A Feasibility Study of Explainable Machine Learning on Small-Scale Postoperative Voice Data

Noura Haddou[1], Najlae Idrissi[2], Sofia Ben Jebara[3]
Faculty of Science and Technology-Intelligence Data and Computing Team,
Sultan Moulay Slimane University, Beni Mellal, Morocco[1,2]
Research Laboratory COSIM-Higher School of Communications of Tunis, University of Carthage, Ariana, 2088, Tunisia[3]

***Abstract*—Voice dysfunction is a common complication following thyroid surgery. However, the application of explainable machine learning for predicting postoperative voice recovery remains largely unexplored. Therefore, an investigation was done to examine voice recovery based on acoustic, objective, and glottal features. Voice recordings were collected from female patients before surgery and one month after surgery. Acoustic and glottal parameters, including Quasi Open Quotient, Speed Quotient, age, and others, were automatically extracted from the recordings. Random Forest, Support Vector Machines, and Logistic Regression with Sequential Feature Selection were applied to examine model behavior and identify feature importance. Model stability and interpretability were evaluated across cross-validation folds. Performance metrics varied over folds, highlighting the exploratory and statistically fragile nature of predictions in small datasets. SHAP (SHapley Additive exPlanations) analysis revealed variability in feature contributions, emphasizing the need for cautious interpretation and detailed methodological reporting. Our findings provide preliminary guidance for applying explainable machine learning to small biomedical datasets. They demonstrate the importance of careful methodological design.**

***Keywords*—XAI; explainable AI; SHAP; glottal features; SVM; thyroidectomy; voice recovery**

I. Introduction

Thyroidectomy is a surgical intervention performed to remove a portion or the entire thyroid gland. It can be performed to treat malignant or benign thyroid diseases [1]. This surgery is mostly followed by a change in voice quality, reported in approximately 30% to 80% of patients [2], [3]. This change is described as a disorder characterized by an impaired pitch, volume, vocal effort, or voice quality, which alters communication or voice-related quality of life [4].

To detect pathological voices, both subjective and objective assessment methods are commonly used [5], [6]. The subjective voice assessment method includes an auditory perceptual assessment and visual examination of the vocal cords in the hospital [5], [7]. However, the diagnosis based on this type of evaluation is difficult because it relies on the experience of doctors, and there is also a high possibility of clinical errors [8]. In contrast, the objective assessment is based on non-invasive computer analysis of acoustic signals to identify pathological voice, which may not even be audible to a human being [5].

Despite these advances, speech pathology remains an underexplored area, and the field has not yet received significant attention from researchers. This creates a significant need for further research and improved diagnostic tools for voice-related disorders [9]. Thyroid cancer is one of these diseases. It is the second most common cancer for both men and women in the 15 to 34 age group and the most common cancer in women [2], [10].

Regrettably, most existing works have focused on voice analysis or classification tasks. In particular, research has aimed to identify acoustic parameters associated with voice changes post-surgery and to develop computer-based diagnostic systems using a combination of these features. These classification techniques have been widely used to build automated detection systems. However, they are not commonly applied to predict voice quality or recovery following thyroidectomy.

In this pilot research, we propose a ML-based system that uses voice recordings from two stages: pre-operative and one month postoperative. Objective, acoustic and glottal features are used to train multiple classifiers. We also apply Sequential feature selection (SFS) for feature selection to improve prediction performance. Finally, we integrate SHapley Additive exPlanations (SHAP) to interpret the decision-making process of the best-performing model. The aim was to address a binary classification problem to differentiate between recovery and non-recovery voice. The main contributions of this research are as follows:

- We implement a leakage-safe, group-stratified machine learning pipeline suitable for extremely small biomedical voice datasets.
- We demonstrate the use of group-stratified nested cross-validation combined with sequential feature selection to reduce evaluation bias in longitudinal voice data.
- We provide an empirical analysis of model instability and performance variance under small-sample regimes using bootstrapped confidence intervals and learning curves.
- We investigate the limitations and reliability of SHAP-based explainability when applied to small-scale tabular datasets.

II. Related Work

Recently, scientists focused on building several systems for detecting pathological voices using machine learning algorithms. These approaches can be classified into two primary techniques: traditional pipeline methods and end-to-end techniques [11]. End-to-end approaches utilize the original raw

voice or transformed speech, such as the spectrogram in its 2D form as input for the classification model, and employ deep learning techniques to generate target labels [12].

On the other hand, the traditional pipeline approaches involve two stages: firstly, the extraction of relevant parameters from a speech signal, and secondly, using these extracted features to train the model and classify the labeled output according to the intended purpose (normal/pathology voice, type of pathology, etc.) [12]. Most studies have used a set of objective and subjective parameters to realize these automatic systems. These measurements contain jitter, glottal parameters [12], etc.

To address this, some researchers have proposed new detection frameworks using ML or deep learning algorithms to forecast voice recovery after thyroid surgery. For example, Lee *et al.* [13] published a study to predict patient voice recovery after three months of postoperative spectrograms with a deep neural network algorithm. They initially proposed an approach that employed pre-operative and first postoperative patients' voice recordings and GRBAS scores to predict the patient's voice GRBAS scores after three months post-surgery. The utilization of GRBAS scores primarily aimed to characterize vocal quality by evaluating the degree of voice impairment during each session. In their study, 114 patients were followed. All patients who have undergone surgery for thyroid cancer. The participants were taken before surgery, two weeks, and three months after the operation. The proposed method has shown that long-term voice disorders can be predicted using pre- and postoperative spectrograms [13].

Another study by Kurt *et al.* [14], focused on developing a detection system using ML techniques to determine patients at risk of developing vocal cord palsy before thyroidectomy [14]. They included a sample of 1039 patients who underwent thyroidectomy in their research. The authors used a set of variables based on clinician reports, which can divided into two categories: continuous variables such as age, tumor size, white blood cell count, etc., and categorical variables including sex of patient, type of resection, presence of comorbidity, and so on. They achieved an accuracy of 100%. However, although Lee *et al.* [13] used spectrograms and GRBAS scores (Grade, Roughness, Breathiness, Asthenia, Strain) in their deep learning approach, this method has not yet been evaluated with traditional machine learning techniques or direct acoustic and glottal parameters. Most existing studies focus on evaluating voice quality before and after surgery using statistical tests [15], [16], [17].

In terms of classification, the majority aim to develop systems that distinguish between pathological and normal voices. To our knowledge, no study has yet explored ML-based prediction of post-thyroidectomy voice recovery using glottal, objective, and acoustic features with SHAP explainability.

III. Materials and Methods

In this section, we outline the methodology and materials used in the present study. Fig. 1 presents the proposed framework for predicting voice recovery after thyroidectomy. The system comprises different steps, including data collection (longitudinal audio recordings), pre-processing, feature extraction, classification, and interpretability analysis using the SHAP method. Classification was performed using the SFS algorithm.

A. Data Collection

Twenty-five females (aged 17-66 years, SD = 14.44) were included in this study. They underwent either partial or total thyroidectomy for malignant or benign thyroid cancer. Among them, 12 underwent partial thyroidectomy, and 13 underwent total thyroidectomy. All participants underwent thyroidectomy for benign thyroid, except for five who underwent surgery for malignant thyroid. The subjects did not have any vocal cord damage or history of alcohol before thyroid surgery. Moreover, no treatments were administered to the patients between the first day and one month after surgery, except for those who experienced vocal fold paralysis on the first day. Voice therapy is recommended in cases of vocal cord paralysis. Thyroidectomy was performed by various surgeons using a conventional cervical method, without the assistance of intraoperative neuromonitoring (IONM) for the identification of either the external branch of the superior laryngeal nerve (EBSLN) or the recurrent laryngeal nerve (RLN). Each patient underwent a laryngeal endoscopic examination conducted by a senior physician or a trained resident during quiet phonation and breathing. Voice recordings were acquired using a microphone integrated into the endoscopy column, positioned approximately 20 cm from the patient's lips. Patients sustained the French vowel /i/ at pre-surgery, and 30 days post-surgery.

B. Data Pre-processing

The duration of audio recordings varied among participants because they were instructed to pronounce the vowel for as long as possible at a comfortable volume and pitch. To ensure consistency across samples, a 3-second segment was extracted from the middle of each recording. This segment length was chosen after initially extracting 3-second segments from the beginning, middle, and end of each recording. Statistical analysis using ANOVA showed no significant effect of segment position on the acoustic parameters. Therefore, the middle 3-second segment was selected, as it represents the most stable portion of the sustained vowel. Acoustic and glottal parameters were then extracted from this segment for each participant.

C. Feature Extraction

In this section, we present the features used in the current study. They are classified into two categories: acoustic features and glottal features. In this work, we used only the parameters extracted on the day before the operation and excluded the other features. Additionally, the VHI-10 parameter was excluded from the training process to avoid any data leakage.

1) Acoustic parameters: From the vocal samples, we extracted the mean values of the fundamental frequency (F0). F0 presents the vocal fold vibration frequency, measured in Hertz, which is defined by the duration of each period T_i. The study also includes information on age, type of operation (partial or total), the nature of the thyroid (malignant or benign), and the Voice Handicap Index (VHI-10). The VHI-10 is a metric employed to record the patient's perception of handicap or impairment due to a speech disease, where scores above 11 are defined as pathology and indicate vocal disability. To further

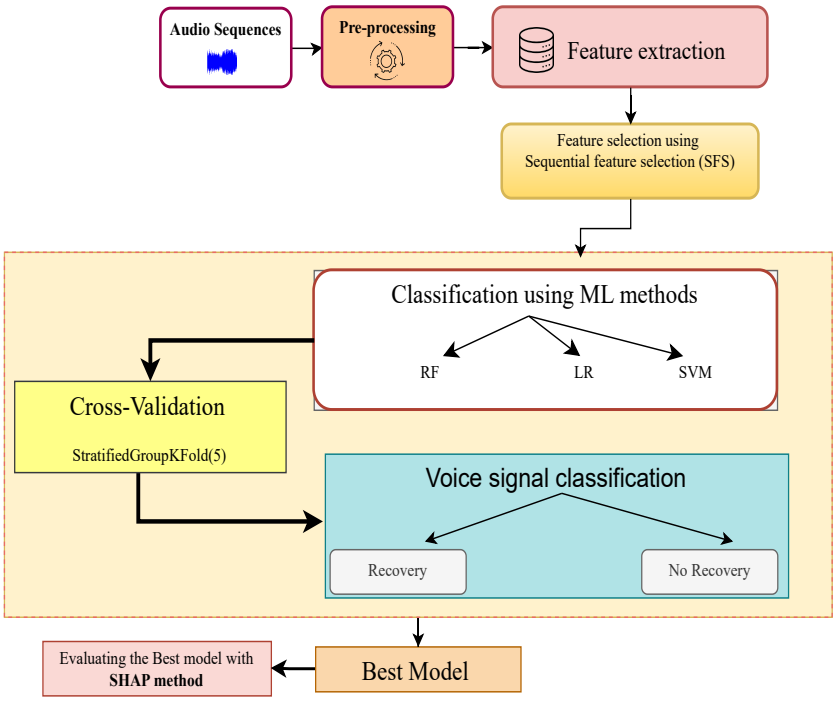


Fig. 1. Proposed method for predicting voice recovery stages following thyroidectomy.

improve the analysis, glottal flow features were used due to their demonstrated effectiveness in capturing the vocal effects of thyroidectomy [18].

2) Glottal parameters: In the literature, many techniques have been developed to obtain the glottal signal features. Among these methods, we find the Iterative Adaptive Inverse Filtering (IAIF) method [19], which estimates the glottal signal from the input speech signal through a two-step iterative process. Initially, the method estimates the glottal waveform and the vocal tract transfer function [18], [20]. It then refines these estimates to improve the performance of glottal signal estimation [18], [20].

In our work, we use the Pitch-Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) method [21] because of its ease of implementation. PSIAIF uses the IAIF method to extract glottal flow from the speech signal. Its main goal is to eliminate the influence of glottal excitation from the speech spectrum of the original signal [22]. The method proceeds by first estimating the vocal fold model using Linear Prediction Coefficients (LPC), and then obtaining the glottal flow through inverse filtering [22]. PSIAIF incorporates pitch-synchronous markers on the glottal waveform, which reduces the effect of formant ripple in comparison to IAIF [22]. This method has been implemented in the MATLAB toolbox Aparat [23]. Below, we provide a summary of the features that were extracted:

- The open quotient (OQ): is the ratio of the duration of the vocal fold open phase to the total duration of a glottal cycle [24].
- OQ1: Open quotient calculated from the primary opening of the glottal flow.
- OQ2: Open quotient calculated from the secondary opening of the glottal flow.
- The Speed Quotient (SQ): is the ratio of the duration of the opening phase to the duration of the closing phase [24]. **SQ1** and **SQ2** represent the speed quotients calculated from the primary and secondary openings, respectively.
- The normalized amplitude quotient (NAQ): is defined as the ratio of the amplitude quotient to the total duration of the glottal pulse [24].

To extract these features we used the methodology described in Alku (1992) [21]. The LPC analysis orders for the vocal tract and glottal source were set based on the sampling frequency, approximately $2 \times \text{round}(f_s/2000) + 4$ and $2 \times \text{round}(f_s/4000)$, respectively.

A leaky integration coefficient of 0.99 was used to model lip radiation. A single pass of a high-pass FIR filter (cutoff frequencies at 40 Hz stopband and 70 Hz passband) was applied prior to inverse filtering to remove low-frequency noise. The analysis was performed on frames windowed with a Hanning window. The audio signal was processed using its original sampling rate.

D. Classification Techniques

In the current study, a change in VHI-10 (ΔVHI_{10}) was calculated for each patient as:

$$\Delta \text{VHI}_{10} = \text{VHI}_{10,1\text{ month}} - \text{VHI}_{10,\text{pre-op}}$$

A negative ΔVHI_{10} (< 0) indicated improvement and was classified as *recovered*, while $\Delta \text{VHI}_{10} \geq 0$ was classified as *not recovered*.

Given the limited dataset size, we employed cross-validation with stratification to evaluate the models. Specifically, we used StratifiedGroupKFold with k = 5. This value was selected after testing k values from 2 to 7, with the best performance achieved at k = 5. To avoid data leakage, groups were defined by patient ID, ensuring that recordings from the same patient did not appear in both training and testing sets. The models were evaluated using a pipeline that applied MinMax scaling to the input features, which normalizes each feature by subtracting the minimum value and dividing by the range (max - min). The models were evaluated using the Sequential Feature Selection (SFS) method. To assess model stability and avoid data leakage, a nested cross-validation procedure was applied inside SFS during the feature selection phase. In addition, a bootstrapped validation with 10000 resamples was performed to calculate the corresponding confidence intervals (CIs) for model performance metrics.

For classification, various machine learning techniques, including SVM, RF, and Logistic Regression (LR) were used. The SVM was chosen as the representative algorithm due to its stability with limited samples. RF and LR were evaluated for comparison and to verify that performance differences were consistent; results are reported for completeness.

E. Hyperparameters

For hyperparameter tuning, GridSearchCV method was applied to each algorithm using the same StratifiedGroupKFold cross-validation process described above. For the SVM model, parameters such as the regularization parameter C (ranging from 1 to 50), kernel type (linear, RBF, polynomial, sigmoid), and gamma (auto or scale) were optimized. RF tuning included the number of estimators (1 to 40), maximum tree depth (1 to 51), and the criterion (Gini or entropy). The selected parameters were then employed in this study. The best hyperparameters were then used for each model. The final hyperparameters for each model were:

- **SVM:** $C = 1$, gamma = `scale`, kernel = `rbf`
- **LR:** max_iter = 1000
- **RF:** max_depth = 2, criterion=`gini`, n_estimators = 3

All analysis were conducted in Python with scikit-learn version 1.6.1, pandas, and matplotlib. The experiments were conducted on an HP laptop with an Intel Core i7-1165G7 processor (2.80 GHz) and 32 GB of RAM.

F. Shapley Additive Explanations (SHAP)

Recently, many studies have applied Explainable Artificial Intelligence (XAI) techniques to understand better the decision-making processes of machine learning models [25]. In this field, researchers have introduced numerous methods to explain how model features, inputs, or components influence model outputs [26]. In particular, these include Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), and SHAP [26]. However, LIME and SHAP are the most widely used because they can be applied to both tabular data and image data. SHAP, introduced by Lundberg and Lee [27], explains model results using concepts from cooperative game theory. SHAP can be applied for both local and global explanations [27]. Local explanations provide insight into individual predictions, while global explanations summarize feature influence across the entire dataset. Shapley values identify the contribution of each feature to the prediction produced by the model.

IV. Results

In this section, we present an analysis of the features and show the results of the classification experiments. In addition, it includes an interpretability analysis of the best-performing model using the SHAP framework.

A. Feature Distribution Analysis

A violin plot is a powerful visualization tool used to display the distribution of data and its probability density function. It summarizes key statistical values, including the median and quartiles, while also showing the full distribution of the dataset. In this study, we present violin plots for the numeric acoustic parameters, grouped by recovery status (Recovery vs. No Recovery).

In Fig. 2, the median value of MeanF0 in the No-Recovery group (171.644 Hz), with data points largely concentrated between 150 Hz and 250 Hz, is slightly higher than that of the Recovery group (168.300 Hz). Similarly, for all other parameters, the No-Recovery group consistently shows higher values compared to the Recovery group. This may suggest that higher values are associated with poorer voice recovery. However, this observation remains preliminary and requires further confirmation.

B. Classifier Performance

Table I summarizes the classification performances for the SVM, LR, and RF algorithms. The experiments reveal that SVM outperforms the other models in terms of accuracy. It reaches an accuracy of 76% with a standard deviation (SD) of 0.09, a recall of 87%, an F1-score of 80%, a precision of 77%, and an AUC score of 70%. This suggests their ability to identify the positive class. However, the SD value indicates that the model's performance is unstable between folds. In contrast, the RF model reaches the highest precision of 77%. However, its lower recall (73%) and accuracy (73%) indicate a lower ability to detect all true positives. Also, its performance is more variable. The lower results were obtained using the LR model, except for AUC (80%), which was higher than that of the SVM and RF methods.

In addition, the CI interval for SVM (accuracy: [0.68, 0.80]) shows moderate stability, indicating relatively consistent performance across folds. However, the wide confidence intervals for all models show that their performance can change significantly with different data.

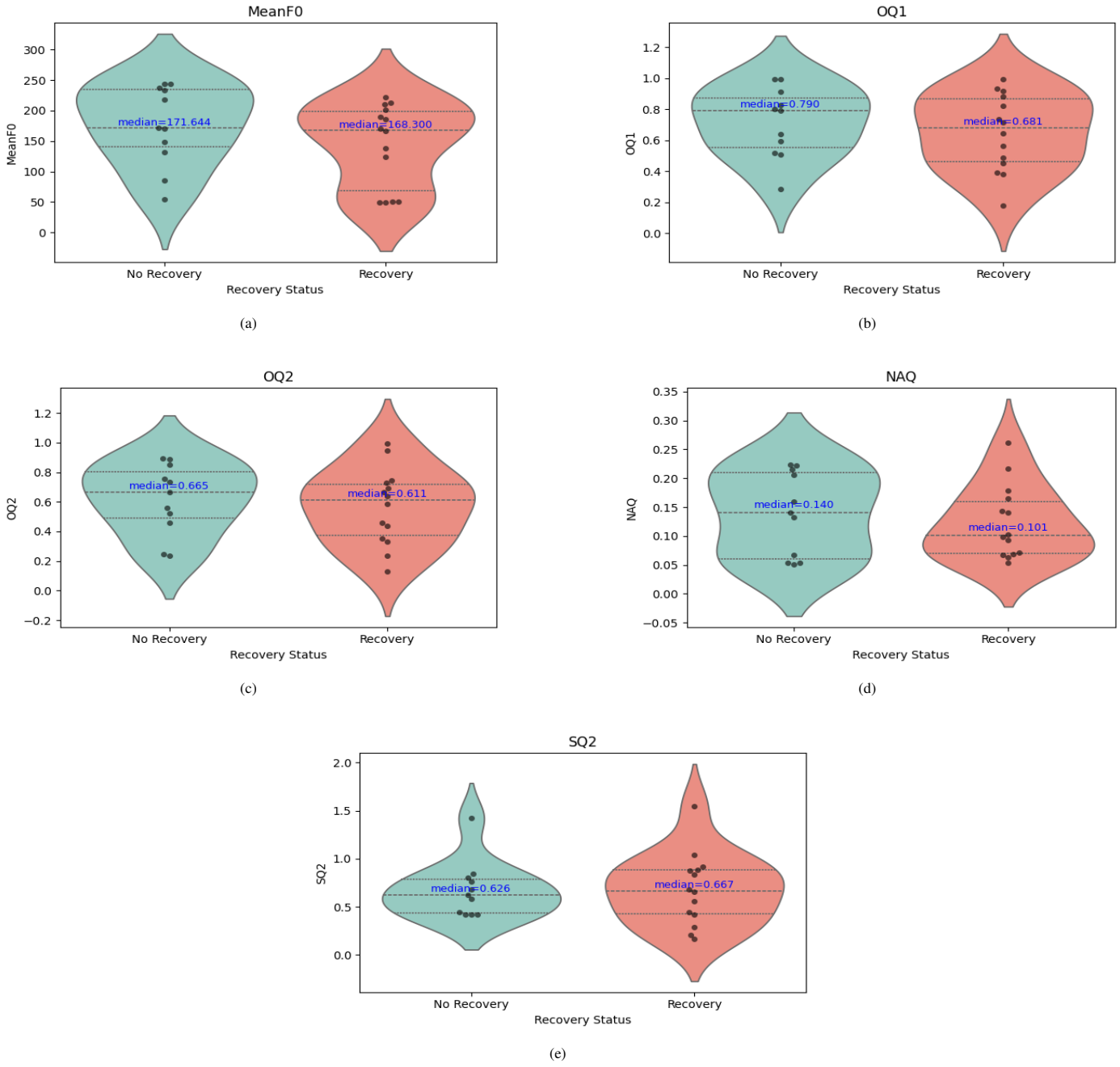


Fig. 2. Violin plots of acoustic and glottal features in the Recovery and No-Recovery groups: a) Mean fundamental frequency (MeanF0), b) Primary Open Quotient (OQ1), c) Secondary Open Quotient (OQ2), d) Normalized Amplitude Quotient (NAQ), and e) Secondary Speed Quotient (SQ2).

TABLE I. CROSS-VALIDATED PERFORMANCE METRICS USING SFS SELECTION METHOD (MEAN ± SD) [95% CI]

Algorithms	Accuracy	Precision	Recall	F1	AUC	Features
SVM	**0.76** ± 0.09 [0.68, 0.80]	**0.77** ± 0.14 [0.68, 0.88]	**0.87** ± 0.18 [0.73, 1.00]	**0.80** ± 0.08 [0.73, 0.85]	0.70 ± 0.15 [0.58, 0.82]	Type_CN, Type
RF	0.72 ± 0.23 [0.56, 0.88]	0.77 ± 0.22 [0.60, 0.93]	0.73 ± 0.28 [0.53, 0.93]	0.73 ± 0.22 [0.56, 0.89]	0.78 ± 0.23 [0.60, 0.95]	Age, Type_CN, Type, NAQ
LR	0.64 ± 0.17 [0.52, 0.76]	0.69 ± 0.19 [0.57, 0.85]	0.73 ± 0.28 [0.53, 0.93]	0.68 ± 0.17 [0.53, 0.79]	**0.80** ± 0.24 [0.62, 0.98]	Age, Type_CN, Type

C. Statistical Comparison

To verify whether the differences were statistically significant among the three classifiers, we conducted a Wilcoxon signed-rank test based on the bootstrapped F1-scores. The obtained results indicated that SVM consistently outperformed both RF and LR, with mean differences of 0.063 (95% CI [0.061, 0.065]) for SVM vs RF and 0.113 (95% CI [0.111, 0.114]) for SVM vs LR. In addition, RF achieved higher F1-scores than LR, with a mean difference of 0.050 (95% CI [0.049, 0.051]). Effect sizes were moderate to large (r = -0.55 for SVM vs RF, $r = -0.87$ for SVM vs. LR, $r = -0.87$ for RF vs. LR), and all comparisons were statistically significant ($p < 0.001$). These findings suggest that SVM demonstrates superior performance in distinguishing between classes on the dataset analyzed.

D. Model Interpretation and Error Analysis

Fig. 3, Fig. 4, and Fig. 5 present the confusion matrix, learning curve, and ROC curve, respectively, for the SVM classifier. Overall, the model demonstrated a strong ability to differentiate between the classes. It performed particularly well in identifying recovery recordings, correctly classifying 12 out of 14 samples. The no-recovery class was classified moderately well, with 7 correct predictions, though some samples were misclassified as recovery.

To better understand the model's behavior, we analyzed the misclassified samples from the SVM model predictions in Table II. The analysis shows that most false positives (FP) (patients predicted as recovered but who did not actually recover) occurred in three patients who underwent partial surgeries for benign conditions. It generally occurs in younger or middle-aged adults (ages 35–53). A false positive was also observed in a patient who had a total thyroidectomy for a malignant thyroid. In contrast, false negatives (patients predicted as not recovered but who actually recovered) were more common among younger patients (ages 31–42) who underwent total surgeries for benign thyroid conditions. These findings suggest that the model's performance is significantly influenced by the type of surgery and the underlying pathology. In addition, ROC analysis supports these findings, with an AUC of 65% achieved for the recovery class. However, due to the small number of misclassified cases, we did not perform statistical tests, as such analyses would not provide reliable or meaningful results.

Dealing with the SVM learning curve (see Fig. 5) displays initial overfitting, with high training accuracy but low validation accuracy. As the training size increases, validation accuracy rises and converges toward the training curve, indicating improved generalization and reduced overfitting.

E. SHAP Results

Fig. 6 presents the SHAP summary plot (mean absolute SHAP values) by class. These values represent the average contribution of each feature to the predictions of the model. In the No recovery class, the most influential variables were Type (partial/total) and Age (patient age), with SHAP values of 0.09 and 0.05, respectively. Other relevant features included Type_CN (benign/malign), OQ1, MeanF0, NAQ, SQ2, and OQ2. In contrast, the Recovery class is mainly associated with

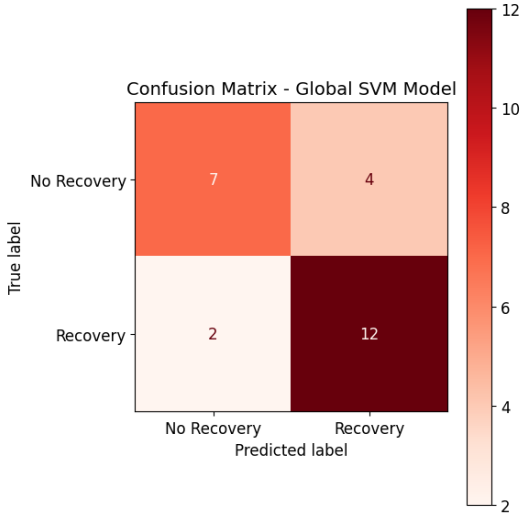


Fig. 3. Confusion matrix for the SVM model predicting post-thyroidectomy voice recovery. The rows represent actual classes and the columns predicted classes.

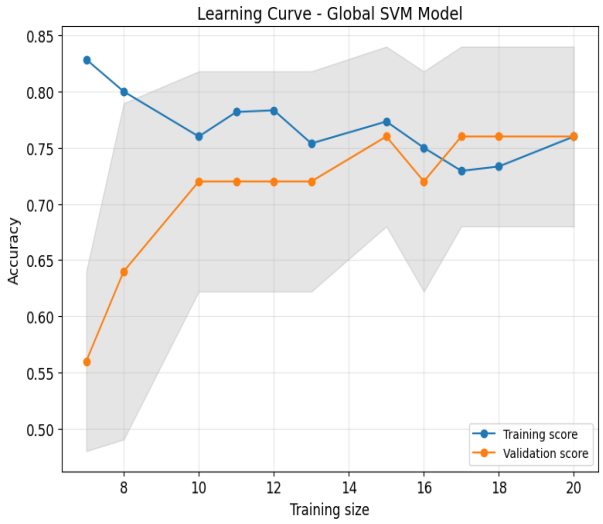


Fig. 4. Learning curve for the SVM method predicting post-thyroidectomy voice recovery. The x-axis represents the number of training samples used, and each point shows the mean accuracy over five cross-validation folds. Shaded regions indicate ±1 standard deviation across folds.

Type (0.08), Type_CN (0.03), and Age (0.03), which have the highest impact.

Fig. 7 presents SHAP-based waterfall plots for three participants selected at random. These plots illustrate the contribution of individual variables to the prediction of the model for each patient. The horizontal axis represents the value predicted by the model, denoted as while each bar illustrates the marginal effect of a specific variable on this prediction. In each plot, red bars indicate positive contributions to the decision, while the blue bars represent a negative contribution. In other words, red means features that increase the model's predicted probability of class, whereas blue bars correspond to features that reduce it.

TABLE II. ERROR ANALYSIS OF MISCLASSIFIED PATIENTS FOR SVM MODEL

Patient_ID	ErrorType	Age	Type	Type_CN	TrueLabel	PredictedLabel
1	False Positive	35	Partial	Benign	0	1
2	False Negative	42	Total	Benign	1	0
3	False Positive	52	Partial	Benign	0	1
4	False Positive	53	Partial	Benign	0	1
5	False Positive	31	Total	Malign	0	1
6	False Negative	31	Total	Benign	1	0

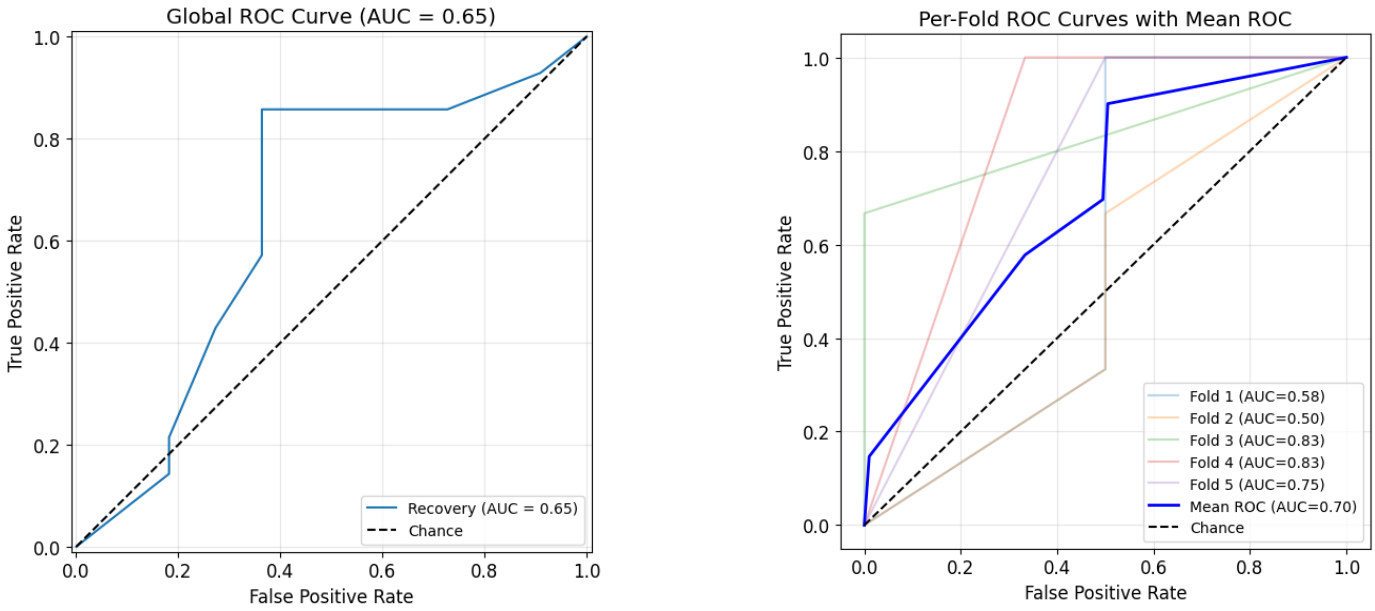


Fig. 5. ROC curve for the SVM method: a) Global ROC curve showing the overall discriminative performance (AUC calculated on all test predictions across folds). b) Per-fold ROC curves with the mean ROC curve (bold line) to illustrate variability across cross-validation folds.

1) Participant 1: is a 35-year-old woman who underwent partial thyroidectomy for a benign condition. For this woman, the predicted score is -0.129 which is below the mean value (-0.004). Type, OQ1, and Age show negative contributions (-0.06, -0.01, and -0.07, respectively), while NAQ contributed positively. Other features had negligible effects.

2) Participant 6: is a 35-year-old woman who underwent total thyroidectomy for a malignant condition. The model predicted a score of 0.016, higher than the baseline value of 0.004. Most features contributed negatively to the prediction, except for Type_CN, MeanF0, and Age, which had positive effects. Among these positive contributors, Type_CN had the largest value (+0.07), in contrast to what was observed for Participant 1.

3) Participant 15: is a 64-year-old woman who underwent a partial thyroidectomy for a benign condition. Most parameters had negligible effects.

V. DISCUSSION

Vocal changes are one of the common problems for subjects undergoing thyroid surgery. In the current work, we predict vocal recovery following thyroidectomy. For this purpose, machine learning techniques (ML) and a set of features are used. In addition, we use various ML algorithms to classify voices into recovery and non-recovery classes. All participants were evaluated before surgery and a month postoperative. This work was designed as a pilot feasibility study rather than a fully powered trial.

The SVM technique achieved the highest performance, with an F1-score of 80%. The SHAP analysis revealed that the most important features contributing to no-recovery predictions were the type of performed operation (partial/total), followed by age, and then the nature of the thyroid (malignant/benign). Various works have also found that speech can be affected by the extent of thyroid surgery [28], [29], [30]. Yılmaz et al. [31] reported that the type of surgical procedure (subtotal thyroidectomy or total), age, histopathologic diagnosis, and sex did not significantly impact vocal quality after thyroidectomy.

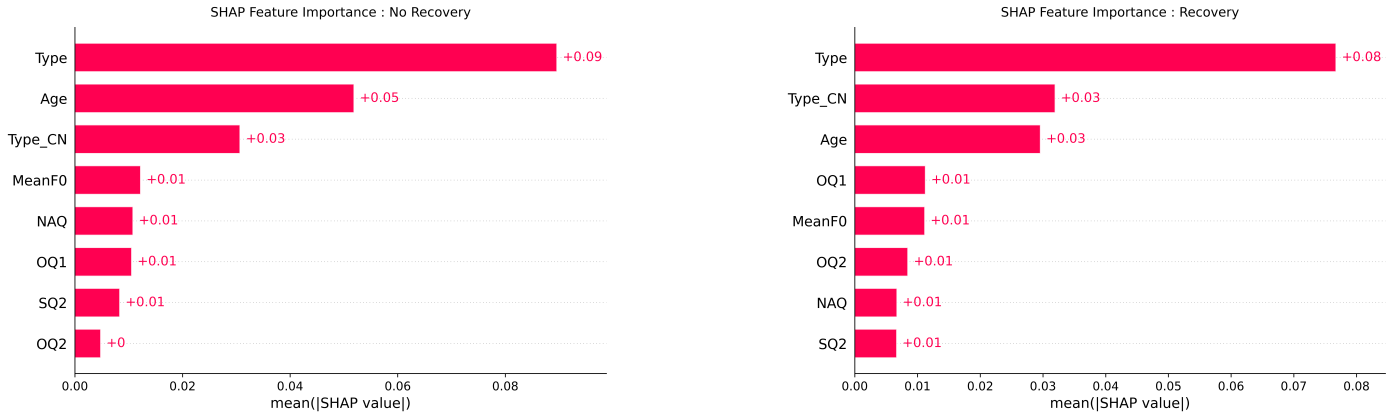
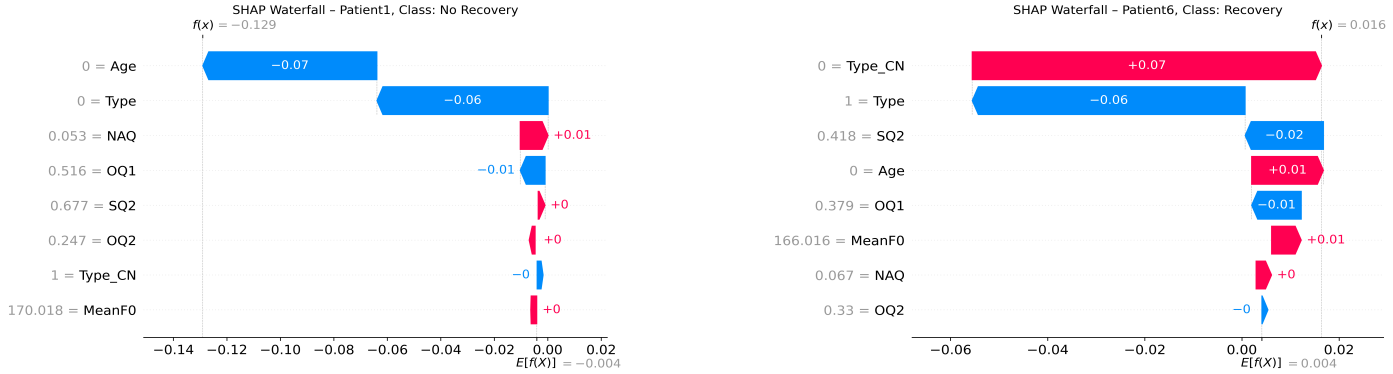
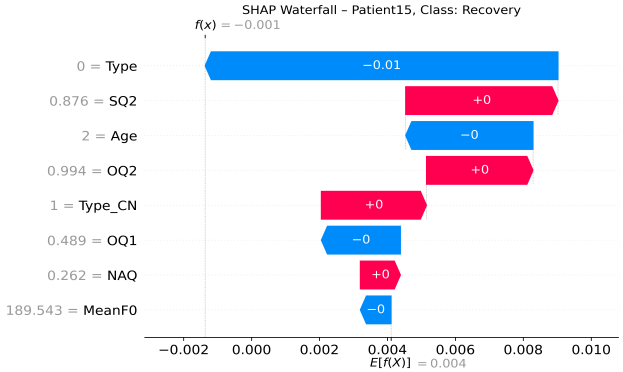


Fig. 6. SHAP summary plot (mean absolute SHAP values) for recovery and no recovery classes.



(a) Participant 1

(b) Participant 6



(c) Participant 15

Fig. 7. Example of waterfall plots for three participants selected randomly from the dataset.

However, they observed that voice changes were associated with tumor type. This may be due to various factors, such as injuries to the Recurrent Laryngeal Nerve (RLN) [32]. In addition, patients' emotions, experiences, and professional occupations may significantly impact these outcomes. Some existing studies have also identified female gender and the volume of the resected gland as risk factors for post-operative speech alterations after total thyroidectomy [28]. Also, a set of glottal features corresponding to different phases of the glottal cycle also contributed to the model's performance. Among these, OQ, and SQ are commonly used time-based parameters, while NAQ represents the shape and timing of the glottal waveform [18].

In our SHAP analysis, the glottal features had a similar impact in predicting both classes. Existing studies showed that in healthy individuals, higher OQ1 and OQ2 values are associated with better voice quality. In contrast, lower values are generally associated with pathological conditions such as nodules and vocal fold paralysis [18]. Moreover, OQ is

inversely correlated with vocal intensity, meaning that lower values correspond to higher vocal intensity [24].

For the NAQ parameter, higher values are indicative of a more symmetrical glottal waveform and are associated with breathy phonation type [33], [34]. OQ and NAQ are also inversely related to vocal intensity [24]. A decrease in NAQ may reflect increased vocal intensity due to a prolonged glottal closure phase [35].

Previous studies have revealed that the SQ2 parameter is lower in healthy voices than in pathological ones. In Addition, decreased values of these parameters have been reported in female patients with heart failure, suggesting their sensitivity to physiological alterations [24]. The glottal parameters are also applied to distinguish between patients with healthy voices and those with vocal fold paralysis [20] or neurological disorders such as Parkinson's disease [36]. These outcomes underline the clinical value of using glottal source features in postoperative voice assessment.

Some recent studies related to this current study have utilized ML techniques to predict thyroid cancer in patients who underwent thyroidectomy. They also achieved promising results in predicting thyroid cancer [14], [13], [37]. Lee *et al.* [13] developed a deep neural network using preoperative and two-week postoperative spectrograms to predict three-month postoperative vocal recovery in 114 patients. Their model achieved high predictive performance, with AUC values of 0.918 for breathiness, 0.735 for asthenia, and 0.894 for grade. In contrast, the present study achieved an AUC of 0.65 for the recovery class using acoustic and glottal features with SVM model on a smaller dataset of 25 patients. The lower performance may be attributed to the smaller sample size and differences in feature representation. Also, our study did not include deep learning models for classification. However, our method provides interpretable predictions using SHAP analysis to identify parameters that most contribute to voice recovery.

In terms of predictive performance, our results show lower accuracy compared with Kurt *et al.* [14] in identifying patients at risk of developing vocal cord palsy before thyroidectomy using ML techniques. However, their study addressed a different clinical problem, focusing on preoperative risk prediction of vocal cord palsy, while the present study predicts postoperative voice recovery. Therefore, direct numerical comparison of accuracy is not appropriate. Although deep learning approaches have demonstrated promising results and the potential to enhance model performance, they were not employed in this study. This decision was primarily due to the limited size of our dataset, which is significantly smaller than those used in comparable studies. Furthermore, this research does not primarily aim to optimize performance measures, but to investigate glottal and acoustic parameters relevant to thyroidectomy. These parameters have previously demonstrated their ability to detect changes in voice after thyroid surgery.

A. Limitations of the Study

Several limitations must be acknowledged. In our study, we employ the PSIAIF method to estimate glottal flow without feature-sensitivity analysis. This method, however, is sensitive to parameter configuration, which can affect the performance of derived features such as OQ and NAQ. In addition, the dataset was relatively small wich may introduce statistical fragility, as model performance and feature importance estimates may be sensitive to minor variations in the data. Although cross-validation was employed to mitigate overfitting, small datasets inherently increase the risk of variability and reduced generalizability. Also, calibration analysis was not performed. Therefore, the agreement between predicted probabilities and actual outcomes remains unassessed. Another limitation is that all participants were female. This approach has reduced the variability of vocal parameters related to physiological differences between the sexes, but it also limits the generalizability of the results to male patients. In fact, male and female voices differ in fundamental frequency and glottal configuration, which can influence these parameters. It is therefore essential to create predictive models that reflect actual clinical diversity. In addition, we did not include external validation because the existing datasets in the literature are private and not publicly accessible. More precisely, no publicly available longitudinal datasets containing preoperative and postoperative voice recordings were identified. As a result, the generalizability of the proposed models to external populations remains unconfirmed. Finally, we used traditional ML models due to their interpretability and suitability for small datasets. However, the deep learning approaches were not explored in this work due to the limited dataset size.

VI. Conclusion

This pilot feasibility study applied machine learning techniques based on different features to predict voice recovery after thyroidectomy. The dataset was collected on the day before surgery and a month after surgery. Preliminary performance varied between models, with the SVM classifier achieving an F1-score of 80%. These findings suggest that machine learning-based prediction is feasible, but not yet conclusive. Future research is encouraged to address the study's limitations through several directions. It is necessary to use a larger, mixed-gender cohort to improve the generalization of the model. In addition, model performance could be improved by applying advanced inverse filtering methods, such as IAIF+, or closed-phase covariance approaches. Another direction is to employ deep learning architectures, which may further improve predictive accuracy. Taken together, these preliminary findings suggest that machine learning may provide a feasible approach for predicting voice recovery after thyroidectomy. However, external validation is required before clinical application.

Declaration of Interest

No competing interests.

Availability of Data

The dataset used in this study cannot be publicly released due to patient privacy regulations and institutional restrictions. However, to support reproducibility, we provide a complete description of the feature extraction process, model architectures, hyperparameters, validation strategy, and evaluation protocol. All experiments were conducted using deterministic random seeds and standard machine learning libraries.

ACKNOWLEDGMENT

The present research has been produced under the Moroccan-Tunisian project titled "Multimodal Detection and Classification of Pathological Voice Disorders". The Ministry of Higher Education and Scientific Research in Tunisia and the Ministry of Higher Education, Scientific Research and Innovation in Morocco sponsor it. We extend our gratitude to Ilhem Charfeddine, Malek Mnejja and, Mariam Ben Ayed from the Department of Otorhinolaryngology-Head and Neck Surgery at Habib Bourguiba University Hospital, Sfax, Tunisia, for their valuable contributions to data collection and for providing essential information on the study participants.

ETHICAL APPROVAL

We conducted our study after approval from the Ethical Committee of the Medical Institution. All patients were informed about the study protocol. Signed consent was obtained from each participant to confirm their agreement to participate. They were informed that their anonymous recording data would be used for acoustic and video analysis.

REFERENCES

[1] K. Van Lierde, E. D'haeseleer, F. L. Wuyts, N. Baudonck, L. Bernaert, and H. Vermeersch, "Impact of thyroidectomy without laryngeal nerve injury on vocal quality characteristics: an objective multiparameter approach," *The Laryngoscope*, vol. 120, no. 2, pp. 338–345, 2010.

[2] C. H. Ryu, S. J. Lee, J.-G. Cho, I. J. Choi, Y. S. Choi, Y. T. Hong, S. Y. Jung, J. W. Kim, D. Y. Lee, D. K. Lee *et al.*, "Care and management of voice change in thyroid surgery: Korean society of laryngology, phoniatrics and logopedics clinical practice guideline," *Clinical and Experimental Otorhinolaryngology*, vol. 15, no. 1, pp. 24–48, 2022.

[3] I. de Pedro Netto, A. Fae, J. G. Vartanian, A. P. B. Barros, L. M. Correia, R. N. Toledo, J. R. G. Testa, I. N. Nishimoto, L. P. Kowalski, and E. C.-d. Angelis, "Voice and vocal self-assessment after thyroidectomy," *Head & neck*, vol. 28, no. 12, pp. 1106–1114, 2006.

[4] S. R. Schwartz, S. M. Cohen, S. H. Dailey, R. M. Rosenfeld, E. S. Deutsch, M. B. Gillespie, E. Granieri, E. R. Hapner, C. E. Kimball, H. J. Krouse *et al.*, "Clinical practice guideline: hoarseness (dysphonia)," *Otolaryngology–Head and Neck Surgery*, vol. 141, no. 1_suppl, pp. 1–31, 2009.

[5] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases," *Neural Computing and Applications*, vol. 32, pp. 15 747–15 757, 2020.

[6] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," *Current opinion in otolaryngology & head and neck surgery*, vol. 16, no. 3, p. 211, 2008.

[7] J. Oates, "Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[8] M. Mizuta, C. Abe, E. Taguchi, T. Takeue, H. Tamaki, and T. Haji, "Validation of cepstral acoustic analysis for normal and pathological voice in the japanese language," *Journal of Voice*, vol. 36, no. 6, pp. 770–776, 2022.

[9] F. T. Al-Dhief, N. M. A. Latiff, N. N. N. A. Malik, N. S. Salim, M. M. Baki, M. A. A. Albadr, and M. A. Mohammed, "A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms," *IEEE Access*, vol. 8, pp. 64 514–64 533, 2020.

[10] K.-W. Jung, Y.-J. Won, H.-J. Kong, and E. S. Lee, "Cancer statistics in korea: incidence, mortality, survival, and prevalence in 2016," *Cancer research and treatment: official journal of Korean Cancer Association*, vol. 51, no. 2, pp. 417–430, 2019.

[11] M. K. Reddy and P. Alku, "A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation," *IEEE Access*, vol. 9, pp. 135 953–135 963, 2021.

[12] N. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, 2020.

[13] J. H. Lee, C. Y. Lee, J. S. Eom, M. Pak, H. S. Jeong, and H. Y. Son, "Predictions for three-month postoperative vocal recovery after thyroid surgery from spectrograms with deep neural network," *Sensors*, vol. 22, no. 17, p. 6387, 2022.

[14] B. Kurt, İ. B. Kırkbir, T. Kurt, A. Güner, and M. Uluşahin, "A novel computer based risk prediction model for vocal cord palsy before thyroidectomy," *Computer Methods and Programs in Biomedicine*, vol. 236, p. 107563, 2023.

[15] L. Soylu, S. Ozbas, H. Y. Uslu, and S. Kocak, "The evaluation of the causes of subjective voice disturbances after thyroid surgery," *The American journal of surgery*, vol. 194, no. 3, pp. 317–322, 2007.

[16] F. Debruyne, F. Ostyn, P. Delaere, and W. Wellens, "Acoustic analysis of the speaking voice after thyroidectomy," *Journal of Voice*, vol. 11, no. 4, pp. 479–482, 1997.

[17] A. Stojadinovic, L. R. Henry, R. S. Howard, J. Gurevich-Uvena, M. J. Makashay, G. L. Coppit, C. D. Shriver, and N. P. Solomon, "Prospective trial of voice outcomes after thyroidectomy: evaluation of patient-reported and clinician-determined voice assessments in identifying postthyroidectomy dysphonia," *Surgery*, vol. 143, no. 6, pp. 732–742, 2008.

[18] M. Mnejja, S. B. Jebara, M. B. Ayed, S. Ayadi, O. Walha, B. Hammami, and I. Charfeddine, "Glottal features in vocal assessment following thyroidectomy," *Journal of Voice*, 2024.

[19] G. P. Kafentzis, Y. Stylianou, and P. Alku, "Glottal inverse filtering using stabilised weighted linear prediction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5408–5411.

[20] A. Bedoui and S. B. Jebara, "On the use of opening phase slopes of the glottal signal to characterize unilateral vocal folds paralysis," in *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2016, pp. 41–46.

[21] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[22] X. Yao, W. Bai, Y. Ren, X. Liu, and Z. Hui, "Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion," *Neurocomputing*, vol. 410, pp. 328–341, 2020.

[23] "Software aparat," November 15, 2022. [Online]. Available: http://aparat.sourceforge.net/index.php/

[24] M. Kohler, M. M. Vellasco, E. Cataldo *et al.*, "Analysis and classification of voice pathologies using glottal signal parameters," *Journal of Voice*, vol. 30, no. 5, pp. 549–556, 2016.

[25] R. Jegan and R. Jayagowri, "Voice pathology detection using optimized convolutional neural networks and explainable artificial intelligence-based analysis," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 27, no. 14, pp. 2041–2057, 2024.

[26] A. M. Salih, I. B. Galazzo, P. Gkontra, E. Rauseo, A. M. Lee, K. Lekadir, P. Radeva, S. E. Petersen, and G. Menegaz, "A review of evaluation approaches for explainable ai with applications in cardiology," *Artificial Intelligence Review*, vol. 57, no. 9, p. 240, 2024.

[27] R. K. Makumbura, L. Mampitiya, N. Rathnayake, D. Meddage, S. Henna, T. L. Dang, Y. Hoshino, and U. Rathnayake, "Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (xai) techniques like shapley additive explanations (shap) for interpreting the black-box nature," *Results in Engineering*, vol. 23, p. 102831, 2024.

[28] J. Ryu, Y. M. Ryu, Y.-S. Jung, S.-j. Kim, Y. J. Lee, E.-K. Lee, S.-K. Kim, T.-S. Kim, T. H. Kim, C. Y. Lee *et al.*, "Extent of thyroidectomy affects vocal and throat functions: a prospective observational study of lobectomy versus total thyroidectomy," *Surgery*, vol. 154, no. 3, pp. 611–620, 2013.

[29] D. A. Vicente, N. P. Solomon, I. Avital, L. R. Henry, R. S. Howard, L. B. Helou, G. L. Coppit, C. D. Shriver, C. C. Buckenmaier, S. K. Libutti *et al.*, "Voice outcomes after total thyroidectomy, partial thyroidectomy, or non-neck surgery using a prospective multifactorial assessment," *Journal of the American College of Surgeons*, vol. 219, no. 1, pp. 152–163, 2014.

[30] A. Stojadinovic, A. R. Shaha, R. F. Orlikoff, A. Nissan, M.-F. Kornak, B. Singh, J. O. Boyle, J. P. Shah, M. F. Brennan, and D. H. Kraus, "Prospective functional voice assessment in patients undergoing thyroid surgery," *Annals of surgery*, vol. 236, no. 6, pp. 823–832, 2002.

[31] B. Yılmaz, S. Bakır, E. E. Yılmaz, E. Şengül, Ö. Uslukaya, A. Gül, F. E. Özkurt, and İ. Topçu, "An analysis on aerodynamic and acoustic changes after thyroidectomy," *International Surgery*, vol. 101, no. 5-6, pp. 233–240, 2016.

[32] P. Aluffi, M. Policarpo, C. Cherovac, M. Olina, R. Dosdegani, and F. Pia, "Post-thyroidectomy superior laryngeal nerve injury," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 9, pp. 451–454, 2001.

[33] A. Palaparthi and I. R. Titze, "Analysis of glottal inverse filtering in the presence of source-filter interaction," *Speech communication*, vol. 123, pp. 98–108, 2020.

[34] K. R. Mittapalle, H. Pohjalainen, P. Helkkula, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku, "Glottal flow characteristics in vowels produced by speakers with heart failure," *Speech Communication*, vol. 137, pp. 35–43, 2022.

[35] P. Alku, M. Airas, E. Björkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 1052–1062, 2006.

[36] M. Novotnỳ, P. Dušek, I. Daly, E. Růžička, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with parkinson's disease: correlation between acoustic speech characteristics and non-speech motor performance," *Biomedical Signal Processing and Control*, vol. 57, p. 101818, 2020.

[37] Y. Xiao, Z. Wu, S. Ruan, Y. Xiong, and T. Huang, "Development and validation of the nomogram for predicting preoperative vocal cord palsy in thyroid cancer patients," *Gland Surgery*, vol. 10, no. 2, p. 541, 2021.