Turkish Archives of Otorhinolaryngology

Original Investigation

Turk Arch Otorhinolaryngol
DOI: 10.4274/tao.2026.2026-6-14

Machine Learning-Based Multiclass Classification of Benign Vocal Fold Lesions Using Acoustic, Clinical, and Patient-Reported Features

Ebru Karakaya Gojayev[1], Zahide Çiler Büyükatalay[2], Arjin Öksüz[3], Gürsel Dursun[2]

[1]University of Health Sciences Türkiye, Ankara Etlik City Hospital, Clinic of Otorhinolaryngology, Ankara, Türkiye
[2]Ankara University Faculty of Medicine, Department of Otorhinolaryngology, Ankara, Türkiye
[3]Artvin Hopa State Hospital, Clinic of Otorhinolaryngology, Artvin, Türkiye

Abstract

Objective: To develop and evaluate a multiclass random forest classifier for distinguishing among five types of benign vocal fold lesions using acoustic and phonatory parameters, patient-reported outcomes (PROs), demographic characteristics, and occupational voice demands.

Methods: This retrospective study included 217 adults with confirmed diagnoses of vocal fold nodules, vocal fold polyps, intracordal cysts, Reinke's edema, or sulcus vocalis at a tertiary otolaryngology center. Features included five acoustic parameters (fundamental frequency, jitter, shimmer, maximum phonation time, s/z ratio), three PRO instruments, the voice handicap index, voice-related quality of life, and reflux symptom index (RSI), and demographic and occupational variables. A class-weighted random forest classifier was evaluated using stratified five-fold cross-validation with bootstrap-derived 95% confidence intervals; multinomial logistic regression and linear support vector machine served as comparators.

Results: The random forest classifier outperformed the comparator models, achieving accuracy of 0.576, balanced accuracy of 0.476, macro-F1 of 0.479, and a macro-averaged area under the receiver operating characteristic curve of 0.76. The highest class-level F1 scores were observed for sulcus vocalis (0.703) and vocal fold polyps (0.625); intracordal cysts showed the lowest recall (0.111), reflecting acoustic overlap with vocal fold polyps. Top 2 accuracy was approximately 74%. Fundamental frequency, RSI, and maximum phonation time were the highest-ranked predictors.

Conclusion: A multidimensional random forest framework incorporating acoustic measures, PROs, and occupational voice demand demonstrated potential for classifying benign vocal fold lesions. Acoustic parameters alone may not adequately distinguish among these lesions, and the inclusion of patient-reported data may improve model performance. A top 2 accuracy of 74% supports utility as a pre-endoscopic decision-support tool.

Keywords: Dysphonia, machine learning, patient-reported outcome measures, random forest, vocal cords, voice disorders

ORCID IDs of the authors:
E.K.G. 0000-0003-4749-3511
Z.Ç.B. 0000-0002-0992-0079
A.Ö. 0000-0002-9213-1627
G.D. 0000-0003-2542-8268

Cite this article as: *Karakaya Gojayev E, Çiler Büyükatalay Z, Öksüz A, Dursun G. Machine learning-based multiclass classification of benign vocal fold lesions integrating acoustic and patient-reported features. Turk Arch Otorhinolaryngol. [Epub Ahead of Print]*

Corresponding Author:
Ebru Karakaya Gojayev, MD;
ebru_karakaya92@hotmail.com

Received Date: 11.06.2026
Accepted Date: 03.08.2026
Epub: 25.09.2026

Introduction

Voice is a fundamental component of communication and social interaction, and its impairment significantly affects an individual's quality of life and functional capacity. Although dysphonia has a prevalence of approximately 1% in treatment-seeking populations, population-based studies have shown that the lifetime risk of voice disorders may reach up to 30%, with higher prevalence observed among women and older adults. Benign vocal fold lesions, including vocal nodules, polyps, Reinke's edema, intracordal cysts, and sulcus vocalis, are among the most common causes of dysphonia and represent a substantial clinical burden. Limited awareness of dysphonia may lead to hoarseness being perceived as a transient condition, resulting in diagnostic delays (1).

In current clinical practice, the diagnosis of benign vocal fold lesions relies primarily on laryngoscopic or videolaryngostroboscopic evaluation. Although these minimally invasive office-based procedures provide the diagnostic reference standard, they require specialized equipment and expertise and are therefore not well suited for large-scale screening, highlighting the need for complementary pre-endoscopic assessment approaches. Acoustic voice analysis has long been used for objective evaluation, and parameters such as fundamental frequency (f0), jitter, shimmer, maximum phonation time (MPT), and the s/z ratio have been associated with various laryngeal pathologies. However, the isolated evaluation of individual acoustic parameters provides limited discriminatory power, underscoring the need for multivariate analytical approaches (2,3).

In recent years, machine learning-based methods have been increasingly applied to the detection and classification of voice disorders. Among these, random forest is an ensemble-based algorithm that combines multiple decision trees, offering robustness against overfitting and the ability to model complex, non-linear relationships (4). Most previous studies have focused on binary classification (pathological versus healthy voice), while multiclass differentiation of benign vocal fold lesions remains underexplored. Moreover, existing models rely predominantly on acoustic features and insufficiently incorporate clinical variables reflecting the functional and demographic context of voice disorders.

Voice disorders cannot be fully characterized by acoustic measurements alone. Patient-reported outcomes (PROs) measures, such as the voice handicap index (VHI), voice-related quality of life (VRQoL), and reflux symptom index (RSI), provide important information regarding functional impact and disease burden. Occupational voice demand classified as high, moderate, or low based on daily vocal use has also been shown to be associated with clinical risk among professional voice users (5-7).

The aim of this study is to develop a random forest-based multiclass machine learning model integrating acoustic parameters, PROs, demographic characteristics, and occupational voice demand, and to evaluate its diagnostic performance in distinguishing benign vocal fold lesions.

Methods

Study Design and Participants

This retrospective observational study was conducted using routinely collected clinical data from adult patients presenting with dysphonia to the Otolaryngology Clinic at Ankara University Faculty of Medicine. Patients evaluated between January 2010 and January 2020 were eligible for inclusion. All patients underwent comprehensive voice assessment and videolaryngostroboscopic examination as part of the routine diagnostic protocol. The reference diagnosis for each patient was established independently by two experienced otolaryngologists based on videolaryngostroboscopic findings. These diagnoses served as the ground-truth labels used for machine learning model development and evaluation.

Patients diagnosed with one of the five most common benign vocal fold lesions were included: vocal fold nodules, vocal fold polyps, intracordal cysts, Reinke's edema, and sulcus vocalis. Patients with rare, malignant, neurological, or inflammatory laryngeal disorders were excluded to ensure adequate sample size per class for multiclass modeling. The final analytic dataset consisted of 217 patients.

Data Collection and Variables

Demographic variables included age, sex (binary: female/male), and smoking status (current smoker versus non-smoker). Occupational voice demand was categorized as high, moderate, or low according to daily vocal load (5-7). High-demand occupations included teachers, singers, imams, call-center workers, and actors; low-demand occupations included homemakers, retirees, and students; all others were categorized as moderate.

PRO instruments included the VHI, RSI, and VRQoL, all administered as part of routine clinical evaluation.

Acoustic Voice Recording and Analysis

Voice recordings were obtained under standardized conditions. Patients were instructed to produce a sustained vowel /a/ for approximately 6 seconds in modal voice at comfortable pitch and loudness. Recordings were performed using a condenser microphone (EM-616, Shure, Chicago, IL) positioned at 20 cm from the mouth in a quiet examination room; ambient noise levels were maintained below 45 dB sound pressure level. Acoustic analysis was conducted using commercially available software (Vocal Assessment, Dr. Speech, Tigers Inc., Seattle, WA). Parameters extracted included fundamental frequency (f0; Hz), jitter (%), shimmer (dB), and s/z ratio. MPT was measured separately as the duration of sustained /a/ phonation following maximum inspiration. All endoscopic diagnoses were made by two experienced otolaryngologists and cases with diagnostic disagreement were resolved by consensus.

Statistical Analysis

All analyses were performed using Python (version 3.10) with the scikit-learn, NumPy, and SciPy libraries. Continuous variables are presented as mean ± standard deviation, and categorical variables as counts and percentages.

The outcome variable was benign lesion category (five classes). The 12 input features were age, sex, smoking status, occupational voice demand, VHI, RSI, VRQoL, f0, MPT, jitter, shimmer, and s/z ratio. Physiologically implausible f0 values were excluded (range: 60-500 Hz). Missing values were handled using median imputation for continuous and most-frequent imputation for categorical variables. All preprocessing steps were implemented within the cross-validation pipeline to prevent data leakage.

Multiclass classification was performed using a class-weighted random forest classifier (n_estimators=500, max_features='sqrt', class_weight='balanced'; all remaining parameters were kept at their scikit-learn default settings) to account for class imbalance. The use of 500 decision trees was a deliberate modeling choice to improve the stability of ensemble predictions and feature importance estimates while reducing model variance. No formal hyperparameter optimization was performed because the primary objective was to evaluate a robust and clinically interpretable baseline model rather than maximize predictive performance. Multinomial logistic regression and linear support vector machine models served as comparators under identical preprocessing and validation conditions. Model performance was evaluated using stratified five-fold cross-validation. Overall performance metrics included accuracy, balanced accuracy, macro-averaged F1-score, and macro one-vs-rest receiver operating characteristic (ROC)-area under the curve (AUC). Non-parametric bootstrap resampling (1,000 iterations) was performed to derive 95% confidence intervals. Top-2 accuracy was additionally calculated. Feature importance was estimated using normalized impurity-based scores and complemented by permutation importance analysis performed within the cross-validation loop to prevent data leakage.

This study was approved by the Ankara University Human Research Ethics Committee (approval no: İ04-493-26, date: 12.05.2026) and conducted in accordance with the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective study design.

Results

A total of 217 patients were included. Their mean age was 43.2±12.6 years, with a slight predominance of female patients (55.3%). Current smoking was present in 56.7%. Occupational voice demand was classified as high in 12.4%, moderate in 42.9%, and low in 44.7%. Vocal fold polyps (33.6%) and sulcus vocalis (23.9%) were the most common diagnoses, followed by Reinke's edema (23.0%), vocal fold nodules (11.1%), and intracordal cysts (8.3%) (Table 1).

The effect of occupational voice demand on model performance is presented in Table 2. Adding occupational voice demand increased overall accuracy slightly (0.567 to 0.576), while balanced accuracy remained unchanged (0.476) and macro-F1 decreased marginally (0.486 to 0.479). Overlapping bootstrap confidence intervals indicated no statistically meaningful global improvement, suggesting a lesion-specific rather than global effect.

Comparative model performance is summarized in Table 3. Random forest achieved the highest accuracy (0.576), balanced accuracy (0.476), macro-F1 (0.479), and macro-ROC-AUC (0.76) among evaluated algorithms. Top 2

Table 1. Demographic and diagnostic characteristics of the study cohort (n=217)

Variable	Value
Age, mean ± SD (years)	43.2±12.6
Sex, female	120 (55.3%)
Sex, male	97 (44.7%)
Current smoker	123 (56.7%)
High voice demand occupation	27 (12.4%)
Moderate voice demand occupation	93 (42.9%)
Low voice demand occupation	97 (44.7%)
Vocal fold nodule	24 (11.1%)
Vocal fold polyp	73 (33.6%)
Intracordal cyst	18 (8.3%)
Reinke's edema	50 (23.0%)
Sulcus vocalis	52 (23.9%)

Values are presented as mean ± SD for continuous variables and n (%) for categorical variables. SD: Standard deviation

Table 2. Effect of occupational voice demand on random forest performance

Model	Accuracy	Balanced accuracy	Macro-F1
Random forest (without occupational voice demand)	0.567	0.476	0.486
Random forest (with occupational voice demand)	0.576	0.476	0.479

Values represent stratified five-fold cross-validation performance; 95% CIs estimated by bootstrap resampling. CIs: Confidence intervals accuracy reached approximately 74%, indicating that the correct diagnosis was captured within the two highest predicted probabilities in nearly three of four cases.

Class-specific performance is presented in Table 4. The highest F1-scores were observed for sulcus vocalis (0.703) and vocal fold polyps (0.625). Reinke's edema demonstrated intermediate performance (F1: 0.545). Vocal fold nodules were more challenging to classify (F1: 0.368), and intracordal cysts showed the lowest recall (0.111) and F1-score (0.154), reflecting substantial acoustic overlap with polyps (Figure 1). Occupational voice demand improved recall for polyps (0.671 to 0.685) and Reinke's edema (0.500 to 0.540). In contrast, recall for intracordal cysts decreased slightly after inclusion of occupational voice demand. This finding is likely attributable to the small number of cyst cases (n=18) and the relatively weak association between occupational voice demand and intracordal cysts, which may have introduced additional variability into this subgroup.

Feature importance analysis is presented in Table 5. Fundamental frequency (f0; 0.120), RSI (0.119), and MPT (0.105) were the leading predictors. VRQoL (0.097) and VHI (0.097) also contributed substantially. Permutation-based importance yielded a comparable ranking, supporting predictor robustness.

Discussion

The random forest model developed in this study offers a multidimensional approach to the classification of benign vocal fold lesions, extending beyond acoustic-only frameworks by integrating PROs and contextual clinical variables. Although overall balanced accuracy (0.476) and macro-F1 (0.479) indicated moderate class-wise discrimination, reflecting the inherent acoustic overlap among benign lesion types, the macro-ROC-AUC of 0.76 suggests meaningful probabilistic ranking ability across lesion categories. In multiclass clinical settings, ROC-AUC and top-2 accuracy is arguably more informative than strict top 1 accuracy, as the clinical objective is not definitive pre-endoscopic diagnosis but rather probability-guided triage

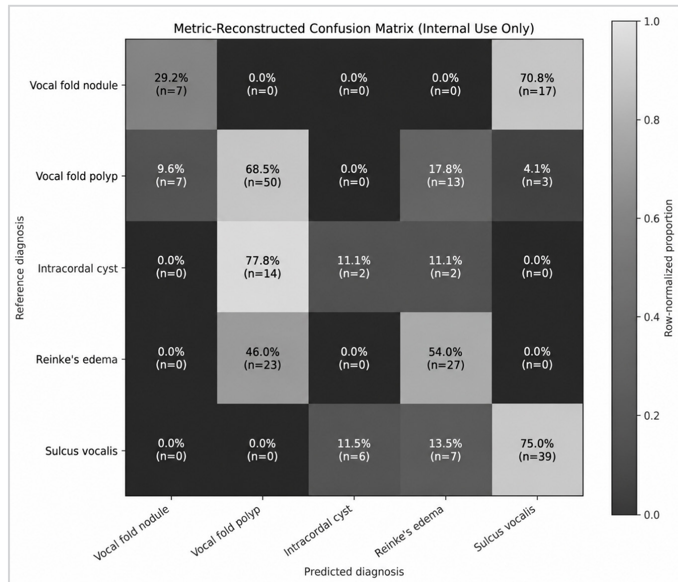

Figure 1. Normalized confusion matrix of the final random forest classifier evaluated using stratified five-fold cross-validation. Rows represent the reference diagnoses and columns represent the predicted diagnoses. Each cell displays the row-normalized proportion together with the corresponding number of patients assigned to each diagnostic category

Table 3. Comparison of classification algorithms (with occupational voice demand)

Model	Accuracy	Balanced acc	Macro-F1	Macro ROC-AUC
Logistic regression	0.51	0.41	0.39	0.72
Linear SVM	0.53	0.42	0.41	0.74
Random forest	0.576	0.476	0.479	0.76

Metrics derived from stratified five-fold cross-validation. Macro ROC-AUC calculated using a one-vs-rest approach. ROC-AUC: Receiver operating characteristic-area under the curve, SVM: Support vector machine

Table 4. Class-specific performance of the final random forest model

Diagnosis	Precision	Recall (without occ.)	Recall (with occ.)	F1-score
Vocal fold nodule	0.500	0.292	0.292	0.368
Vocal fold polyp	0.575	0.671	0.685	0.625
Intracordal cyst	0.250	0.167	0.111	0.154
Reinke's edema	0.551	0.500	0.540	0.545
Sulcus vocalis	0.661	0.750	0.750	0.703

Values derived from stratified five-fold cross-validation of the class-weighted random forest model. Occ.: Occupational voice demand

Table 5. Feature importance ranking (random forest model)

Predictor	Relative importance
Fundamental frequency (f0)	0.120
Reflux symptom index	0.119
Maximum phonation time	0.105
Voice-related quality of life	0.097
Voice handicap index	0.097
Shimmer	0.091
s/z ratio	0.091
Jitter	0.072
Age	0.065
Occupational voice demand	0.052
Smoking	0.051
Sex	0.041

Normalized impurity-based (mean decrease in impurity) scores; values sum to 1.0

that can narrow the differential before laryngoscopy. A model with balanced accuracy below 0.50 in a five-class problem should not be interpreted as performing at chance level; rather, class-weighted training and the inherent difficulty of acoustically overlapping benign lesions set a naturally lower performance ceiling than binary classification tasks.

Most previous machine learning studies in voice pathology have focused on binary classification, primarily distinguishing healthy voices from pathological ones, rather than differentiating among multiple benign vocal fold lesions. Consequently, the high accuracy values reported in these studies (approximately 86-96%) are not directly comparable with the presented study because binary classification represents a substantially less complex task than multiclass classification involving five diagnostically overlapping lesion categories (2,3). In contrast, our model addressed the clinically more challenging problem of discriminating among vocal fold nodules, polyps, intracordal cysts, Reinke's edema, and sulcus vocalis, all of which may exhibit similar acoustic characteristics. Therefore, the moderate overall classification performance observed in the presented study should be interpreted within the context of this considerably greater classification complexity rather than compared directly with binary classification studies. This distinction is clinically important because the aim of our model is not simply to distinguish healthy voices from pathological ones, but to support the pre-endoscopic differential diagnosis of benign vocal fold lesions encountered in routine laryngology practice.

The high importance scores of VRQoL, RSI, and VHI in the model indicate that PROs capture diagnostically informative dimensions beyond what acoustic parameters alone can provide (8,9). Notably, RSI emerged as the second most important predictor overall. This finding is biologically plausible: laryngopharyngeal reflux is differentially associated with specific lesion types, particularly Reinke's edema and vocal fold polyps, and RSI scores may therefore carry lesion-specific signal beyond generic voice handicap measures (6). The superior classification of sulcus vocalis, reflected in the highest F1-score, is consistent with the distinctive effects of this pathology on phonation efficiency and patient-perceived voice quality. The lesion-specific recall improvements attributed to occupational voice demand support prior findings associating high vocal load with phonotraumatic lesion development among teachers and professional voice users (5-7).

Random forest's performance advantage over linear comparators reflects its capacity to model non-linear interactions among acoustic, patient-reported, and demographic variables (4). The top 2 accuracy of approximately 74% is particularly relevant clinically: by narrowing the differential diagnosis to two likely categories in nearly three of four cases, the model may support pre-endoscopic patient counseling and prioritization of specialist referral.

Importantly, the proposed model is not intended to establish an autonomous diagnosis or replace comprehensive clinical evaluation. Rather, it is designed as a clinical decision-support tool that complements detailed history taking, physical examination, videolaryngostroboscopic assessment, and physician expertise. The final diagnosis should always be established through comprehensive otolaryngological evaluation.

The relatively low proportion of vocal fold nodules compared with sulcus vocalis likely reflects the referral pattern of our university-based tertiary laryngology clinic. Patients with uncomplicated phonotraumatic nodules are frequently diagnosed and managed conservatively in primary or secondary care, whereas patients with persistent dysphonia, diagnostic uncertainty, recurrent symptoms, or suspected structural vocal fold abnormalities are more likely to be referred to our center. Consequently, sulcus vocalis and other structurally complex lesions may have been relatively overrepresented in our study cohort.

Study Limitations

This study has several limitations. The retrospective single-center design may limit generalizability, although standardized protocols and routine clinical data enhance ecological validity. Model performance was assessed using internal cross-validation without external validation. Therefore, external validation using independent patient cohorts from different institutions, recording devices, and recording environments is essential before clinical implementation. Such validation will help determine the generalizability, robustness, and transportability of the proposed model across diverse clinical settings. Notably, the intracordal cyst subgroup consisted of only 18 patients (8.3%), resulting in fewer than four cases per cross-validation fold on average; class-specific metrics for this group should therefore be interpreted with caution, as they carry substantial statistical uncertainty. The analysis was restricted to five common benign lesions and did not encompass rare or malignant pathologies. A normalized confusion matrix illustrating class-specific prediction patterns is provided as Figure 1. Finally, acoustic analysis was based on sustained vowel phonation, which may not fully represent voice behavior during connected speech.

Conclusion

This study demonstrates that a multidimensional random forest model integrating acoustic biomarkers, PROs, demographic variables, and occupational voice demand provides a more clinically comprehensive framework for benign vocal fold lesion classification than acoustic-only approaches. Although class-wise performance was moderate, lesion-specific improvements and a top 2 accuracy of 74% indicate that the proposed framework has potential as an adjunctive pre-endoscopic clinical decision-support tool. Future prospective multicenter validation studies are required before routine clinical implementation. Importantly, the proposed model should be regarded as a complementary clinical decision-support tool rather than a replacement for comprehensive otolaryngological evaluation, including detailed history taking, physical examination, videolaryngostroboscopic assessment, and physician judgment.

Ethics

Ethics Committee Approval: This study was approved by the Ankara University Human Research Ethics Committee (approval no: İ04-493-26, date: 12.05.2026).

Informed Consent: The requirement for informed consent was waived because of the retrospective study design.

Footnotes

Authorship Contributions

Surgical and Medical Practices: E.K.G., Z.Ç.B., Concept: E.K.G., Z.Ç.B., Design: E.K.G., Data Collection and/or Processing: E.K.G., A.Ö., Analysis or Interpretation: E.K.G., A.Ö., G.D., Literature Search: E.K.G., Writing: E.K.G., Z.Ç.B., G.D.

Conflict of Interest: The authors declare that they have no conflict of interest.

Financial Disclosure: The authors declare that this study has received no financial support.

Main Points

- Distinguishing benign vocal fold lesions before laryngoscopy remains a clinical challenge; machine learning offers a promising diagnostic support approach.
- A random forest classifier trained on acoustic parameters combined with patient-reported outcomes (voice handicap index, voice-related quality of life, reflux symptom index) achieved a macro-receiver operating characteristic-area under the curve of 0.76 and top 2 accuracy of ~74% in 217 dysphonic patients.
- Integration of patient-reported data significantly improved classification performance over acoustic features alone, particularly for sulcus vocalis and Reinke's edema.
- This model may serve as a practical pre-laryngoscopic triage tool in tertiary otolaryngology settings.

References

1. Stachler RJ, Francis DO, Schwartz SR, Damask CC, Digoy GP, Krouse HJ, et al. Clinical practice guideline: hoarseness (Dysphonia) (Update). Otolaryngol Head Neck Surg. 2018; 158(1 Suppl): S1-42. Erratum in: Otolaryngol Head Neck Surg. 2018; 159: 403. [Crossref]
2. Godino-Llorente JI, Gómez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Trans Biomed Eng. 2004; 51: 380-4. [Crossref]
3. Verde L, De Pietro G, Sannino G. Voice disorder identification by using machine learning techniques. IEEE Access. 2018; 6: 16246-55. [Crossref]
4. Breiman L. Random forests. Mach Learn. 2001; 45: 5-32. [Crossref]
5. Paniagua MS, Pérez CJ, Calle-Alonso F, Salazar C. An acoustic-signal-based preventive program for university lecturers' vocal health. J Voice. 2020; 34: 88-99. [Crossref]
6. Koufman JA, Isaacson G. The spectrum of vocal dysfunction. Otolaryngol Clin North Am. 1991; 24: 985-8. [Crossref]
7. Vilkman E. Occupational safety and health aspects of voice and speech professions. Folia Phoniatr Logop. 2004; 56: 220-53. [Crossref]
8. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, et al. The voice handicap index (VHI): development and validation. Am J Speech Lang Pathol. 1997; 6: 66-70. [Crossref]
9. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). J Voice. 1999; 13: 557-69. [Crossref]