Medical & Biological Engineering & Computing manuscript No.
(will be inserted by the editor)

Learned and handcrafted features for early-stage laryngeal SCC diagnosis

Tiago Araújo, Cristina P. Santos, Elena De Momi, Sara Moccia

Received: date / Accepted: date

Abstract Squamous cell carcinoma (SCC) is the most common and malignant laryngeal cancer. An early-stage diagnosis is of crucial importance to lower patient mortality and preserve both the laryngeal anatomy and vocal-fold function. However, this may be challenging as the initial larynx modifications, mainly concerning the mucosa vascular tree and the epithelium texture and color, are small and can pass unnoticed to the human eye. The primary goal of this paper was to investigate a learning-based approach to early-stage SCC diagnosis, and compare the use of (i) texture-based global descriptors, such as local binary patterns, and (ii) deep-learning-based descriptors. These features, extracted from endoscopic narrow-band images of the larynx, were classified with support vector machines as to discriminate healthy, precancerous and early-stage SCC tissues. When tested on a benchmark dataset, a median classification recall of 98% was obtained with the best feature combination, outperforming the state of the art (recall = 95%). Despite further investigation is needed (e.g. testing on a larger dataset), the achieved results support the use of the developed methodology in the actual clinical practice to provide accurate early-stage SCC diagnosis.

Tiago Araújo
Center for MicroElectroMechanical Systems (CMEMs), Informatics Department, University of Minho, Braga, Portugal
E-mail: a71346@alunos.uminho.pt

Cristina P. Santos
Center for MicroElectroMechanical Systems (CMEMs), Industrial electronics Department, University of Minho, Guimarães , Portugal

E. De Momi
Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

S. Moccia
Department of Information Engineering, Universitá Politecnica delle Marche, Ancona, Italy and with the Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy

Keywords Convolutional Neural Networks · Deep learning · Laryngeal cancer · Machine Learning · Tissue classification · Textural analysis.

1 Introduction

Nowadays, laryngeal cancer is the 14^{th} most common cancer in the world with more than 157,000 estimated new cases [1]. Laryngeal cancer refers to a malignant tumor that affects the larynx, an organ that has a key role in breathing, speaking and swallowing. Laryngeal cancer most commonly takes the form of Squamous Cell Carcinoma (SCC), which normally origins in the squamous epithelium [2].

To decrease mortality rate and preserve both laryngeal anatomy and vocal-cord function, an early-stage SCC diagnosis is crucial. Recently, optical-biopsy techniques for screening purposes are progressively spreading to allow early diagnosis [3]. Screening is commonly performed using laryngoscopy with narrow-band imaging (NBI), which supports visual biopsy by contrasting superficial vessels better than standard white-light endoscopy.

An early-stage diagnosis may not be trivial as the small laryngeal-tissue changes may pass unnoticed to the human eye [4]. Main changes occur in the (i) mucosa vascular tree, with the presence of longitudinal hypertrophic vessels and dot-like vessels (known as intraepithelial papillary capillary loops (IPCL)) [5], and (ii) epithelium, with the thickening and whitening of the epithelial layer (condition known as leukoplakia) [6]. The changes in (i) are indicative early-stage cancerous tissue while the change in (ii) indicates precancerous tissue.

In [7] and [8], some initial efforts of computer-assisted SCC diagnosis are proposed, with [7] specifically focusing 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. However, the classification is strongly sensitive to a-priori set parameters and does not take into account epithelial modifications that do not affect vascular tree (such as the leukoplakia). This algorithm achieved a median classification recall of 42% on the Laryngeal dataset [9].

A more advanced solution has been proposed in [9], where a learning approach based on handcrafted features and support vector machines (SVM) was used to classify healthy, precancerous and early-stage SCC tissues from patches of NBI images, achieving a median classification recall of 93%. The work in [10] provides an analysis of both handcrafted features and learned features, using features extracted from pre-trained convolutional neural networks (CNNs) and achieving a classification recall of 95%.

Outside the field of laryngoscopy other researchers heavily made use of machine learning algorithms to classify tissues according to texture-based information. In [11] and [4], the histogram from the local binary pattern (LBP) was combined with intensity-based features to classify abdominal tissues in laparoscopic images by means of SVM. Other successful handcrafted features are grey-level co-occurrence matrix (GLCM)-based features, [12] and [13], and Gabor filter-based features, [14]. In [15], a CNN was used to classify interstitial lung diseases, achieving a classification performance of 85%, and in [16] CNNs paired with a neighboring ensemble predictor were used to classify cell nuclei in histopathology images of cancerous tissue. In addition, [17] and [18] also use deep neural networks with optical images of skin lesions and retinal fundus to classify skin cancer and predict cardiovascular risk factors, respectively.

A summary of the state of the art for laryngeal cancer diagnosis is presented in Table 1.

In this work, mainly inspired by the work in [9], a system was implemented where feature extraction was applied to the Laryngeal dataset [20]. The dataset consists of 1320 patches, relative to 4 laryngeal tissue classes: healthy tissue, tissue with hypertrophic vessels, leukoplakia and tissue with IPCL-like vessels. Tissue samples for each class are shown in Fig. 1. Here, feature extraction was accomplished with:

- Handcrafted texture-based features, such as LBP and first-order statistics;
- Learned features obtained from pre-trained and fine-tuned CNN models by using transfer-learning.

Table 1 Recent work on laryngeal-tissue classification from optical images.

Author	Methods	Implementation goal
Barbalata et al., 2016 [7]	Linear Discriminant Analysis (LDA) and Matched Filtering (MF)	Discriminate between malignant and benign issue
Moccia, et al., 2017 [9]	LBP, GLCM-based features, SVM and Gini Coefficient (GC)	Classification of laryngeal tissues captured in NBI images
Nanni, et al., 2018 [10]	Deep learning and handcrafted features	Analyze impact of handcrafted and learned features for computer vision

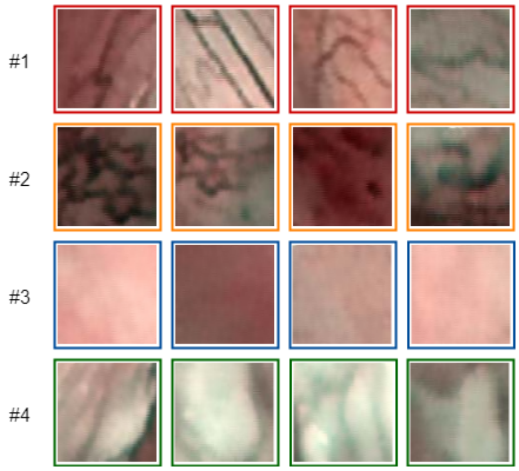

Fig. 1 Sixteen sample patches, four for each of the four analyzed laryngeal tissue classes, are shown. Red: tissue with hypertrophic vessels; Blue: healthy tissue; Orange: tissue with intraepithelial papillary capillary loop-like vessels; Green: tissue with leukoplakia.

Principal components analysis (PCA), [21], was used before classification for dimensionality reduction. Feature classification was performed with multi-class SVMs. With respect to [10], several pre-trained CNN models were studied (also performing feature extraction from various layers for each CNN) and fine-tuning approaches were investigated.

Our main research questions were:

- Research question 1 (RQ1)

Are CNN-based features more powerful than handcrafted ones for laryngeal tissue SVM-based classification?

- Research question 2 (RQ2)
  Can the performance of CNN-based features be improved by performing fine-tuning on pre-trained CNN models?

This paper is organized as follows: Sec. 2 explains the main methods exploited in this paper and the dataset used to validate it; Sec. 3 explains the evaluation protocol used to investigate our two research questions; Sec. 4 and Sec. 5 presents and discusses the main results of this work, respectively.

2 Methods

The workflow of the proposed method is shown in Fig. 2.

2.1 Pre-processing and feature extraction

Three main classes of features were extracted from the patches of the Laryngeal dataset: texture-based global descriptors, first-order statistics and CNN-based learned features. Prior to feature extraction, the patches underwent Gaussian smoothing.

2.1.1 Texture-based global descriptors

Inspired by [9], the texture-based global descriptor chosen for this work was LBP, as they are considered the state of the art for medical image texture analysis. LBP are grey-scale invariant and of low-complexity.

The base formulation of LBP ($LBP_{R,P}$) requires the definition, for a pixel $c = (c_x, c_y)$, a spatial circular neighborhood of radius R with P equally-spaced neighbor points ($\{P_n\}_{n \in (0,P-1)}$):

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

where g_c and g_{p_n}, denote the grey values of pixel c and of its n^th neighbor p_n, respectively and s is defined as:

$$s(g_{p_n} - g_c) = \begin{cases} 1 & g_{p_n} \geq g_c \\ 0 & g_{p_n} < g_c \end{cases} \tag{2}$$

The formulation for LBP adopted in this work, the one used more often, was the uniform rotation-invariant LBP, as rotation invariance is appropriate in this case since the endoscope pose during the larynx inspection is constantly shifting.

2.1.2 First-order statistics

For each patch the intensity *mean*, *variance* and *entropy* are computed and concatenated to form a single intensity-based feature set (STAT), as per [9]. The entropy is defined as:

$$entropy = -\sum_{i=0}^{255} h_i \log_2(h_i) \tag{3}$$

where h_i is the image histogram counts.

2.1.3 CNN-based learned features with transfer learning

CNNs are a class of deep, feedforward artificial neural networks. CNNs are composed of interconnected neurons with learnable weights, biases and activation functions. These networks are essentially built with four type of layers: convolutional, activation, pooling and fully-connected. The convolutional layer is the core building block of a CNN, pooling layers perform non-linear downsampling, activation layers apply activation functions (e.g. rectified linear unit) and fully-connected layers are made of neurons that have full connections to the previous layer.

For fine-tuning, the layer weights of the original CNN were frozen up to a certain layer and our data was used to train the unfrozen part of the network. Each model has a different separating layer, which will be described below.

Here, the following pre-trained CNN models were investigated: ResNet V2 with 101 layers [22], Inception V4 [23], and Inception-ResNet V2 [24].

The ResNet V2 with 101 layers model architecture and freezing method are shown in Fig. 3. This model consists of four groups of multiple building blocks each. A building block has 3 convolutional layers. To fine tune this model, our data was used to train only layers from the last group forward. Additionally, for each of the models with origin in this model the features from earlier layers were obtained: right after the pooling layer (Layer A for easier addressing) and after the spatial squeeze layer (Layer B).

The Inception V4-model architecture as well as the freezing method are depicted in Fig. 4. For this model, besides the final layer, the following stopping points were considered for feature extraction: right after the 7x Inception-B layers (Layer C), after the Average Pooling layer (Layer D) and after the Dropout layer (Layer E).

Finally, Inception-ResNet V2 model architecture and freezing method are shown in Fig. 5. As it can be seen in Fig. 5 the model was trained from the last block of layers forward and the following ending points used:

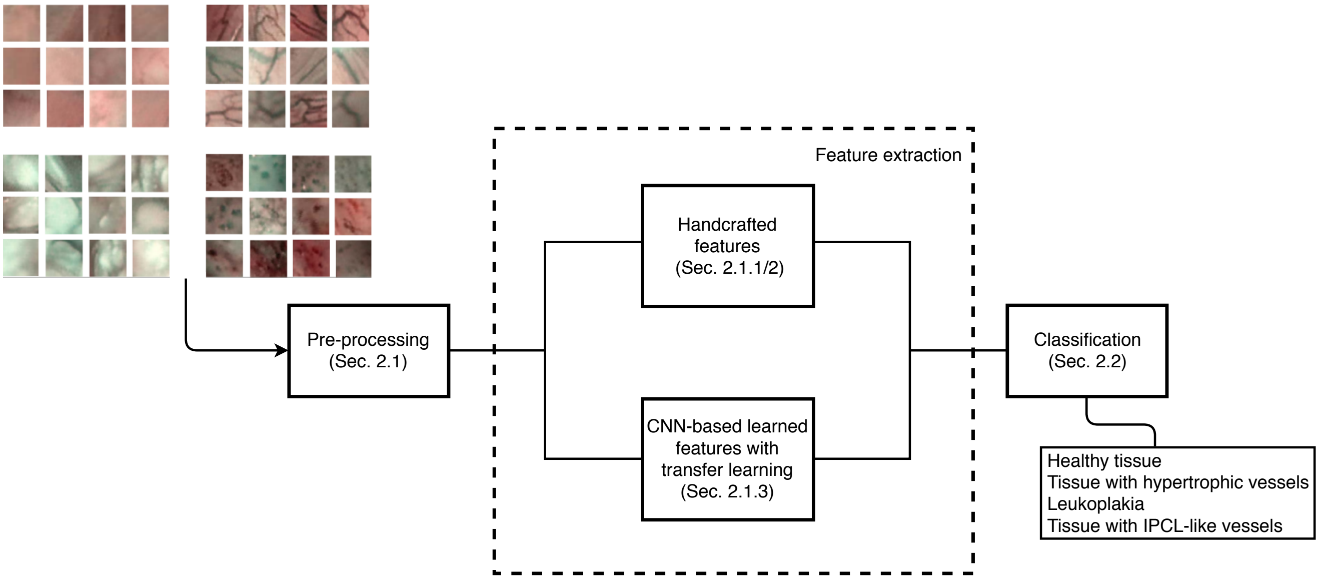
Feature extraction
Handcrafted features (Sec. 2.1.1/2)
Pre-processing (Sec. 2.1)
Classification (Sec. 2.2)
CNN-based learned features with transfer learning (Sec. 2.1.3)
Healthy tissue
Tissue with hypertrophic vessels
Leukoplakia
Tissue with IPCL-like vessels

Fig. 2 Workflow of the proposed solution for early-stage laryngeal cancer diagnosis.

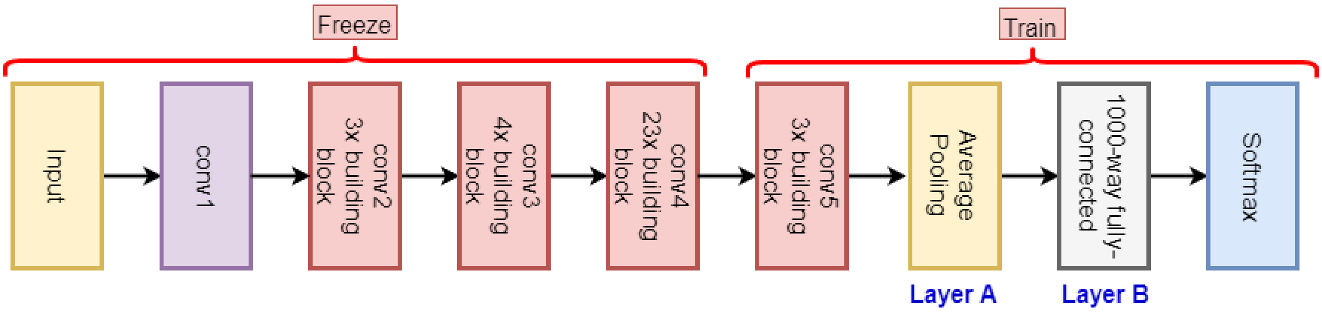

Fig. 3 Architecture of ResNet v2 with 101 layers. Frozen layers and layers that are fine tuned are shown too.

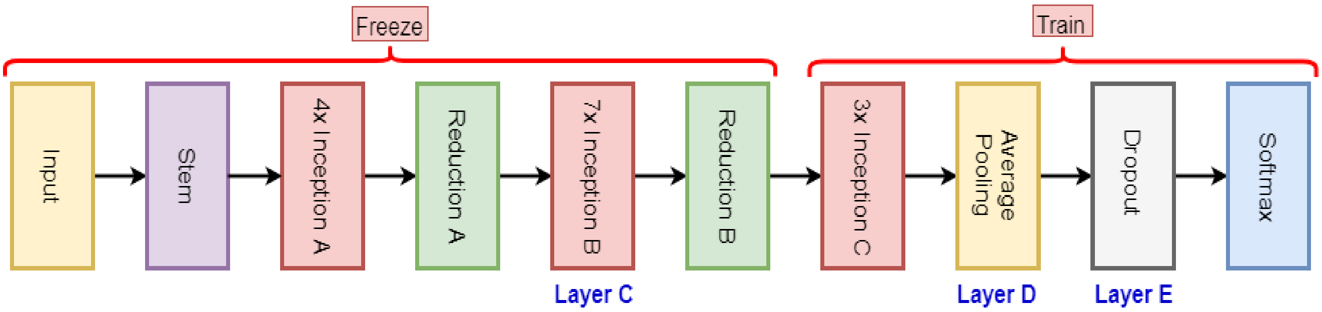

Fig. 4 Architecture of Inception V4. Frozen layers and layers that are fine tuned are shown too.

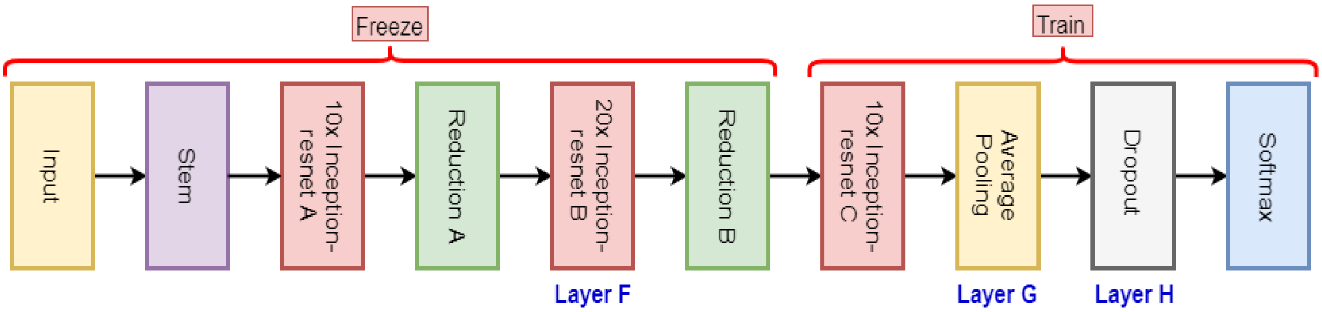

Fig. 5 Architecture of Inception-ResNet V2. Frozen layers and layers that are fine tuned are shown too.

Table 2 Dataset characteristics

	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 (4 per image)	440 (4 per image)	440 (4 per image)	1320

right after the 20-layer block (Layer F, which can be seen "diverging" from the main network), after the average pooling layer (Layer G), after the dropout layer (Layer H) and of course the last layer of the model.

2.2 Classification

In this paper SVM, [25], was used for tissue classification in four classes. The basis of SVM is finding a hyperplane that best divides a dataset into two classes (binary problem). However, they can also be adjusted to work with multi class problems. SVM handles the high dimensionality, characteristic of our problem and with the *kernel-trick* it prevents parameter proliferation, limiting over-fitting and lowering computational complexity, [26] and [27]. The radial basis function kernel (Gaussian kernel) was used. Additionally SVM is very robust to noise in training data. To implement multi-class SVM classification, the one-vs-rest scheme was used.

3 Evaluation protocol

In this work, as introduced in Sec. 1, the publicly available Laryngeal dataset [20] was used. This dataset consists of images from 33 NBI videos, corresponding to 33 patients affected by SCC, acquired by an NBI endoscopic system (Olympus Visera Elite S190 video processor and an ENF-VH rhino-laryngo videoscope) with a frame rate of 25 frames per second and image size of 1920 pixels x 1072 pixels. For these videos, 10 images were manually selected from each video obtaining a total of 330 images.

For each of those images, 4 patches were cropped with a size of 100 pixels x 100 pixels, for a total of 1320 patches equally distributed between four classes, namely, healthy tissue, tissue with hypertrophic vessels, leukoplakia and tissue with IPCL-like vessels, as mentioned in Sec.1. Each patch was cropped from a portion of the tissue relative to only one of the four considered classes. Both the previous image and the following patch selection were performed under the supervision of an expert clinician (otolaryngologist specialized in head and neck oncology). The dataset characteristics are shown in Table 2.

The patches of the dataset were initially pre-processed with a Gaussian filter (standard deviation (σ) = 0.8). In the feature extraction step, LBPs were computed with the following (R, P) combinations: (1,8), (2,16), (3,24) for each RGB channel, and the corresponding L2-normalized histograms were concatenated. This choice allowed multi-scale and, therefore, a more accurate description of the texture information. Adding the STAT features, for each patch, a 172-feature long vector was obtained.

As for CNN-based features, transfer learning was used with CNN models pre-trained on the ImageNet dataset [28]. These models are the ones mentioned previously in Sec. 2.1.3, namely ResNet V2 with 101 layers, Inception V4 and Inception-ResNet V2. These models provided a 1000-feature long vector.

PCA [21] was used for feature dimensionality reduction by selecting the principal components that described 95% of the data variance. This process was followed by the classification.

For the classification, SVM with the radial basis function and the one-vs-rest scheme was used. 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.

To obtain a robust estimation of the classification performance, 3-fold cross validation was performed, separating data at patient level to prevent data leakage, as per [9]. The 1320 patch dataset was split to obtain well-balanced folds both patient and tissue wise, as shown previously in Table 2. Each time, two folds were used for training and the other one for testing only. Therefore, this evaluation does not lead to biased results.

3.1 Evaluation Metrics

The following metrics were used to evaluate the classification performance. It was calculated the class-specific recall:

$$\mathrm{Rec}_{\mathrm{class_j}} = \frac{\mathrm{TP_j}}{\mathrm{TP_j} + \mathrm{FN_j}} \tag{4}$$

where $\mathrm{TP_j}$ is the number of elements of the $\mathrm{j^{th}}$ class correctly classified (true positive of the $\mathrm{j^{th}}$ class) and $\mathrm{FN_j}$ the number of elements of the $\mathrm{j^{th}}$ class wrongly assigned to one of the other classes (false negative of the j$^{\text{th}}$ class). The class-specific precision was evaluated, where:

$$\mathrm{Prec}_{\mathrm{class_j}} = \frac{\mathrm{TP_j}}{\mathrm{TP_j} + \mathrm{FP_j}} \tag{5}$$

being $\mathrm{FP_j}$ the number of false positive of the j$^{\text{th}}$ class, the F1 score was computed, where:

$$\mathrm{F1}_{\mathrm{class_j}} = 2 \times \frac{\mathrm{Prec}_{\mathrm{class_j}} \times \mathrm{Rec}_{\mathrm{class_j}}}{\mathrm{Prec}_{\mathrm{class_j}} + \mathrm{Rec}_{\mathrm{class_j}}} \tag{6}$$

finally, the accuracy of the model was also calculated:

$$\mathrm{Acc}_{\mathrm{class_j}} = \frac{\mathrm{TP_j} + \mathrm{TN_j}}{\mathrm{TP_j} + \mathrm{TN_j} + \mathrm{FP_j} + \mathrm{FN_j}} \tag{7}$$

where TP is the number of true positives in the model, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

The implementation of LBP feature extraction, PCA and classification was performed with scikit-learn.

Tensorflow-Slim was used for the CNN feature extraction. All the computational efforts were done with an Intel Core i7-6700K CPU @ 4.00GHz.

3.2 Experiments

To answer our research questions, the following investigations were carried out:

3.2.1 Investigation of RQ1

To assess our hypothesis that CNN-based features are more powerful than handcrafted texture-based features for laryngeal tissue classification the SVM results using LBP and STAT (LBPS), CNN-based features, and LBPS paired with the CNN-based feature were compared. The tests were performed with the models mentioned in Sec. 2.1.3. For each of these models, the features were extracted from the last layer of each CNN.

3.2.2 Investigation of RQ2

To investigate the possibility of improving the performance of the laryngeal tissue classification by fine-tuning the pre-trained models, two cases were considered. The former, freezes part of the model and train the rest with our data, and the latter trains the entire network from scratch.

Also, for each of the original models and the ones created in case one and two features from different layers were extracted, as stated in Sec. 2.

Table 3 Research question 1 results. The first four columns consist in the percentages of the metrics mentioned before, the median across the three folds, and the last one is the total classification time in seconds. The first row is feature extraction with LBPS only, from 2-4 with CNN only and the rest with LBPS paired with CNN. The case in bold corresponds to the best result in the table. The other highlighted cases (italic) correspond to cases important in the comparison with the best result.

	Rec (%)	Prec (%)	F1 (%)	Acc (%)	Total time (s)
LBPS	*94*	*94*	*94*	*94*	*14*
Inception-v4	93	94	93	93	25
ResNet V2 101 layers	*95*	*95*	*95*	*95*	*36*
Inception-ResNet V2	94	95	94	94	18
LBPS + Inception-v4	96	96	96	96	27
LBPS + ResNet V2 101 layers	**98**	**98**	**98**	**98**	**41**
LBPS +Inception-Resnet V2	97	98	98	98	20

4 Results

The results are organized according to the research questions.

4.1 RQ1 results

CNN based features proved to be superior than handcrafted features as it can be seen in Table 3 where LBPS had a median classification recall of 94% and ResNet V2 with 101 layers achieved 95%. The other CNN models, i.e. Inception-v4 and Inception-ResNet V2, had similar performance to LBPS, achieving a classification median recall of 93% and 94% respectively. However, the best result was achieved with a combination of both: LBPS and CNN features extracted with ResNet V2 with 101 layers with a median classification recall of 98% as seen in Fig. 6.

4.2 RQ2 results

With the combination of LBPS and FT INC Layer D (or E) features, a median classification recall of 97% was achieved, overcoming the SVM-based results in Table 3 with 96%.

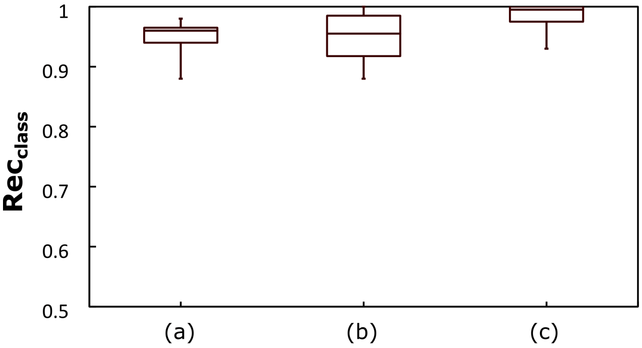

Fig. 6 Research question 1 boxplots of classification recall (Rec_{Class}) obtained when using (a) local binary pattern and first order statistics features, (b) ResNet V2 with 101 layers based features (c) and both.

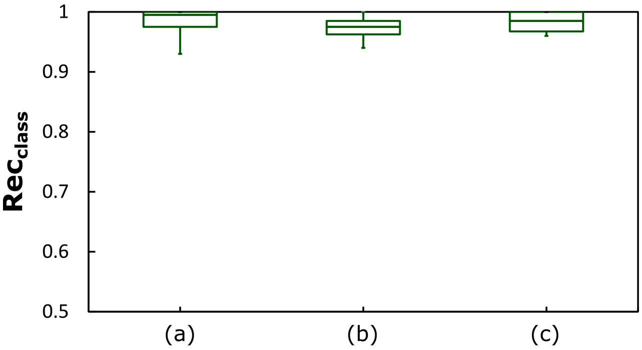

Fig. 7 Boxplots of classification recall (Rec_{Class}) for research question 2. The boxplots are obtained using (a) local binary pattern and first order statistics features (LBPS) with Resnet V2 with 101 layers based features, (b) LBPS features with features based on fine-tuned Inception V4 Layer D and (c) LBPS features combined with features extracted from Inception-ResNet V2 Layer F .

The best results, as seen in Fig. 7, were:

- Features from LBPS in combination with ResNet V2 with 101 layers Layer B, with median classification recall of 98%
- LBPS with FT Inception V4 Layer D features, with median classification recall of 97%
- LBPS and Inception-ResNet V2 Layer F features, with a median classification recall of 97%

5 Discussion

From the results of Table 3, learned features outperformed handcrafted ones. Additionally a combination of handcrafted features and learned features contributed to further improve the classification performance demonstrating that the two methods extract different information from the input images, reflecting the results of Nanni et al. [10]. As for the CNN models, the more complex ones, Inception-ResNet V2 and ResNet V2 with 101

Table 4 Research question 2 results. The first four columns consist in the percentages of the metrics mentioned before, the median across the three folds, and the last one is the total classification time in seconds. The rows from 1-6 correspond to the RN network, from 7-13 to the INC network and the rest to the INC-RN network. The cases in bold correspond to the best result for each network in the table. The other highlighted cases (italic) correspond to cases important in the comparison with the best result for each network.

	Rec (%)	Prec (%)	F1 (%)	Acc (%)	Total time (s)
LBPS + RN original Layer A	97	97	97	97	163
LBPS + RN original Layer B	**98**	**98**	**98**	**98**	**41**
LBPS + FT RN last layer	*94*	*95*	*94*	*94*	*13*
LBPS + FT RN model Layer B	94	95	93	94	13
LBPS + Scratch RN last layer	*93*	*94*	*92*	*93*	*14*
LBPS + Scratch RN Layer B	93	94	93	93	14
LBPS + INC original	96	96	96	96	27
LBPS + INC original Layer C	95	95	95	95	25
LBPS + INC original Layer D	97	97	97	97	96
LBPS + INC original Layer E	97	97	97	97	97
LBPS + FT INC Layer D	**97**	**97**	**97**	**97**	**101**
LBPS + FT INC Layer E	97	97	97	97	111
LBPS + Scratch INC last layer	*93*	*93*	*93*	*93*	*13*
LBPS + INC-RN original Layer F	**97**	**98**	**98**	**98**	**70**
LBPS + INC-RN original Layer G	97	97	97	97	69
LBPS + INC-RN original Layer H	97	97	97	97	70
LBPS + FT INC-RN Layer F	93	94	93	93	14
LBP + FT INC-RN Layer G	*97*	*97*	*97*	*97*	*72*
LBPS + FT INC-RN Layer H	97	97	97	97	72
LBPS + Scratch INC-RN Layer G	*94*	*94*	*94*	*94*	*13*

Table 5 Comparison between the best results in literature for laryngeal cancer classification. Barbalata et al., 2016 algorithm applied to the Laryngeal dataset.

	Barbalata et al., 2016 [7]	Moccia et al. [9]	Nanni et al. [10]	LBPS + RN original Layer B
Recall %	42	93	94	98

layers, demonstrated superior results than the Inception-V4. This is probably because the deeper the network, the more detailed the features it is able to extract from the image.

From the results relative to RQ2 in Table 4, the models trained from scratch with our dataset underperformed the other cases. This can be due to the fact that our dataset was relatively small. The models resulting from fine-tuning with layers which also suffered a reduction in the number of features, namely the last layers, layer B, Layer C and Layer F, had a huge hit in their performance. However, the cases related to the other layers (Layer A, D, E, G and H) were on par with the results of Table 3. Nonetheless, we expect that a bigger dataset coupled with fine-tuning could lead to a further increase in the overall performance of the laryngeal-tissue classification.

As for executing the feature extraction in different layers for each model it increased the overall performance for most networks by a small margin, confirming that the layer where features are extracted can lead to a better laryngeal tissue classification performance. Lastly the performance obtained using transfer learning was higher than standard methods in the literature. A comparison of the best results in literature for laryngeal cancer classification can be seen in Table 5.

Finally, the results reported were visually evaluated by three expert clinicians, which agreed with our conclusions.

6 Conclusion

In this project, a learning-based system, inspired by [9, 10], for early-stage detection of laryngeal cancer was implemented. The system processed NBI images using transfer learning with pre-trained CNN models.

Assisted by a comprehensive evaluation and an exhaustive analysis of the results, we demonstrated that the proposed approach can lead to high-performance classification results, overcoming the state of the art. Such high performance was achieved independently from patient-specific parameters or arbitrary thresholds. This makes the proposed solution a proper tool to face the complexity and high diversity of laryngeal pathological tissues.

For future work, we intend to perform image segmentation to define tumoral margins. This would be helpful with a view to assist surgeons with robotic procedures,e.g implementing virtual fixture. Challenges in this direction will consist in the definition of a proper CNN architecture for reliable, accurate and real-time segmentation (e.g. [29], [30] and [31]).

References

1. S. McGuire. World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015 (2016)
2. K. Markou, A. Christoforidou, I. Karasmanis, G. Tsiropoulos, S. Triaridis, I. Constantinidis, V. Vital, A. Nikolaou, Hippokratia **17**(4), 313 (2013)
3. J. Unger, J. Lohscheller, M. Reiter, K. Eder, C.S. Betz, M. Schuster, Cancer Research (2014)
4. P. Liang, Y. Cong, M. Guan, in *IEEE International Conference on Information and Automation* (IEEE, 2012), pp. 871–875
5. C. Piazza, F. Del Bon, G. Peretti, P. Nicolai, Current Opinion in Otolaryngology & Head and Neck Surgery **20**(6), 472 (2012)
6. J.S. Isenberg, D.L. Crozier, S.H. Dailey, Annals of Otology, Rhinology & Laryngology **117**(1), 74 (2008)
7. C. Barbalata, L.S. Mattos, IEEE Journal of Biomedical and Health Informatics **20**(1), 322 (2016)
8. H.I. Turkmen, M.E. Karsligil, I. Kocak, Computers in Biology and Medicine **62**, 76 (2015)
9. S. Moccia, E. De Momi, M. Guarnaschelli, M. Savazzi, A. Laborai, L. Guastini, G. Peretti, L.S. Mattos, Journal of Medical Imaging **4**(3), 034502 (2017)
10. L. Nanni, S. Ghidoni, S. Brahnam, Applied Computing and Informatics (2018). DOI https://doi.org/10.1016/j.aci.2018.06.002
11. Y. Zhang, S.J. Wirkert, J. Iszatt, H. Kenngott, M. Wagner, B. Mayer, C. Stock, N.T. Clancy, D.S. Elson, L. Maier-Hein, in *Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9786 (International Society for Optics and Photonics, 2016), vol. 9786, p. 978619
12. X. Shen, K. Sun, S. Zhang, S. Cheng, in *IEEE International Conference on Signal Processing, Communication and Computing* (IEEE, 2012), pp. 756–759
13. M. Misawa, S.e. Kudo, Y. Mori, K. Takeda, Y. Maeda, S. Kataoka, H. Nakamura, T. Kudo, K. Wakamura, T. Hayashi, et al., International Journal of Computer Assisted Radiology and Surgery **12**(5), 757 (2017)
14. F. Van Der Sommen, S. Zinger, E.J. Schoon, et al., in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670 (International Society for Optics and Photonics, 2013), vol. 8670, p. 86700V
15. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, IEEE Transactions on Medical Imaging **35**(5), 1207 (2016). DOI 10.1109/TMI.2016.2535865
16. K. Sirinukunwattana, S.E.A. Raza, Y. Tsang, D.R.J. Snead, I.A. Cree, N.M. Rajpoot, IEEE Transactions on Medical Imaging **35**(5), 1196 (2016). DOI 10.1109/TMI.2016.2525803

17. A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Nature **542**, 115 EP (2017)
18. R. Poplin, A.V. Varadarajan, K. Blumer, Y. Liu, M.V. McConnell, G.S. Corrado, L. Peng, D.R. Webster, Nature Biomedical Engineering **2**(3), 158 (2018). DOI 10.1038/s41551-018-0195-0
19. G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, Medical Image Analysis **42**, 60 (2017)
20. S. Moccia, E.D. Momi, L.S. Mattos. Laryngeal dataset (2017). DOI 10.5281/zenodo.1003200. URL https://doi.org/10.5281/zenodo.1003200
21. I. Jolliffe. Principal component analysis (2011). DOI 10.1007/978-3-642-04898-2_455
22. K. He, X. Zhang, S. Ren, J. Sun, in *IEEE Conference on Computer Vision and Pattern Recognition* (2016)
23. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *IEEE Conference on Computer Vision and Pattern Recognition* (2015)
24. C. Szegedy, S. Ioffe, V. Vanhoucke, Computing Research Repository **abs/1602.07261** (2016). URL http://arxiv.org/abs/1602.07261
25. C.J. Burges, Data Mining and Knowledge Discovery **2**(2), 121 (1998). DOI 10.1023/A:1009715923555
26. G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, in *Workshop on Statistical Learning in Computer Vision* (2004), pp. 1–22
27. Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, T. Huang, in *Conference on Computer Vision and Pattern Recognition* (2011), pp. 1689–1696. DOI 10.1109/CVPR.2011.5995477
28. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, International Journal of Computer Vision **115**(3), 211 (2015). DOI 10.1007/s11263-015-0816-y
29. R. Vivanti, L. Joskowicz, N. Lev-Cohain, A. Ephrat, J. Sosna, Medical & Biological Engineering & Computing **56**(9), 1699 (2018). DOI 10.1007/s11517-018-1803-6
30. N. Hatipoglu, G. Bilgin, Medical & Biological Engineering & Computing **55**(10), 1829 (2017). DOI 10.1007/s11517-017-1630-1
31. S. Moccia, S. Foti, A. Routray, F. Prudente, A. Perin, R. F. Sekula, L. S. Mattos, J. R. Balzer, W. Fellows, E. De Momi, C. Riviere, Annals of Biomedical Engineering **46** (2018). DOI 10.1007/s10439-018-2091-x

Tiago Araújo, is a Master's Student at the University of Minho, at the Department of Informatics. His Master's research, under the supervision of Professor Cristina P Santos at the University of Minho, is on early-stage laryngeal cancer segmentation from endoscopic key images using deep-learning approaches. His current research interests are machine learning and medical imaging processing.

Cristina P Santos (DEng, Msc, PhD, Habil), is an Assistant Professor at UMinho, at the Department of Industrial Electronics & and a researcher at Research Center CMEMs at the University of Minho, Portugal. She was Head Director of Integrated Master in Biomedical Engineering (2013-2015) and is currently a member of the Directive Board of the Doctoral Program in biomedical Engineering. Her work focuses on methods to characterize human motion, the study of human locomotion, Human-Robot Interaction & Collaboration, medical devices, and the neuro-rehabilitation of patients suffering from motor problems by means of bio-inspired robotics and neuroscience technologies.

Sara Moccia achieved the European Ph.D. degree cum laude in Bioengineering on May, 16th 2018, with a thesis entitled "Supervised tissue classification in optical images: Towards new applications of surgical data science". Sara pursued her Ph.D in collaboration with the Department of Electronics, Information and Bioengineering at Politecnico di Milano (Milan, Italy) and the Department of Advanced Robotics at Istituto Italiano di Tecnologia (Genoa, Italy). During her Ph.D, she spent six months at the "Computer-Assisted Medical Intervention" laboratory at the German Cancer Research Centre (Heidelberg, Germany).Sara is now PostDoc in the Department of Information Engineering at Università Politecnica delle Marche (Ancona, Italy) and research fellow at the Department of Advanced Robotics at Istituto Italiano di Tecnologia (Genoa, Italy). Her main research activity deals with developing machine-learning and deep-learning algorithms for medical-image analysis to provide diagnostic support and context awareness.

Elena De Momi, MSc in Biomedical Engineering in 2002, PhD in Bioengineering in 2006, currently Associate Professor in Electronic Information and Bioengineering Department (DEIB) of Politecnico di Milano. She is co-founder of the Neuroengineering and Medical Robotics Laboratory, in 2008, being responsible of the Medical Robotics section. IEEE Senior Member, she is currently Associate Editor of the Journal of Medical Robotics Research, of the International Journal of Advanced Robotic Systems, Frontiers in Robotics and AI and Medical & Biological Engineering & Computing. From 2016 she has been an Associated Editor of IEEE ICRA, IROS and BioRob and she is currently Publication Co-Chair of ICRA 2019. She is responsible for the lab course in Medical Robotics and of the course on Clinical Technology Assessment of the MSc degree in Biom. Eng. at Politecnico di Milano and she serves in the board committee of the PhD course in Bioengineering. Her academic interests include image-processing, virtual environments, augmented reality and simulators, teleoperation, haptics, medical robotics, human robot interaction.