calculates how often pair of pixels $(\mathbf{c}, \mathbf{q})$ with specific values and in a specified spatial relationship occur in an image. The spatial relationship is defined by θ and d, which are the angle and distance between $\mathbf{c}$ and $\mathbf{q}$. The GLCM width (W), equal to the GLCM height (H), corresponds to the number of quantized image intensity gray-levels. For the $w = h$ intensity gray-level, the GLCM computed

with θ and d is defined as:

$$GLCM_{\theta,d}(h,w) = \begin{cases} 1, & I(\mathbf{c}) = h \text{ and } I(c_x + d \cdot cos(\theta),\ c_y + d \cdot sin(\theta)) = w \\ 1, & I(\mathbf{c}) = h \text{ and } I(c_x - d \cdot cos(\theta),\ c_y - d \cdot sin(\theta)) = w \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

From the normalized $GLCM_{\theta,d}$, as suggested in Ref. 22, a feature set (F_{GLCM}) is extracted, which consists of GLCM *contrast*, *correlation*, *energy* and *homogeneity*. The normalized $GLCM_{\theta,d}$, which expresses the probability of gray-level occurrences, is obtained by dividing each $GLCM_{\theta,d}$ entry by the sum of all entries.

First order statistics Intensity *mean*, *variance* and *entropy* (Eq. 4) in each patch are computed and concatenated to form a single intensity-based feature set ($Stat_1$). The *entropy* is defined as:

$$entropy = -\sum_i h_i \log_2(h_i) \tag{4}$$

where h_i refers to the image histogram counts of the $i(= 0:255)$ bin. As recommended in Ref. 23, such features are adopted to integrate the texture-based informations encoded in $LBP^{riu2}_{R,P}$.

In addition to these descriptors, we tested two feature combinations (F_{GLCM}+$Stat_1$, $H_{LBP_{riu2}}$+ $Stat_1$), as suggested in Ref. 23 for applications in colorectal image analysis.

2.1.3 *Classification*

To perform tissue classification, SVM are used.[24] SVM are chosen since they allow overcoming the *curse-of-dimensionality* that arises analyzing our high-dimensional feature space.[25,26] The *kernel-trick* prevents parameter proliferation, lowering computational complexity and limiting over-fitting. Moreover, the SVM decisions are only determined by the support vectors, which makes SVM robust to noise in training data. Here, SVM with the Gaussian kernel (Ψ) are used. For a binary classification problem, given a training set of N data $\{y_k, \mathbf{x_k}\}_{k=1}^{N}$, where $\mathbf{x_k}$ is the k^{th} input feature vector and y_k is the k^{th} output label, the SVM decision function takes the form of:

$$f(\mathbf{x}) = sign\Big[\sum_{k=1}^{N} a_k^* y_k \Psi(\mathbf{x}, \mathbf{x_k}) + b\Big] \tag{5}$$

where:

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

b is a real constant and a_k^* is retrieved as follow:

$$a_k^* = \max\Big\{-\frac{1}{2}\sum_{k,l=1}^{N} y_k y_l \Psi(\mathbf{x_k}, \mathbf{x_l}) a_k a_l + \sum_{k=1}^{N} a_k\Big\} \tag{7}$$

with:

$$\sum_{k=1}^{N} a_k y_k = 0, \qquad 0 \leq a_k \leq C, \qquad k = 1, ..., N \tag{8}$$

In this paper, γ and C are retrieved with grid search, as explained in Sec. 2.2. To implement multi-class SVM classification, the *one-vs-one* scheme is used.

For the sake of completeness, the performance of other classifiers, such us k-nearest neighbors

(kNN),[27] naive Bayes (NB),[28] and random forest (RF),[29] are also investigated.

Prior to classification, the feature matrices are normalized within each feature dimension. Specifically, the feature matrices are pre-processed by removing the mean (centering) and scaling to unit variance.

2.1.4 Confidence estimation

As a pre-requisite for our confidence estimation, we compute the probability ($Pr_i(j)$) of the i^{th} patch to belong to the jth class, with j $\in$ $[1, J]$ and J the number of considered tissue classes. For the probability computation, the Platt scaling method revised for multi-class classification problems is used.[30] The Platt scaling method consists of training the parameters of an additional sigmoid function to map SVM outputs to probabilities.

To estimate the reliability of the SVM classification of the i^{th} patch, inspired by the work in Ref. 31 for abdominal tissue classification applications, we evaluate the dispersion of Pr_i among the J classes using the Gini coefficient (GC):[32]

$$\text{GC} = 1 - 2\int_0^1 L(x)dx \tag{9}$$

where L is the Lorentz curve, which is the cumulative probability among laryngeal classes rank-ordered by decreasing values of their individual probabilities. The GC has value 0 if all the probabilities are equally distributed (maximum uncertainty) and 1 for maximum inequality (the classifier is 100% confident in assigning the label). The classification of a patch is considered to be confident

Table 1: Evaluation dataset. For each of the 33 patients' video, 10 images are used for a total of 330 images. From each image, 4 tissue patches are extracted for a total of 1320 patches relative to the 4 considered tissue classes: healthy tissue, tissue with hypertrophic vessels, leukoplakia, tissue with intraepithelial papillary capillary loop-like vessels. For a robust evaluation, the dataset is split at patient level to perform 3-fold cross-validation. In each fold, 11 patients are included, for a total of 110 images per fold. Each fold contains 440 patches equally distributed among the laryngeal tissue classes.

	Fold 1	Fold 2	Fold 3	Total
patient ID	1-11	12-22	23-33	33
n. of images	110 (10 per patient)	110 (10 per patient)	110 (10 per patient)	330
n. of patches	440 patches (4 per image)	440 patches (4 per images)	440 patches (4 per image)	1320

if GC is higher that a threshold (τ):

$$\begin{cases} \text{Patch}(i) \text{ is confident}, & \text{GC} \geq \tau; \\ \text{Patch}(i) \text{ is not confident}, & \text{otherwise.} \end{cases}$$

2.2 Evaluation

In this study, four tissue classes, which are typically evaluated during early-stage diagnosis with NBI laryngoscopy, are considered: (i) tissue with IPCL-like vessels, (ii) leukoplakia, (iii) tissue with hypertrophic vessels, and (iv) healthy tissue. We retrospectively analyzed 33 NBI videos, which refer to 33 different patients affected by SCC. SCC was diagnosed with histopathological examination. Videos were acquired with a NBI endoscopic system (Olympus Visera Elite S190 video processor and an ENF-VH rhino-laryngo videoscope) with frame rate of $25 fps$ and image size of 1920×1072 pixels.

A total number of 330 in-focus images (10 per video) was manually selected from the videos, in such a way that the distance between the endoscope and the tissue could be considered constant and approximatively equal to 1 mm for all the images. This distance is suggested in clinics for correct evaluation of tissues during NBI endoscopy examination.[33]

The NBI images were pre-processed as in Sec. 2.1.1. The parameters used for anisotropic

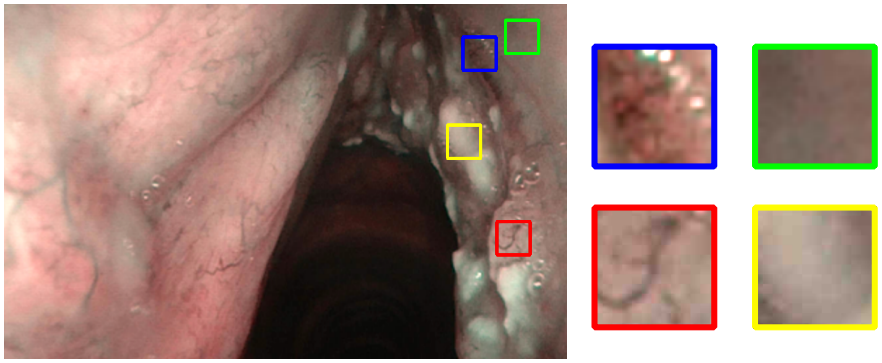


Fig 3: Four patches, relative to the four analyzed laryngeal tissue classes, are manually cropped from the image. Blue: tissue with intraepithelial papillary capillary loop-like vessels; Yellow: tissue with Leukoplakia; Green: healthy tissue; Red: tissue with hypertrophic vessels.

Table 2: Tested feature vectors and corresponding number of features. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Normalized histogram of rotation-invariant uniform local binary patterns.

Feature vector	$Stat_1$	F_{GLCM}	F_{GLCM} + $Stat_1$	$H_{LBP_{riu2}}$	$H_{LBP_{riu2}}$ + $Stat_1$
Number of features	9	144	153	162	171

diffusion filtering were set as in Ref. 34. The saturation and brightness thresholding values used to mask SR were set as in Ref. 6.

For each of the 330 images, 4 patches were manually cropped with a size of 100×100 pixels, for a total of 1320 patches, equally distributed among the four classes (Table 1). Each patch was cropped from a portion of tissue relative to only one of the four considered classes (tissue with IPCL-like vessels, leukoplakia, hypertrophic vessels and healthy tissue), thus avoiding tissue overlap in one patch. The selection was performed under the supervision of an expert clinician (otolaryngologist specialized in head and neck oncology). A visual example of 4 patches cropped from a NBI frame is shown in Fig. 3. We decided to select only one patch per tissue class because, in most of the images, we were not able to select more than a single patch for the IPCL-like class. This is due to the small extension of this vascular alteration in early-stage cancer.

For the feature extraction described in Sec. 2.1.2, the $LBP^{riu2}_{R,P}$ were computed with the following $(R; P)$ combinations: $(1; 8)$, $(2; 16)$, $(3; 24)$, and the corresponding $H_{LBP_{riu2}}$ were concate-

nated. Such choice allows a multi-scale, and therefore a more accurate description of the texture, as suggested in Ref. 10. Twelve $GLCM_{\theta,d}$ were computed using all the possible combinations of (θ, d), with $\theta \in \{0°, 45°, 90°, 135°\}$ and $d \in \{1, 2, 3\}$, and the corresponding F_{GLCM} sets were concatenated. The chosen interval of θ allows to approximate rotation invariance, as suggested in Ref. 22. The values of d were chosen to be consistent with the scale used to compute $LBP^{riu2}_{R,P}$. $LBP^{riu2}_{R,P}$, $GLCM_{\theta,d}$ and $Stat_1$ were computed for each channel in the NBI image. All the tested feature vectors and their length are reported in Table 2.

As for performing the classification presented in Sec. 2.1.3, the SVM hyper-parameters (γ, C) were retrieved via grid-search and cross-validation on the training set. The grid-search space for γ and C was set to [10^{-7}, 10^{-1}] and [10^{-3}, 10^{3}], respectively, with six values spaced evenly on log_{10} scale in both cases. Similarly, we retrieved the number of neighbors for kNN with a grid-search space set to [2,10] with nine values spaced evenly, and the number of trees in the forest for RF with a grid-search space set to [40,100] with six values spaced evenly.

The computation of $H_{LBP_{riu2}}$, F_{GLCM} and $Stat_1$ was implemented using OpenCv [1]. The classification was implemented with scikit-learn [2].

Investigation of H1 In order to assess our hypothesis that machine-learning techniques can characterize laryngeal tissues in NBI images by exploiting textural information, we first evaluated the classification performance of the texture descriptors without confidence estimation (*Base* case).

To obtain a robust estimation of the classification performance, 3-fold cross-validation was performed, separating data at patient level to prevent data leakage. The 1320-patch dataset was split to obtain well-balanced folds both at patient-level and tissue-level, as shown in Table 1. Each time,

[1] http://docs.opencv.org /3.1.0/index.html
[2] http://scikit-learn.org/stable/ index.html

two folds were used for training and the remaining one for testing purpose only. This evaluation does not lead to biased results since our dataset is balanced over the three folds.

Inspired by Ref. 10, we computed the class-specific recall ($\mathbf{Rec_{class}} = \{Rec_{class_j}\}_{j\in[1,J=4]}$) to evaluate the classification performance, where:

$$Rec_{class_j} = \frac{TP_j}{TP_j + FN_j} \tag{10}$$

being TP_j the number of elements of the j^{th} class correctly classified (true positive of the j^{th} class) and FN_j the number of elements of the j^{th} class wrongly assigned to one of the three left classes (false negative of the j^{th} class). We further evaluated the class-specific precision ($\mathbf{Prec_{class}} = \{Prec_{class_j}\}_{j\in[1,J=4]}$), where:

$$Prec_{class_j} = \frac{TP_j}{TP_j + FP_j} \tag{11}$$

being FP_j the number of false positive of the j^{th} class, and the F1 score ($\mathbf{F1_{class}}\{F1_{class_j}\}_{j\in[1,J=4]}$), where:

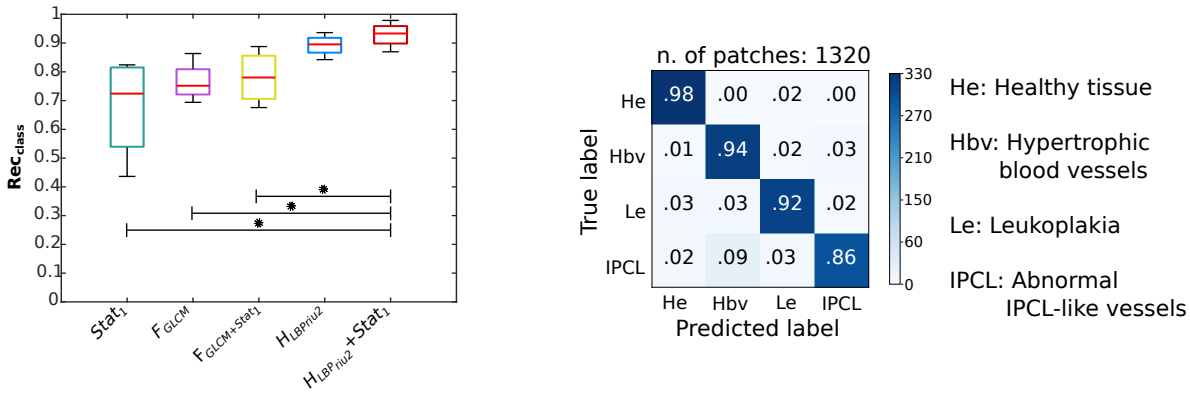
$$F1_{class_j} = 2\frac{Prec_{class_j} \times Rec_{class_j}}{Prec_{class_j} + Rec_{class_j}} \tag{12}$$

For a comprehensive analysis, we computed the area (AUC) under the receiver operating characteristic (ROC) curve. Since our task is a multi-class classification problem and our dataset is balanced, we computed the macro-average ROC curve. The gold-standard classification was obtained by labeling the patches under the supervision of an expert clinician.

We used the Wilcoxon signed-rank test (significance level $\alpha = 0.05$) for paired sample to assess whether the classification achieved with our best performing (highest $\mathbf{Rec_{class}}$ median value) feature vector significantly differs from the ones achieved with the other feature sets in Table 2.

Table 3: Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec_{class}}$), precision ($\mathbf{Prec_{class}}$), and F1-score ($\mathbf{F1_{class}}$) obtained testing different feature vectors for the *Base* case (i.e., without the inclusion of confidence on classification estimation). Classification is obtained with support vector machines. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Normalized histogram of rotation-invariant uniform local binary patterns.

	$Stat_1$	F_{GLCM}	F_{GLCM} + $Stat_1$	$H_{LBP_{riu2}}$	$H_{LBP_{riu2}}$ + $Stat_1$
$\mathbf{Rec_{class}}$	72 (54-82)	75 (72-81)	78 (71-86)	90 (87-92)	93 (90-96)
$\mathbf{Prec_{class}}$	67 (57-80)	75 (71-80)	78 (72- 84)	90 (88-92)	94 (91-95)
$\mathbf{F1_{class}}$	70 (56-81)	74 (71-80)	79 (72-85)	90 (89-91)	92 (91-95)

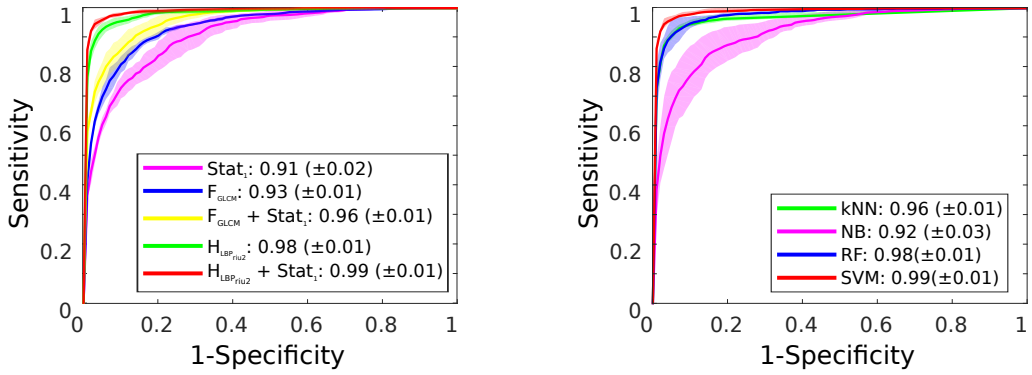


(a) Boxplots of $\mathbf{Rec_{class}}$ for the tested features (b) Confusion matrix for $H_{LBP_{riu2}}$

Fig 4: Comparison of different features without including the classification confidence estimation. Classification is obtained with support vector machines. (a) Boxplots of class-specific recall ($\mathbf{Rec_{class}}$) for different features. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Histogram of rotation-invariant uniform local binary patterns. The stars indicate significant differences (Wilcoxon test, $\alpha = 0.05$). (b) Normalized confusion matrix for $H_{LBP_{riu2}} + Stat_1$. The colorbar indicates the number of patches. The total number of patches (n. of patches) is reported.

Similarly, we evaluated whether the classification achieved with SVM differs (Wilcoxon signed-rank test with $\alpha = 0.05$) from the ones achieved with the other tested classifiers (kNN, NB, RF).

For the sake of completeness, we compared the performance of our best-performing feature set with those of the most recent - and so far the only one - method (Ref. 6) published on the topic of laryngeal tissue classification in NBI endoscopy, applying the latter to our dataset. As introduced in Sec. 1, the method requires to set the vessel segmentation parameters, which were here set as in Ref. 6. The feature classification was performed with SVM, instead of LDA, for fair comparison. The comparison was repeated excluding the leukoplakia class, to avoid privileging the proposed



(a) ROC curves for the tested features. (b) ROC curves for the the tested classifiers.

Fig 5: Macro-average receiver operating characteristic (ROC) curves. The mean (± standard deviation) curves obtained from the 3 cross-validation folds are reported in bold (transparent area). The mean (± standard deviation) area under the ROC curve is reported in the legend. (a) ROC curves for the tested features. Classification is obtained using support vector machines. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Histogram of rotation-invariant uniform local binary patterns. (b) ROC curves for the tested classifiers. Classification is obtained using the histogram of local binary pattern and first order statistics. kNN: k-nearest neighbors, NB: naive Bayes, RF: random forest, SVM: support vector machines.

method. Indeed, the method in Ref. 6 focuses on the analysis of vessels, which however are not visible in case of leukoplakia due to the thickening of the epithelial layer.

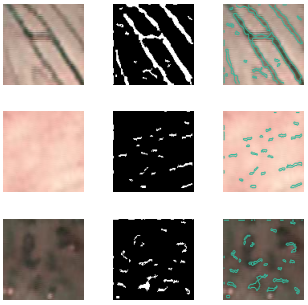
Investigation of H2 To investigate the hypothesis that, by estimating the level of classification confidence and discarding low-confidence samples, the number of incorrectly classified cases can be lowered, we evaluated how $\mathbf{Rec_{class}}$, $\mathbf{Prec_{class}}$, $\mathbf{F1_{class}}$ obtained with our best performing feature vector change considering different thresholds ($\tau \in [0.6 : 0.1 : 1)$) on the GC value. Since, once the low-confidence patches are excluded, the balance between classed did not hold, we computed the ROC curves for each of the four laryngeal classes (and not the macro-average ones as for **H1**).

3 Results

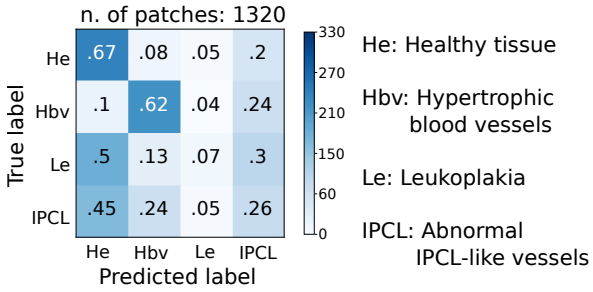
For the *Base* case, the best performance (median $\mathbf{Rec_{class}}$ = 93%, inter-quartile range (IQR) = 6%) was obtained with $H_{LBP_{riu2}} + Stat_1$ and SVM classification, as shown in Table 3. The same

Table 4: Comparison of different classifiers. Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec}_{\mathbf{class}}$), precision ($\mathbf{Prec}_{\mathbf{class}}$), and F1 score ($\mathbf{F1}_{\mathbf{class}}$) are reported for the four different tissue classes. Classification is obtained using the histogram of local binary patterns and first order statistics. kNN: k-nearest neighbors, NB: naive Bayes, RF: random forest, SVM: support vector machines.

	kNN	NB	RF	SVM
$\mathbf{Rec}_{\mathbf{class}}$	90 (84-93)	78 (74-82)	89 (84-91)	93 (90-96)
$\mathbf{Prec}_{\mathbf{class}}$	89 (86-91)	81 (73-84)	87 (86-89)	94 (91-95)
$\mathbf{F1}_{\mathbf{class}}$	89 (86-91)	79 (74-83)	89 (86-90)	92 (91-95)



(a) Vessel segmentation



(b) Confusion matrix for the method in Ref. 6

Fig 6: Performance of the state of art. (a) Visual samples of the vessel segmentation obtained applying Barbalata et al.[6] algorithm to patches with hypertrophic vessels (first row), healthy tissue (second row) and intraepithelial papillary capillary loop-like vessels (third row). From left to right, original patch, vessel mask and vessel mask superimposed on the original patch. (b) Normalized confusion matrix obtained applying Barbalata et al.[6] algorithm to our dataset. Colorbar indicates the number of patches. The total number of patches (n. of patches) is reported.

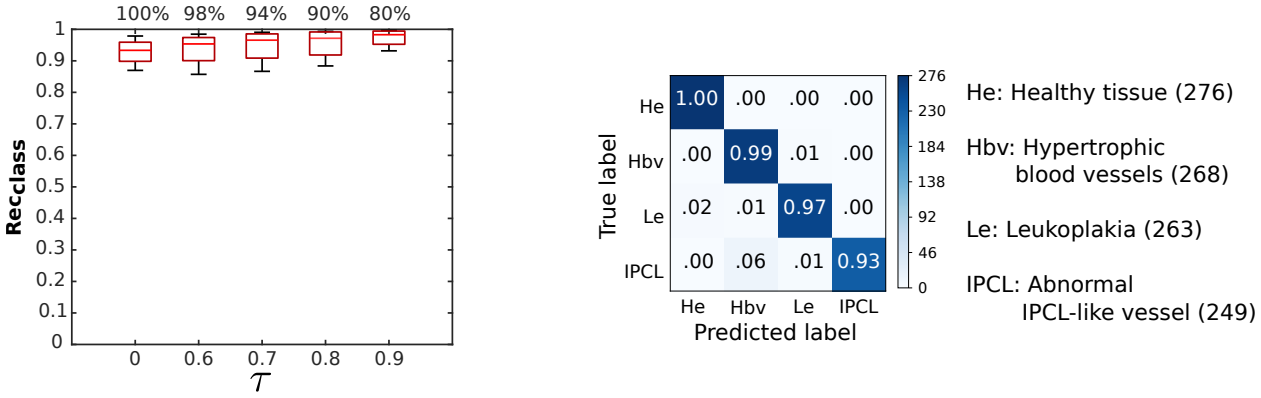
was observed also when considering $\mathbf{Prec}_{\mathbf{class}}$ (median = 94%, IQR = 4%) and $\mathbf{F1}_{\mathbf{class}}$ (median = 92%, IQR = 4%). The classification statistics relative to all the analyzed features are reported in Fig. 4a. Significant differences (p-value < 0.05) were found when comparing $H_{LBP_{riu2}} + Stat_1$ with $Stat_1$, F_{GLCM}, and F_{GLCM} + $Stat_1$. The normalized confusion matrix for $H_{LBP_{riu2}} + Stat_1$ is shown in Fig. 4b. In Fig. 5a, the macro-average ROC curves are reported for all tested features and SVM classification. The mean AUC across the three folds was 0.99 for $H_{LBP_{riu2}} + Stat_1$.

As shown in Table 4, SVM has shown comparable performance with respect to kNN and RF in terms of $\mathbf{Rec}_{\mathbf{class}}$, $\mathbf{Prec}_{\mathbf{class}}$, $\mathbf{F1}_{\mathbf{class}}$, while SVM outperformed (p-value < 0.05) NB. The same can be noticed from the ROC curve analysis in Fig. 5b.

When applying the algorithm proposed by Barbalata et al.[6] to our dataset, a median $\mathbf{Rec}_{\mathbf{class}}$

Table 5: Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec_{class}}$), precision ($\mathbf{Prec_{class}}$), and F1 score ($\mathbf{F1_{class}}$) are reported at different level of confidence (τ) on support vector machines classification.

	$\tau = 0$	$\tau = 0.60$	$\tau = 0.70$	$\tau = 0.80$	$\tau = 0.90$
$\mathbf{Rec_{class}}$	93 (90-96)	95 (91-97)	96 (92-99)	98 (93-99)	98 (95-100)
$\mathbf{Prec_{class}}$	94 (91-95)	95 (92-96)	97 (93-98)	97 (95-98)	99 (96-100)
$\mathbf{F1_{class}}$	92 (91-95)	94 (93-96)	95 (94-97)	96 (95-97)	98 (97-99)

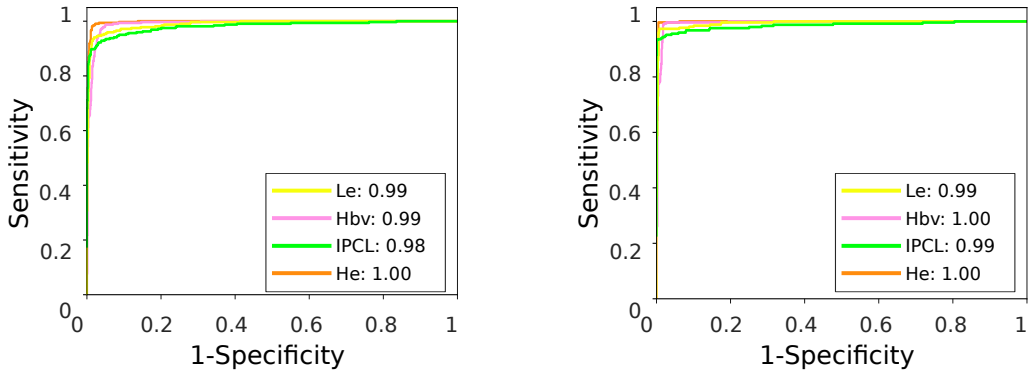


(a) $\mathbf{Rec_{class}}$ for different values of τ (b) Confusion matrix for $\tau = 0.9$

Fig 7: Effect of varying the threshold (τ) on the classification confidence level. Classification is obtained using local binary pattern and first order statistics with support vector machines. (a) Boxplot of the class-specific accuracy-rate ($\mathbf{Rec_{class}}$) for different τ. The percentage of confident patches for each τ is reported above each boxplot. $\tau = 0$ refers to classification without confidence estimation. (b) Normalized confusion matrix for $\tau = 0.9$. The number of patches for each class is reported in parenthesis.

value of 42% was obtained, with IQR of 48%. Significant differences (p-value $\ll 0.05$) were found when comparing the algorithm results with those obtained exploiting $H_{LBP_{riu2}} + Stat_1$. Visual examples of the vessel segmentation obtained with the method proposed in Ref. 6 are reported in Fig. 6a for patches with hypertrophic vessels, healthy tissue and IPCL-like vessels. The confusion matrix for the classification obtained with the method in Ref. 6 is reported in Fig. 6b. Barbalata et al. algorithm correctly labeled leukoplakias and abnormal IPCL only in the 7% and 26% of all cases, respectively. Almost half of leukoplakias and abnormal IPCL were misclassified as healthy tissues. When excluding the leukoplakia class, the $\mathbf{Rec_{class}}$ was: 62% (healthy tissue), 70% (tissue with hypertrophic vessels), 28% (tissue with ICPL-like vessels).

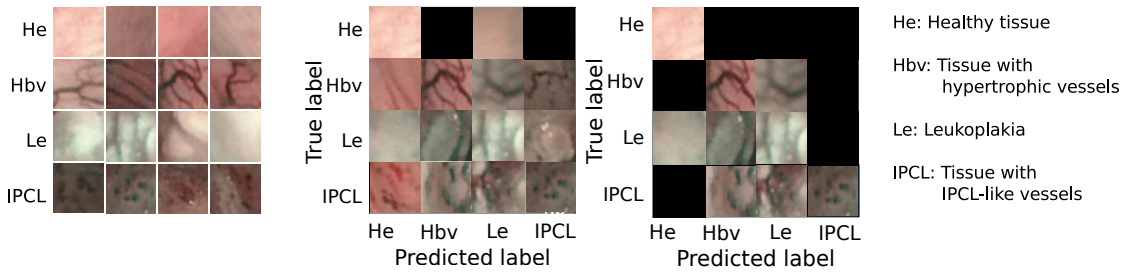
As shown in Table 5, when varying τ in $[0.6 : 0.1 : 1)$, the median $\mathbf{Rec_{class}}$ for $H_{LBP_{riu2}} + Stat_1$



(a) ROC curves for $\tau = 0.6$. (b) ROC curves for $\tau = 0.9$.

Fig 8: Receiver operating characteristic (ROC) curves at different level τ of confidence on classification. Each curve refers to one of the laryngeal tissue classes. He: healthy tissue; Hbv: tissue with hypertrophic vessels; Le: leukoplakia; IPCL: tissue with intraepithelial papillary capillary loop-like vessels. The area under the ROC curve, for each curve, is reported in the legend. Classification is obtained using local binary pattern and first order statistics with support vector machines. (a) ROC curves for $\tau = 0.6$. (b) ROC curves for for $\tau = 0.9$.

monotonically increased from 93% (*Base* case) to 98% (τ = 0.9). The corresponding statistics are shown in Fig. 7a. In particular, as can be seen by comparing the confusion matrices in Fig. 4b and Fig. 7b, the classification recall increased from 98% (healthy tissue), 94% (hypertrophic vessels), 92% (leukoplakia), 86% (IPCL-like vessels) to 100% (healthy tissue), 99% (hypertrophic vessels), 97% (leukoplakia), 93% (IPCL-like vessels). Despite the fact that a slightly lower improvement was observed for the IPCL class with respect to the other classes, it is worth noting that, with the proposed approach, no misclassification occurred between the IPCL class and the healthy tissue one. The same happened also for the hypertrophic vessel class, while only the 2% of samples with leukoplakia was misclassified as healthy. The same trend was observed for $\mathbf{Prec}_{\mathbf{class}}$ and $\mathbf{F1}_{\mathbf{class}}$. The ROC curves for $\tau = 0.6$ and $\tau = 0.9$ are shown in Fig. 8. The AUC is reported for each of the analyzed classes (AUC = 0.99 ± 0.01 ($\tau = 0.9$)). The $\mathbf{Rec}_{\mathbf{class}}$ increment came at the cost of a reduction of the percentage of confident patches to 80% ($\tau = 0.9$) of all the patches in the testing set, which corresponds to $\sim$ 1056 patches. However, as shown in Fig. 6b, with the exclusion of low-confidence patches, even in the worst case (classification of tissue with IPCL-like vessels), the



(a) (b) (c)

Fig 9: Visual samples of classification results. Classification is obtained using local binary pattern and first order statistics with support vector machines. (a) Examples of patches for the four tissue classes in our dataset. Visual confusion matrices for the *Base* case (b), i.e., without the inclusion of confidence estimation, and after including the confidence estimation with $\tau = 0.9$ (c). Black squares indicate the absence of misclassification between the true and predicted label.

accuracy still reached 93%.

Fig. 9 shows visual samples of patches in our dataset (Fig. 9a), as well as samples of patch classification results at the *Base* case (Fig. 9b) and after the introduction of the confidence measure ($\tau = 0.9$) (Fig. 9c).

4 Discussion

In this paper, we presented and fully evaluated an innovative approach to the computer-aided classification of laryngeal tissues in NBI laryngoscopy. Different textural features were tested to investigate the best feature set to characterize malignant and healthy laryngeal tissues: texture-based global descriptors (F_{GLCM} and $H_{LBP_{riu2}}$) and first order statistics ($Stat_1$). A confidence measure on the SVM-based classification was used to estimate the reliability of the classification results.

When comparing non-combined features (F_{GLCM}, $Stat_1$, $H_{LBP_{riu2}}$), the highest classification performance was obtained with $H_{LBP_{riu2}}$. In general, F_{GLCM} performed worse with respect to $H_{LBP_{riu2}}$. This is probably due to the GLCM lack of robustness to illumination condition changes, which are typically encountered during endoscopic examination.

SVM has shown comparable performance with respect to RF and kNN, while significant differ-

ences (p-value < 0.05) where found with respect to NB. This is probably due to NB not being able to handle high-dimensional feature spaces such as ours. This is in accord with previous findings in the literature, e.g., Refs. 25, 35, 36. Accordingly, for the tested dataset, we could not conclude that SVM performance was better than the one obtained with RF and kNN.

When comparing the proposed method with the state of the art, the classification based on $H_{LBP_{riu2}}$ significantly outperformed (p-value $\ll 0.05$) the one proposed by Barbalata et al.,[6] also when excluding the leukoplakia class. Since the method in Ref. 6 relies on accurate vessel segmentation (to extract vascular shape-based features), a possible reason of such result could be related to the challenging nature of our validation dataset, which however well summarizes the diagnostic scenario. Indeed, vessel segmentation was not trivial (Fig. 6a) due to (i) the noisy nature of NBI data, (ii) the low contrast of vessels in patches with healthy tissue and leukoplakia and (iii) the irregular shape of IPCL-like vessels. With texture-based features, higher classification performance was achieved with respect to shape-based features since texture-based feature computation does not require vessel segmentation. Moreover, the texture-based features here used are invariant to illumination changes and endoscope pose, which makes them suitable for the analyzed scenario.

The classification performance obtained with $H_{LBP_{riu2}}$ was further increased by estimating the confidence of the SVM classification, with few misclassification of confident patches that mainly occurred with high-challenging vascular patterns, whose classification is not trivial also for the human eye (Fig. 9c). Such results support our hypothesis that the proposed approach is suitable for classifying laryngeal tissues with high reliability, since it automatically estimates its own confidence level and provides high classification accuracy for confident patches.

A limitation of the proposed study could be seen in its patch-based nature. Note, however, that the choice of focusing on patches manually extracted under the supervision of an expert clini-

cian was driven by the necessity of having a controlled and representative dataset to fairly evaluate different features. As future work, instead of manually selecting squared patches, we plan to implement more automatic strategies, such as superpixel segmentation.[37] The features cold be directly extracted from superpixels, as to classify each superpixel as belonging to one of the analyzed laryngeal classes. Moreover, considering that recent researches on gastrointestinal image classification (e.g., Refs. 38, 39) are focusing more and more on convolutional neural networks (CNN), it would be interesting to exploit also CNN as feature extractor for comparison.

Our expectation is that research on the classification of laryngeal tissues will be empowered by the proposed work, becoming a topic of interest for the scientific community, which until now has mainly focused on other anatomical sites, such as the gastro-intestinal tract. Moreover, we hope this study will motivate a more structured and widespread data collection in clinics and the sharing of such data through public databases. Despite the dimension of the analyzed dataset (330 images) is comparable with that of similar researches (e.g., Barbalata et al.[6] with 120 images, Turkmen et al.[7] with 70 images), larger amounts of data would bring the possibility of further exploring machine-learning classification algorithms, e.g., to classify a larger number of laryngeal malignant tissues.

In conclusion, the most significant contribution of this work is showing that LBP-based features and SVM can differentiate laryngeal tissues accurately. This is highly beneficial for practical uses. Comparing with other state-of-the-art method in the area, the proposed method is simpler and the result is more accurate. It is acknowledged that further research is required to further ameliorate the algorithm as to offer all possible support for diagnosis, but the results presented here are surely a promising step towards a helpful endoscope-integrated processing system to support the diagnosis of early-stage SCC.

Disclosures

The authors have no conflict of interest to disclose.

References

1 K. Markou, A. Christoforidou, I. Karasmanis, G. Tsiropoulos, S. Triaridis, I. Constantinidis, V. Vital, and A. Nikolaou, "Laryngeal cancer: epidemiological data from Northern Greece and review of the literature," *Hippokratia* **17**(4), pp. 313–318, 2013.

2 J. Unger, J. Lohscheller, M. Reiter, K. Eder, C. S. Betz, and M. Schuster, "A noninvasive procedure for early-stage discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis," *Cancer research* **75**(1), pp. 31–39, 2015.

3 C. Piazza, F. Del Bon, G. Peretti, and P. Nicolai, "Narrow band imaging in endoscopic evaluation of the larynx," *Current Opinion in Otolaryngology & Head and Neck Surgery* **20**(6), pp. 472–476, 2012.

4 P. J. Poels, F. I. de Jong, and H. K. Schutte, "Consistency of the preoperative and intraoperative diagnosis of benign vocal fold lesions," *Journal of Voice* **17**(3), pp. 425–433, 2003.

5 J. S. Isenberg, D. L. Crozier, and S. H. Dailey, "Institutional and comprehensive review of laryngeal leukoplakia," *Annals of Otology, Rhinology & Laryngology* **117**(1), pp. 74–79, 2008.

6 C. Barbalata and L. S. Mattos, "Laryngeal tumor detection and classification in endoscopic video," *IEEE Journal of Biomedical and Health Informatics* **20**(1), pp. 322–332, 2016.

7 H. I. Turkmen, M. E. Karsligil, and I. Kocak, "Classification of laryngeal disorders based on shape and vascular defects of vocal folds," *Computers in Biology and Medicine* **62**, pp. 76–85, 2015.

8 L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, *et al.*, "Surgical Data Science: Enabling Next-Generation Surgery," *arXiv preprint, arXiv:1701.06482* , 2017.

9 B. Li and M. Q.-H. Meng, "Texture analysis for ulcer detection in capsule endoscopy images," *Image and Vision computing* **27**(9), pp. 1336–1342, 2009.

10 Y. Zhang, S. J. Wirkert, J. Iszatt, H. Kenngott, M. Wagner, B. Mayer, C. Stock, N. T. Clancy, D. S. Elson, and L. Maier-Hein, "Tissue classification for laparoscopic image understanding based on multispectral texture analysis," in *SPIE Medical Imaging*, pp. 978619–978619, International Society for Optics and Photonics, 2016.

11 P. Liang, Y. Cong, and M. Guan, "A computer-aided lesion diagnose method based on gastroscopeimage," in *Information and Automation (ICIA), 2012 International Conference on*, pp. 871–875, IEEE, 2012.

12 X. Shen, K. Sun, S. Zhang, and S. Cheng, "Lesion detection of electronic gastroscope images based on multiscale texture feature," in *Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on*, pp. 756–759, IEEE, 2012.

13 F. Van Der Sommen, S. Zinger, E. J. Schoon, *et al.*, "Computer-aided detection of early cancer in the esophagus using HD endoscopy images," in *SPIE Medical Imaging*, pp. 86700V–86700V, International Society for Optics and Photonics, 2013.

14 M. Misawa, S.-e. Kudo, Y. Mori, K. Takeda, Y. Maeda, S. Kataoka, H. Nakamura, T. Kudo, K. Wakamura, T. Hayashi, *et al.*, "Accuracy of computer-aided diagnosis based on narrow-band imaging endocytoscopy for diagnosing colorectal lesions: comparison with experts," *International Journal of Computer Assisted Radiology and Surgery* , pp. 1–10, 2017.

15 J. Orozco, O. Rudovic, F. X. Roca, and J. Gonzalez, "Confidence assessment on eyelid and eyebrow expression recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pp. 1–8, IEEE, 2008.

16 S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating estimates of classification confidence for a case-based spam filter," in *International Conference on Case-Based Reasoning (ICCBR)*, **3620**, pp. 177–190, Springer, 2005.

17 C. Zhang and R. L. Kodell, "Subpopulation-specific confidence designation for more informative biomedical classification," *Artificial Intelligence in Medicine* **58**(3), pp. 155–163, 2013.

18 B. McLaren and K. Ashley, "Helping a CBR program know what it knows," *Case-Based Reasoning Research and Development* , pp. 377–391, 2001.

19 A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, and B. van Ginneken, "Noise reduction in computed tomography scans using 3-D anisotropic hybrid diffusion with continuous switch," *IEEE Transactions on Medical Imaging* **28**(10), pp. 1585–1594, 2009.

20 L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis," *Artificial Intelligence in Medicine* **49**(2), pp. 117–125, 2010.

21 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* **24**(7), pp. 971–987, 2002.

22 R. M. Haralick, K. Shanmugam, *et al.*, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics* **3**(6), pp. 610–621, 1973.

23 D. Onder, S. Sarioglu, and B. Karacali, "Automated labelling of cancer textures in colorectal histopathology slides using quasi-supervised learning," *Micron* **47**, pp. 33–42, 2013.

24 C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**(2), pp. 121–167, 1998.

25 G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, **1**(1-22), pp. 1–2, Prague, 2004.

26 Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: fast feature extraction and SVM training," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1689–1696, IEEE, 2011.

27 J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics* (4), pp. 580–585, 1985.

28 D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *European Conference on Machine Learning*, pp. 4–15, Springer, 1998.

29 L. Breiman, "Random forests," *Machine learning* **45**(1), pp. 5–32, 2001.

30 T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research* **5**(Aug), pp. 975–1005, 2004.

31 S. Moccia, S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L. S. Mattos, and L. Maier-Hein, "Uncertainty-Aware Organ Classification for Surgical Data Science Applications in Laparoscopy," *arXiv preprint arXiv:1706.07002* , 2017.

32 B. G. Marcot, "Metrics for evaluating performance and uncertainty of Bayesian network models," *Ecological Modelling* **230**, pp. 50–62, 2012.

33 P. Lukes, M. Zabrodsky, E. Lukesova, M. Chovanec, J. Astl, J. A. Betka, and J. Plzak, "The role of NBI HDTV magnifying endoscopy in the prehistologic diagnosis of laryngeal papillomatosis and spinocellular cancer," *BioMed Research International* **2014**, 2014.

34 S. Moccia, V. Penza, G. O. Vanone, E. De Momi, and L. S. Mattos, "Automatic workflow for narrow-band laryngeal video stitching," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pp. 1188–1191, IEEE, 2016.

35 D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sensing of Environment* **118**, pp. 259–272, 2012.

36 A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

37 Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1356–1363, IEEE, 2015.

38 R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features," in *Image and Signal Processing (CISP), 2015 8th International Congress on*, pp. 372–376, IEEE, 2015.

39 S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural net-

works," in *SPIE Medical Imaging*, pp. 978528–978528, International Society for Optics and Photonics, 2016.

Sara Moccia (BSc 2012, MSc 2014) is a PhD student at the Department of Advanced Robotics at the Istituto Italiano di Tecnologia (IIT) and at the Department of Information, Electronics and Bioengineering at Politecnico di Milano. Her research interests include computer vision, image processing and machine learning applied to the medical field.

Elena De Momi (MSc 2002, PhD 2006) is Assistant Professor at the Electronic, Information, and Bioengineering Department (DEIB) of Politecnico di Milano. She was co-founder of the Neuro-engineering and Medical Robotics Laboratory, in 2008, being responsible of the Medical Robotics section. She has been Associate Editor of the Journal of Medical Robotics Research, of the International Journal of Advanced Robotic Systems and of Frontiers in Robotics and AI. Since 2016 she has been an Associated Editor of the IEEE International Conference on Robotics and Automation.

Marco Guarnaschelli (BSc 2014, MSc 2017) graduated from Politecnico di Milano in Biomedical Engineering. His research interests include computer vision, image processing and machine learning.

Matteo Savazzi (BSc 2014, MSc 2017) graduated from Politecnico di Milano in Biomedical Engineering. His research interests include computer vision, image processing and machine learning.

Dr. Andrea Laborai (MSc 2014) has been working as a Resident in Otorhinolaryngology at the Ospedale Policlinico San Martino University of Genova since 2014. He spent six months of the

residency program at the Spedali Civili of Brescia.

Prof. Luca Guastini has been working since 2012 both as a clinician of Genoa's IRCCS San Martino-IST University Hospital and as an Assistant Professor of Otorhinolaryngology at University of Genoa's Medical and Postgraduate Schools. So far, he has published 38 papers for a selection of prestigious national and international journals, 12 book chapters, and as many as 110 congress and symposium-related publications. He is member of the European Laryngological Society and Italian Otorhinolaryngology Society.

Prof. Giorgio Peretti (MSc 1983) is Full Professor at Genoa University and the Director of the Department of Otorhinolaryngology of the University of Genoa, Italy. So far, he has authored 92 papers for a selection of national and international journals, 45 book chapters, and as many as 355 congress and symposium-related publications. His main fields of research include angiogenesis in early-stage laryngeal carcinoma, narrow-band imaging in diagnosis and surveillance of head and neck cancer, combination of endoscopy and imaging in staging and surveillance of head and neck cancer.

Leonardo S. Mattos (BSc 1998, MSc 2003, PhD 2007) is a Team Leader at the Istituto Italiano di Tecnologia (IIT). His research interests include robotic microsurgery, user interfaces and systems for safe and efficient teleoperation, computer vision and automation. He has been a researcher at IIT's Advanced Robotics Department since 2007. He was the PI and Coordinator of the European project μRALP and is currently the PI and Coordinator of the project TEEP-SLA.

List of Figures

5 Macro-average receiver operating characteristic (ROC) curves. The mean ($\pm$ standard deviation) curves obtained from the 3 cross-validation folds are reported in bold (transparent area). The mean ($\pm$ standard deviation) area under the ROC curve is reported in the legend. (a) ROC curves for the tested features. Classification is obtained using support vector machines. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Histogram of rotation-invariant uniform local binary patterns. (b) ROC curves for the tested classifiers. Classification is obtained using the histogram of local binary pattern and first order statistics. kNN: k-nearest neighbors, NB: naive Bayes, RF: random forest, SVM: support vector machines.

6 Performance of the state of art. (a) Visual samples of the vessel segmentation obtained applying Barbalata et al.[6] algorithm to patches with hypertrophic vessels (first row), healthy tissue (second row) and intraepithelial papillary capillary loop-like vessels (third row). From left to right, original patch, vessel mask and vessel mask superimposed on the original patch. (b) Normalized confusion matrix obtained applying Barbalata et al.[6] algorithm to our dataset. Colorbar indicates the number of patches. The total number of patches (n. of patches) is reported.

7 Short caption

List of Tables

3 Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec}_{\mathbf{class}}$), precision ($\mathbf{Prec}_{\mathbf{class}}$), and F1-score ($\mathbf{F1}_{\mathbf{class}}$) obtained testing different feature vectors for the *Base* case (i.e., without the inclusion of confidence on classification estimation). Classification is obtained with support vector machines. $Stat_1$: Intensity *mean, variance, entropy*; F_{GLCM}: Gray-level co-occurrence matrix-based descriptors; $H_{LBP_{riu2}}$: Normalized histogram of rotation-invariant uniform local binary patterns.

4 Comparison of different classifiers. Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec}_{\mathbf{class}}$), precision ($\mathbf{Prec}_{\mathbf{class}}$), and F1 score ($\mathbf{F1}_{\mathbf{class}}$) are reported for the four different tissue classes. Classification is obtained using the histogram of local binary patterns and first order statistics. kNN: k-nearest neighbors, NB: naive Bayes, RF: random forest, SVM: support vector machines.

5 Median (first quartile - third quartile) class-specific recall ($\mathbf{Rec}_{\mathbf{class}}$), precision ($\mathbf{Prec}_{\mathbf{class}}$), and F1 score ($\mathbf{F1}_{\mathbf{class}}$) are reported at different level of confidence (τ) on support vector machines classification.