Classifying Vocal Folds Fixation from Endoscopic Videos with Machine Learning

Francesca Pia Villani[1], Alberto Paderno[2], Maria Chiara Fiorentino[3], Alessandro Casella[4], Cesare Piazza[2] and Sara Moccia[5]

***Abstract*— Vocal folds motility evaluation is paramount in both the assessment of functional deficits and in the accurate staging of neoplastic disease of the glottis. Diagnostic endoscopy, and in particular videoendoscopy, is nowadays the method through which the motility is estimated. The clinical diagnosis, however, relies on the examination of the videoendoscopic frames, which is a subjective and professional-dependent task. Hence, a more rigorous, objective, reliable, and repeatable method is needed. To support clinicians, this paper proposes a machine learning (ML) approach for vocal cords motility classification. From the endoscopic videos of 186 patients with both vocal cords preserved motility and fixation, a dataset of 558 images relative to the two classes was extracted. Successively, a number of features was retrieved from the images and used to train and test four well-grounded ML classifiers. From test results, the best performance was achieved using XGBoost, with precision = 0.82, recall = 0.82, F1 score = 0.82, and accuracy = 0.82. After comparing the most relevant ML models, we believe that this approach could provide precise and reliable support to clinical evaluation.**

***Clinical Relevance*— This research represents an important advancement in the state-of-the-art of computer-assisted oto-laryngology, to develop an effective tool for motility assessment in the clinical practice.**

I. INTRODUCTION

With the advent of artificial intelligence (AI), the last decade has seen a revolution in the field of medical-image analysis, with applications ranging from diagnosis to treatment, guidance, and follow-up [1]. While promising results were obtained for processing anatomical images, such as computerized-tomography [2], ultrasound [3], and magnetic-resonance [4] images, the analysis of endoscopic videos still represents a challenge [5] and only few commercially-available solutions exist [6]. This may be explained considering the peculiar challenges of endoscopic videos, including poor contrast, low signal-to-noise ratio, presence of motion blurring, and tissue motion. The field of otolaryngology and head and neck surgery makes not an exception [7]. Videoendoscopy is largely used in clinical practice for a number of applications, among which the assessment of vocal cords motility. Paralysis of one or both vocal folds may jeopardize key physiological functions of the larynx, from breathing to airway protection and phonation [8]. The clinical diagnosis relies on the subjective examination and interpretation of vocal folds motion during real-time viewing or playback of videos captured through videoendoscopy. This evaluation is time-consuming and requires a skilled professional to be performed, and is characterized by high inter- and intra-rater variability [9]. In this context, machine learning (ML) has the potential to tackle the variability of videoendoscopic frames, and to provide a quantitative perspective to the analysis of vocal folds motility. In this paper, we focus on the analysis of endoscopic frames extracted from endoscopic videos, proposing a ML algorithm for the assessment of vocal folds motility.

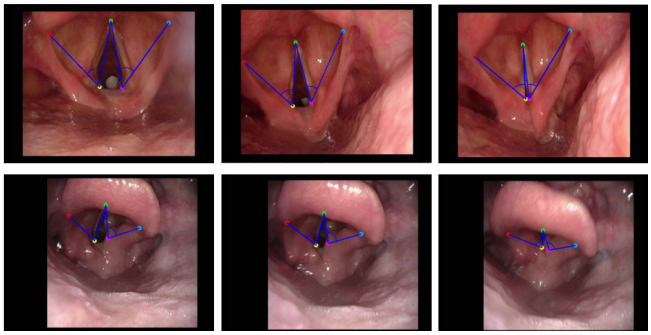

Fig. 1: Representation of three consecutive frames (from left: abducted, normal, and adducted vocal cords, respectively) with ground truth keypoints annotation. The images in the first row refer to a subject with preserved motility, while the ones in the second row to a subject with fixation. The colored points represent the keypoints: left epiglottic in red, left vocal fold in yellow, anterior commissure in green, right vocal fold in magenta, right epiglottic in cyan.

The literature on ML algorithms for laryngeal videoendoscopic image analysis has been growing since 2017. The work in [10] is among the first to investigate the use of ML algorithms for early stage cancerous laryngeal tissue classification. Since then, several studies have been published, including a recent review [7]. Motility assessment, instead, is mostly addressed with deep learning (DL) methods based

*This work was not supported by any organization

[1] F.P. Villani is with the Department of Humanities, Università degli Studi di Macerata, Italy f.villani2@unimc.it

[2]A. Paderno and C. Piazza are with the Department of Otorhinolaryngology – Head and Neck Surgery, ASST Spedali Civili of Brescia, University of Brescia, Italy

[3]M.C. Fiorentino is with the Department of Information Engineering, Università Politecnica delle Marche, Italy

[4]A. Casella is with the Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy and the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[5]S. Moccia is with The BioRobotics Institute, Scuola Superiore Sant'Anna and the Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Italy

on glottal segmentation, from which the motility is evaluated based on the movement of each fold with respect to the midline; or through region of interest detection, and glottal gap delimitation [11]. Hamad et al. [12] developed a DL system for automatic segmentation of the glottal region in laryngoscopy videos using a fully convolutional regression network. More recently, Yousef et al [13] studied vocal folds kinematics during the running speech, analyzing vocal folds vibrations in adductor spasmodic dysphonia. A U-Net was deployed for glottal area segmentation in high-speed videoendoscopy to quantitatively analyze vibrations in both healthy and unhealthy patients. Similarly, in [14] vocal folds dynamics is evaluated in association with voice disorders. They trained a deep neural network with data from laryngeal high-speed videoendoscopy with the aim of segmenting the glottal area, from which the glottal edges are derived during connected speech. Other studies make use of phasegram [15] (a visualization method of system dynamics that can be interpreted as a bifurcation diagram in time) or phonovibrogram [9], [16] (a graphical representation of the vocal folds deflections, automatically extracted from laryngeal high speed recordings) to evaluate vocal folds motility related with voice disorders. However, unlike videoendoscopy, these kinds of tests are not usually performed in clinical practice.

Differently from the work in the literature, we rely on ML for vocal folds motility estimation. We propose, in fact, a method to classify motility into two classes (namely: preserved motility and fixation) based on keypoints. This method is advantageous as it allows to directly obtain a classification, without the need of post-processing, as in the case of glottal segmentation. Each of the selected keypoints represents an important clinical landmark for the analysis, providing a close approximation of both glottic and arytenoid movements. Starting from the coordinates of the five keypoints, clinically relevant features were handcrafted to train the classification models.

II. METHODS

A. Vocal Folds Model and Keypoints Annotation

The dataset used for this analysis is made of videoendoscopic frames of patients treated at the Unit of Otorhinolaryngology - Head and Neck Surgery, University of Brescia, Italy. Data were acquired following the principles of the Helsinki Declaration, and approval was obtained by the local ethical committee of Spedali Civili of Brescia. A total of 558 endoscopic images from 186 patients was collected from a dedicated archive and anonymized, and for each video three representative frames were selected. The motility was estimated among these three endoscopic frames from five keypoints chosen according to the clinical experience of the clinicians, and located at specific sites of the larynx: the epiglottic insertion point of the left and right aryepiglottic folds (LE and RE, respectively), the posterior angle of the left and right vocal folds (LV and RV, respectively), and the anterior commissure (A), as shown in Fig. 1. RE and LE represent the insertion of the aryepiglottic folds in the epiglottis. In particular, they are one of the pivot points that remain fixed when the arytenoid moves (together with the aryepiglottic fold). Hence, they are suitable reference points when trying to assess the movements of the supraglottic larynx, using the angle formed by them and the vocal folds.

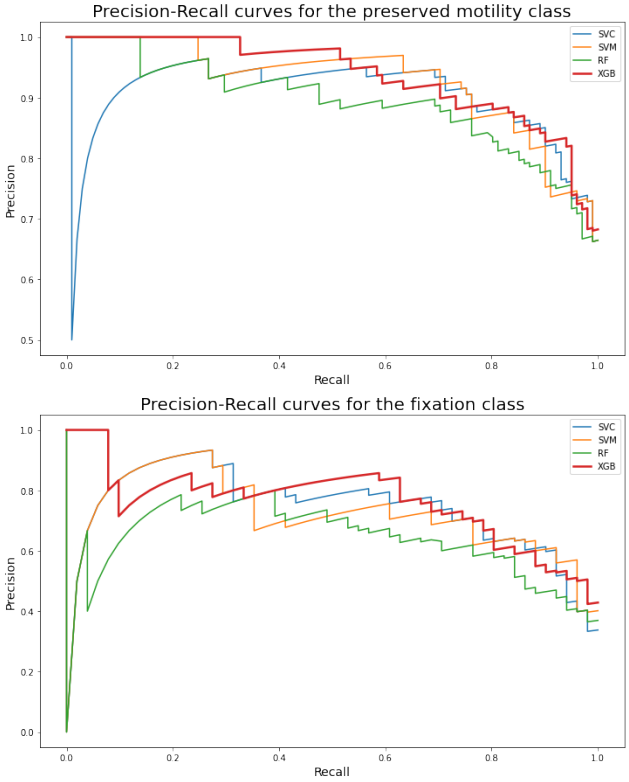

Fig. 2: Precision-Recall curves calculated on the test set, for all the classifiers. The best-performing classifier resulted to be XGB, showing the highest average precision and area under the precision-recall curve for both classes.

Frames annotation was performed by an expert (more than 10 years of experience) laryngologist using LabelMe[1]. Only subjects for which three frames representing a specific vocal folds position (abducted, neutral, adducted) were available, were included in the study. After this process of data selection, the collected dataset counted 101 subjects with preserved motility and 51 subjects with fixation. The dataset includes both oncologic and non-oncologic patients.

B. Features Extraction and Classification

To assess vocal folds motility, we extracted the following features from the labeled frames:

- The central and the two external angles for each frame (as shown in Fig. 1).
- The static index: the difference between the two external angles for each of the three frames.
- The dynamic index: the ratio between the difference of the right angle in the first and third frames and the difference of the left angle in the first and third frames.

[1]https://github.com/wkentaro/labelme

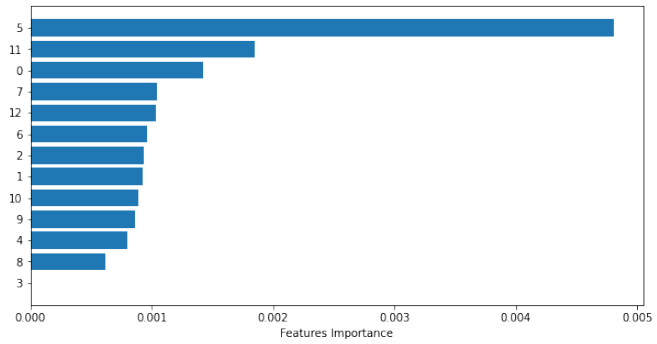

Fig. 3: Features importance of the XGB classifier. Features from 0 to 8 refer to the three angles (central and externals) of the three successive frames, features from 9 to 11 refer to the static indexes of the three frames, and feature 12 refers to the dynamic index.

We investigated common ML classification algorithms, including support vector machines with linear (SVC) and non linear (SVM) kernels, XGBoost (XGB), and random forest (RF). The optimal hyperparameters for each classifier were retrieved via grid-search and cross validation on the training set, using stratified three-fold cross validation. This ensures that every patient in our dataset appears at least once in the testing set. In particular, the three-fold cross validation cyclically splits the dataset into three equally sized folds, of which two are used to train and one to validate and tune the parameters. Before classification, features were normalized by removing the mean (centering) and scaling to unit variance. Given the unbalance between the two classes, the minority class was over-sampled using the synthetic minority oversampling technique (SMOTE). Also class weights were balanced according to the number of samples of each class.

C. Experimental Analysis

The performance of the classifiers was evaluated using classification precision (Prec), recall (Rec), and F1-score (F1) on the test set. Considering the unbalance of our dataset, the area under the precision-recall curve (AUC) and the average precision (AP) were also computed.

III. RESULTS

The performance of all the classifiers is shown in Table I, results are reported in terms of the metrics computed on the test set. Fig. 2 shows the precision-recall curves of all the classifiers. All the tested models showed comparable results, however, the best-performing classification algorithm resulted to be the XGB, with an AP of 0.76 and 0.94, and an AUC of 0.76 and 0.93 for the fixation and preserved motility class, respectively. Features importance of the XGB classifier is reported in Fig. 3. Specifically, on 152 test subjects (among the three cross validation folds), XGB achieved the lowest number of incorrect predictions (27 subjects). Samples of misclassified frames are shown in Fig. 4.

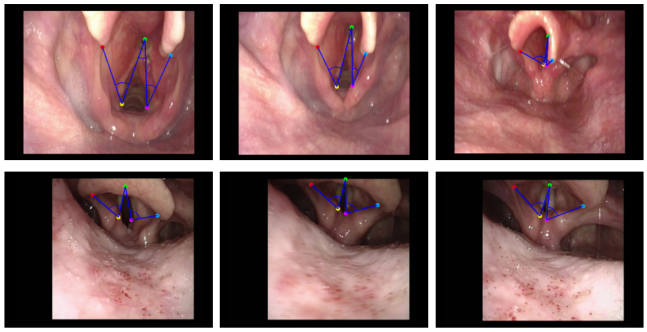

Fig. 4: Visual samples of misclassified frames. The images in the first row were erroneously predicted as belonging to the fixation class, while the images in the second row were erroneously predicted as belonging to the preserved motility class. In the latter case, the vocal folds area occupies a small portion of the frame, which makes the prediction more challenging.

IV. DISCUSSION

The main objective of this study was to evaluate the ability of ML algorithms to discriminate between vocal cords preserved motility and fixation. To do so, we extracted a number of relevant features from triplets of videoendoscopic frames, representing specific vocal folds positions. The extracted features were used to train and test four different classifiers, which showed good results, and the best-performing resulted to be the XGB. Even though the results of this model do not depart from the others, the use of this specific ML classifier could be useful in case of some not labeled keypoints, as it is able to handle missing values [17]. From the results, it is also possible to appreciate the ability of all the tested models in assessing vocal cords motility. This is an expected behavior [18], [19] and confirms that the application on ML may have a positive impact to assist clinicians in their practice.

To the best of our knowledge, this is the first study to rely on keypoints to evaluate vocal cords motility. Previous work in literature, in fact, focused on the segmentation of the glottis to evaluate the motility. The advantage of relying on keypoints, as already demonstrated in precedent work from

TABLE I: Performance evaluation metrics. Precision (Prec), recall (Rec), F1-score (F1), accuracy (Acc), average precision (AP), and area under the precision-recall curve (AUC) are reported. For each classifier, the first row refers to the class fixation, while the second to the class preserved motility.

Classifier	Prec	Rec	F1	Acc	AP	AUC
SVC	0.73	0.73	0.73	0.82	0.75	0.73
	0.86	0.86	0.86		0.90	0.90
SVM	0.71	0.76	0.74	0.82	0.72	0.71
	0.88	0.84	0.86		0.93	0.92
RF	0.67	0.59	0.62	0.76	0.64	0.63
	0.80	0.85	0.83		0.89	0.89
XGB	0.76	0.69	0.72	0.82	0.76	0.76
	0.85	0.89	0.87		0.94	0.93

Fig. 5: Visual samples of frames from the used dataset. It is characterized by high variability among the frames, which reflects also on the variability of the features used to train the models.

other fields [20], [21], is the possibility to obtain a direct classification. Methods relying on segmentation, in fact, need a post-processing step to obtain a diagnosis.

A limitation of the proposed work could be seen in the relatively limited size of the dataset, which is due to the time needed to label each frame, and to the lack of available annotated dataset online. The time consuming annotation procedure also makes it difficult, at the moment, to evaluate intra-observer variability. Moreover, the dataset used in this work includes frames with very high variability among each other, as shown in Fig. 5, which is typical of videoendoscopic frames. This characteristic of the dataset reflects also on the extracted features and on the achieved results. For this reason, adding the classification algorithm downstream of a frame selection process might improve the results.

As future work, to support clinicians in the actual clinical practice, the classification model could be included within other computer assisted algorithms for diagnostic support, e.g., frames selection and automatic keypoints regression.

V. CONCLUSION

Vocal folds fixation is typically assessed by visually evaluating videoendoscopic frames. This process is time consuming and requires an expert eye. To make the evaluation more objective, in this paper we compared four ML models to classify vocal cords motility into two classes: preserved motility and fixation. The best-performing model, XGB, proved to be a useful tool to investigate vocal cords motility in a more objective and reliable way. It is, in fact, able to distinguish between the two classes, which makes it a potential tool to support clinicians in the clinical practice.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] I. Domingues, G. Pereira, P. Martins, H. Duarte, J. Santos, and P. H. Abreu, "Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4093–4160, 2020.

[3] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, "A review on deep-learning algorithms for fetal ultrasound-image analysis," *Medical Image Analysis*, 2022.

[4] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 939–954, 2019.

[5] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni *et al.*, "Surgical data science–from concepts toward clinical translation," *Medical Image Analysis*, vol. 76, p. 102306, 2022.

[6] A. Repici, M. Badalamenti, R. Maselli, L. Correale, F. Radaelli, E. Rondonotti, E. Ferrara, M. Spadaccini, A. Alkandari, A. Fugazza *et al.*, "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial," *Gastroenterology*, vol. 159, no. 2, pp. 512–520, 2020.

[7] A. Paderno, F. Gennaini, A. Sordi, C. Montenegro, D. Lancini, F. P. Villani, S. Moccia, and C. Piazza, "Artificial intelligence in clinical endoscopy: insights in the field of videomics," *Frontiers in Surgery*, p. 1361, 2022.

[8] J. K. R. Menon, R. M. Nair, and S. Priyanka, "Unilateral vocal fold paralysis: can laryngoscopy predict recovery? a prospective study," *The Journal of Laryngology Otology*, vol. 128, no. 12, pp. 1095–1104, 2014.

[9] D. Voigt, M. Döllinger, A. Yang, U. Eysholdt, and J. Lohscheller, "Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 3, pp. 275–288, 2010.

[10] S. Moccia, E. De Momi, M. Guarnaschelli, M. Savazzi, A. Laborai, L. Guastini, G. Peretti, and L. S. Mattos, "Confident texture-based laryngeal tissue classification for early stage diagnosis support," *Journal of Medical Imaging*, vol. 4, no. 3, p. 034502, 2017.

[11] G. Andrade-Miranda, Y. Stylianou, D. D. Deliyski, J. I. Godino-Llorente, and N. Henrich Bernardoni, "Laryngeal image processing of vocal folds motion," *Applied Sciences*, vol. 10, no. 5, 2020.

[12] A. S. Hamad, M. M. Haney, T. E. Lever, and F. Bunyak, "Automated segmentation of the vocal folds in laryngeal endoscopy videos using deep convolutional regression networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 140–148, 2019.

[13] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, and M. Naghibolhosseini, "Deep-learning-based representation of vocal fold dynamics in adductor spasmodic dysphonia during connected speech in high-speed videoendoscopy," *Journal of Voice*, 2022.

[14] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, A. de Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "A deep learning approach for quantifying vocal fold dynamics during connected speech using laryngeal high-speed videoendoscopy," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 6, p. 2098–2113, 2022.

[15] C. T. Herbst, J. Unger, H. Herzel, J. G. Švec, and J. Lohscheller, "Phasegram analysis of vocal fold vibration documented with laryngeal high-speed video endoscopy," *Journal of Voice*, vol. 30, no. 6, 2016.

[16] J. Lohscheller, "Towards evidence based diagnosis of voice disorders using phonovibrograms," in *2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, 2009, pp. 1–4.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," 2016, pp. 785–794.

[18] V. Singh, M. K. Gourisaria, and H. Das, "Performance analysis of machine learning algorithms for prediction of liver disease," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 2021, pp. 1–7.

[19] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2021, pp. 654–659.

[20] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.

[21] S. Moccia, L. Migliorelli, V. Carnielli, and E. Frontoni, "Preterm infants' pose estimation with spatio-temporal features," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 8, pp. 2370–2380, 2020.