Vocal Pathology Recognition Using Acoustic Features

Karen Itsuki Taniguchi
Federal University of Technology – Paraná (UTFPR)
Cornélio Procópio, Paraná, Brazil
karenitaniguchi@gmail.com

Laís Aya Taniguchi
Federal University of Technology – Paraná (UTFPR)
Pato Branco, Paraná, Brazil
laisaya.taniguchi@gmail.com

Maiara Mitiko Taniguchi
Serviço Nacional de Aprendizagem Industrial (SENAI)
Florianópolis, Santa Catarina, Brazil
maiara.taniguchi@sc.senai.br

Daniel Prado Campos
Federal University of Technology – Paraná (UTFPR)
Apucarana, Paraná, Brazil
danielcampos@utfpr.edu.br

Rafael Gomes Mantovani
Federal University of Technology – Paraná (UTFPR)
Apucarana, Paraná, Brazil
rafaelmantovani@utfpr.edu.br

ABSTRACT

Voice disorders impact communication, employability, and quality of life; however, gold-standard assessments based on laryngoscopy remain invasive and resource-intensive. This work examines the impact of vowel choice, pitch condition, speaker sex, and feature design on the performance of machine-learning models for automatic voice pathology detection. Using the Saarbrücken Voice Database (SVD), we build a benchmark on sustained vowels by deriving 120 binary classification tasks from the factorial combination of vowel (*/a/*, */i/*, */u/*, all), pitch condition (low, high, normal, low-high-low, all), and speaker sex (male, female, both). We compared a baseline acoustic representation inspired by recent work with an extended feature set. Four off-the-shelf classifiers were evaluated using AUC as the main performance metric. Results show that XGBoost and SVM consistently achieve the best ranks, with median AUC values around 0.80 and a maximum of 0.866 for the configuration combining novel features, all pitch conditions, male speakers, and vowel */a/*. Sex-specific models consistently surpass mixed-sex models, and feature-importance analysis highlights spectral bandwidth, jitter, and shimmer as key descriptors. The proposed feature set outperforms the baseline, and using only sustained */a/* is as effective as using all vowels, simplifying acquisition. Future work may improve the pre-processing step, expand the feature set, and employ deep features.

KEYWORDS

voice pathology, audio signal, machine learning, telemedicine

1 INTRODUCTION

Voice disorders impair communication, work ability, and quality of life. Clinical diagnosis frequently relies on laryngoscopy and stroboscopy, which are invasive, resource-intensive, and uncomfortable for patients. These constraints have driven the development of non-invasive, signal-based screening methods that analyze sustained vowels or brief speech segments and flag likely pathology. Recent studies formalize this pipeline as acquisition, preprocessing, feature extraction, supervised classification into healthy vs. pathological (and, in some cases, into specific disorder categories) [1, 2].

Machine learning (ML) methods and, more recently, deep neural networks (DNNs) have accelerated progress by learning discriminative representations directly from acoustic signals and, when available, complementary electroglottographic (EGG) traces [1]. Convolutional architectures trained on sustained /a/ phonemes have reported competitive results. Moreover, hybrid CNN-RNN models, residual networks, and transfer learning have also been explored, often using time-frequency representations such as mel-spectrograms or alternatives handcrafted to speech perception [3–5]. Despite advances, performance remains sensitive to feature design, model choice, dataset composition, and evaluation protocol.

Despite the steady progress, two aspects remain insufficiently quantified. First, experimental protocols vary widely across studies, particularly in terms of which vowel is analyzed and under what pitch condition. Many works focus on a single vowel, which obscures the influence of phonetic content and phonatory setting on performance [6]. Second, the contribution of fusing metadata (such as sex and age) with audio data is not well established [1, 2].

Accordingly, this study has two objectives. Firstly, to quantify how sustained-vowel choice and pitch condition affect binary voice-pathology classification. Second, assess the contribution of audio-only models against models augmented with sex and age metadata. Throughout, we exclude electroglottographic (EGG) and read-sentence materials, and we report ROC/AUC alongside precision, recall, F_1, and confusion matrices to support clinically oriented interpretation.

2 RELATED WORK

Research on non-invasive screening of vocal disorders has evolved along two complementary lines: (i) handcrafted acoustic descriptors followed by classical classifiers, and (ii) deep representations learned from time-frequency images (or raw waveforms), sometimes complemented by electroglottographic (EGG) or clinical information.

2.1 Handcrafted features and classical ML

Early and still widely used pipelines extract short-time descriptors, such as Mel-Frequency Cepstral Coefficients (MFCCs), energy and zero-crossing statistics, and perturbation measures like jitter, shimmer, and harmonic-to-noise ratio (HNR), and feed them to supervised learners (SVM, decision trees, Naïve Bayes, etc.). A recent comparative study across different voice databases reaffirmed MFCCs as a strong baseline while highlighting the importance of balanced splits and richer metrics beyond accuracy [2]. In the

same study, Online Sequential ELM (OSELM) stood out among classical models, reaching over 86% accuracy in single database evaluation and keeping consistent performance (over 81%) under cross-database testing [2].

More recently, a comprehensive and reproducible benchmark on the Saarbrücken Voice Database (SVD) focused on handcrafted descriptors for binary vocal-pathology detection [7]. In that work, the authors restricted the analysis to adult speakers producing sustained /a:/ at normal pitch, carefully selecting at most one healthy and one pathological recording per subject and trimming leading/trailing silence.

From each recording, they extracted a set of acoustic and meta-acoustic features, including mean and variability of the fundamental frequency, classical perturbation measures (jitter, shimmer, HNR), spectral descriptors (centroid, contrast, flatness, roll-off, zero-crossing rate), and cepstral statistics based on MFCCs and LFCCs. Two additional features were proposed to better capture irregular phonation: a binary indicator of unreliable f0 estimation (NaN feature) and a pitch-difference descriptor that quantifies intra-utterance variability of f0. Age and sex were also incorporated as metadata.

On top of these descriptors, [7] systematically evaluated thousands of feature subsets combined with six classical classifiers (k-NN, Naïve Bayes, decision trees, random forests, AdaBoost, and SVM with RBF kernel). The evaluation protocol relied on stratified 10-fold cross-validation with minority-class oversampling via k-means SMOTE, applied only to the training folds, followed by feature scaling and a grid search over classifier hyperparameters. Instead of accuracy, model selection and reporting emphasized metrics that are more robust to class imbalance, such as Matthews Correlation Coefficient (MCC), sensitivity, specificity, geometric mean, and unweighted average recall. The best random forest and AdaBoost configurations achieved UAR values around 85% on SVD, establishing a strong and transparent baseline for future work [7].

2.2 Deep time-frequency representations

Convolutional architectures trained on spectro-temporal images (e.g., mel-spectrograms, gammatonegrams, scalograms) have delivered consistent gains over purely handcrafted features in several settings [3]. Beyond standard mel-spectrograms, alternatives that better approximate human auditory filtering or capture non-stationarity have been explored. For instance, a multidimensional feature extraction approach combining gammatone-based representations with Teager-Kaiser energy (TKEO) scalograms, coupled with a modified ResNet backbone, reported high accuracy in both binary and multi-class setups [5]. Hybrid CNN-RNN designs have also been proposed to model local spectral patterns and longer-range temporal dependencies within the same framework [4].

3 EXPERIMENTAL METHODOLOGY

This section details the experimental methodology adopted in this study. An end-to-end overview of the pipeline—from data acquisition to model evaluation—is depicted in Figure 1.

3.1 Dataset

The Saarbrücken Voice Database (SVD) is a clinically curated corpus of German speech developed by the Phonetics group at Saarland University and made publicly available for research on normal and pathological phonation.[1] It contains recordings from 1002 pathological speakers and 851 controls, corresponding to 2225 recording sessions and covering a broad range of voice disorders. Among the pathological speakers, 454 are male, and 548 are female; in the control group, 423 are male, and 428 are female, as shown in Figure 2. Each session includes sustained vowels */i/*, */a/*, */u/* produced at typical, higher, lower, rising, followed by falling pitch, plus a short read sentence. Microphone audio and simultaneous electroglottograph (EGG) are provided as high-quality WAV signals, recorded at 50 kHz, 16-bit precision [6, 8].

The SVD database comprises functional and organic disorders, featuring 71 diagnosis labels. Among singly-labeled pathological sessions, the most frequent diagnoses include vocal fold paralysis, hyperfunctional dysphonia, and laryngitis [6]. Recordings are organized by speaker/session and diagnosis category, and metadata fields include speaker's sex, age, and clinical diagnosis, enabling either a binary (healthy vs. pathological) aggregation or diagnosis-specific stratification.

To illustrate the acoustic variability present in the curated subset, Figure 3 shows waveforms and log-magnitude spectrograms for two female speakers: a healthy control (session ID 1, panels a–b) and a pathological case diagnosed with laryngitis (session ID 493, panels c–d). Both samples correspond to the sustained vowel /a/ at neutral pitch, after silence trimming and preprocessing as described above.

The spectrogram of the healthy speaker (panel a) presents a clear harmonic stack with well-defined formant bands and relatively low broadband noise between harmonics, while the corresponding waveform (panel b) exhibits a highly regular, quasi-periodic oscillation with relatively stable cycle-to-cycle amplitude. In contrast, the pathological sample with laryngitis (panels c–d) exhibits a blurrier harmonic structure with increased aperiodic energy at mid-to-high frequencies, accompanied by more pronounced amplitude fluctuations and irregularities in the time-domain signal.

In line with Vrba *et al.* [7], we restrict our experiments to microphone recordings of the sustained vowel */a/* produced at neutral pitch, exclude EGG signals and read sentences from all experiments, and collapse the original diagnosis categories into a binary label of healthy versus pathological.

We adopt the machine-readable metadata table scraped from the SVD website and released by [7] *et al.* as part of their study and companion repository[2]. The same filtering strategy is applied: recordings from subjects younger than 18 years are removed; sessions reported as corrupted or containing artifacts in the original comments are excluded; recordings marked as produced with a singing voice are discarded; and, for each subject, at most one healthy and one pathological */a/* recording are retained, selecting the oldest session by date and recording ID to avoid data leakage [7]. For the remaining signals, we reuse the silence-trimmed waveforms

[1] https://stimmdb.coli.uni-saarland.de/
[2] https://github.com/aailab-uct/Automated-Robust-and-Reproducible-Voice-Pathology-Detection

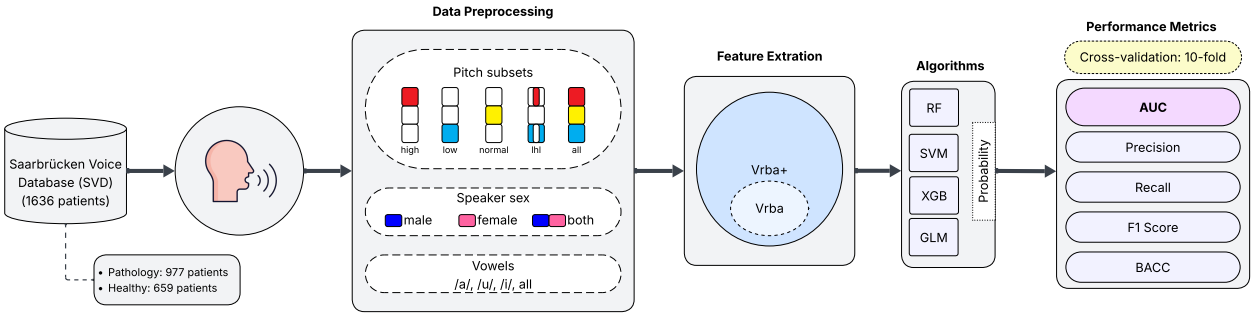


Figure 1: Experimental methodology pipeline for voice pathology detection.

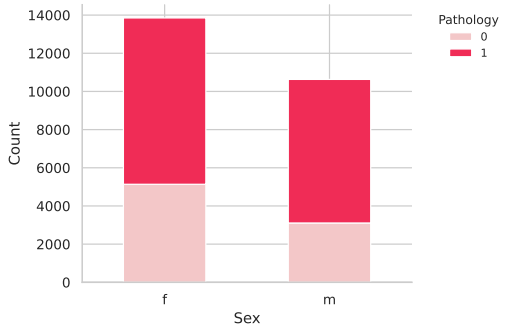


Figure 2: Distribution of healthy and pathological recordings by speaker sex in the curated SVD dataset.

provided in the repository, where leading and trailing segments that are more than 15 dB below the maximum root mean square amplitude are removed using `librosa.effects.trim`. This results in the same curated subset of 1636 trimmed */a/* recordings described by Vrba *et al.* [7].

3.2 Data preprocessing and feature extraction

Starting from the curated and silence-trimmed subset described above, we compute frame-level acoustic features directly on the trimmed sustained */a/* segment provided by Vrba *et al.* [7]. The original SVD speech channel is recorded at $f_s = 50\,\text{kHz}$; signals are processed at their native sampling rate (or resampled to 50 kHz when needed). No additional denoising, silence removal, or loudness normalization is applied beyond the preprocessing pipeline of Vrba *et al.*, so that the data presented to the classifiers are identical to those used in [7]. Speaker sex and age are taken from the machine-readable SVD metadata compiled by Vrba *et al.* and used only as optional covariates. The target is a binary label (healthy vs. pathological) aggregated from the database diagnosis categories.

Table 1 summarizes the acoustic descriptors computed on each trimmed segment. Frame-level features are extracted with a window of approximately 41 ms and a hop of approximately 10 ms, then aggregated as indicated (means and standard deviations where applicable). Pitch is estimated with pYIN; perturbation measures (local, RAP and PPQ5 jitter; local, APQ3 and APQ5 shimmer) and harmonics-to-noise ratio follow the standard procedures implemented in Praat; cepstral peak prominence (CPP) is computed from the log-spectral cepstrum by measuring the prominence of the main peak in the quefrency region corresponding to 60–300 Hz. These families cover source-related cues (periodicity and perturbation) and filter-related cues (spectral envelope), which are known to be informative for dysphonia screening.

We consider two handcrafted acoustic representations. The *Vrba (baseline)* feature set reproduces the descriptors used by Vrba *et al.* [7], including mean f_0, harmonics-to-noise ratio (HNR), jitter, shimmer, spectral centroid, spectral roll-off, zero-crossing rate, LFCC, MFCC-based statistics, skewness and shannon entropy. The *Vrba+ (extended)* feature set augments this baseline with minimum and maximum f_0, spectral bandwidth, RMS energy, band-wise spectral contrast (seven octave bands rather than a single average), additional jitter and shimmer variants (local, RAP, PPQ5, APQ3, APQ5), and a simple CPP measure. Table 1 lists all descriptors in the Vrba+ set; the Vrba baseline corresponds to the subset originally considered in [7].

3.3 Classification learning tasks

Two acoustic feature datasets were used in this study: the feature set proposed by Vrba *et. al* [7] and the expanded feature set (Vrba+). Each dataset was independently expanded into a structured collection of binary classification tasks derived from the factorial combination of three experimental factors:

- **Vowel**: */a/*, */i/*, */u/*, and all vowels combined;
- **Pitch condition**: low pitch, high pitch, normal pronunciation, low-high-low intonation, and all conditions combined;
- **Speaker sex**: male only, female only, and both.

This design resulted in $5 \times 3 \times 4 = 60$ tasks per dataset, totaling 120 binary classification tasks across the two datasets. Each task represents an independent classification problem aimed at discriminating healthy from pathological phonation.

Before model training, all tasks underwent a standardized preprocessing workflow. The procedure consisted of the following steps: i) removal of near-zero-variance features; ii) conversion of all non-numeric predictors (factor or character types) into numeric

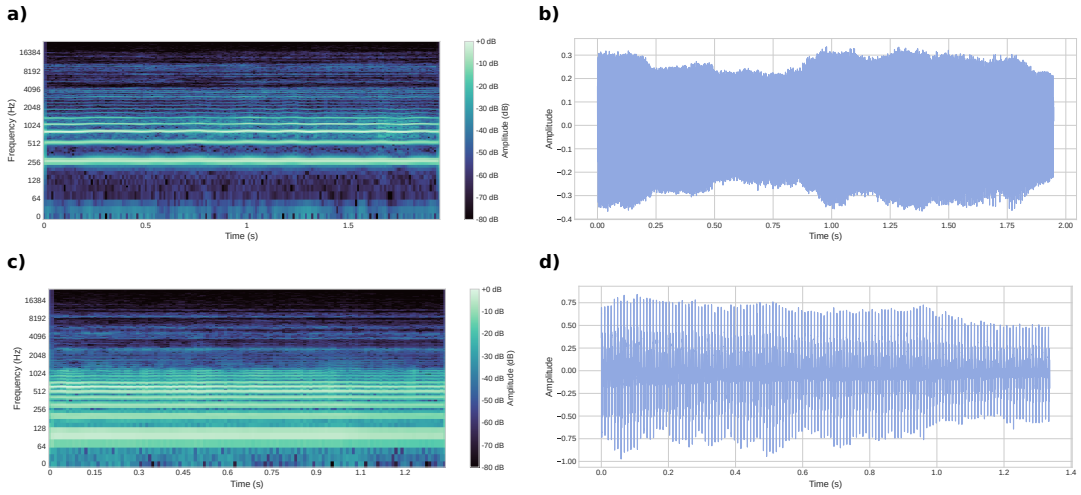


Figure 3: Examples of sustained /a/ at neutral pitch from female speakers in the SVD dataset. Panels (a)–(b) show spectrogram and waveform for a healthy control, and panels (c)–(d) show the corresponding signals for a pathological case diagnosed with laryngitis.

Table 1: Acoustic features in the extended feature set (Vrba+) extracted from the trimmed sustained vowel segment.

Group	Feature (symbol)	Key parameters		Ref.
Pitch	Fundamental frequency (f_0)	pYIN; $f_{\min} = 50$ Hz, $f_{\max} = 500$ Hz	μ_{f_0}, σ_{f_0}, min f_0, max f_0	[9]
Cepstral	MFCCs ($c_i,\ i = 1 \ldots 20$)	41 ms window, 10 ms hop; 128 mel bands; 13 coeffs	$\mu(c_i), \sigma(c_i)$	[10]
Cepstral	LFCCs ($l_i,\ i = 1 \ldots 20$)	linear-scale filterbank; FFT=2048, hop=512	$\mu(l_i)$	
Cepstral	Deltas (Δc_i, $\Delta^2 c_i$)	same as MFCCs	$\mu(\Delta c_i), \mu(\Delta^2 c_i)$	[10]
Cepstral	Cepstral peak prominence (CPP)	log-spectral cepstrum; quefrency for 60–300 Hz	prominence	[11]
Spectral shape	Centroid (C), bandwidth (BW), roll-off ($R_{0.85}$), zero-crossing rate (ZCR), RMS energy (E_{rms})	41/10 ms; roll-off at 85%	mean	
Spectral contrast	$\text{SC}_b, SCB_b,\ b = 1 \ldots 7$	7 octave bands	mean per band ($b = 1 \ldots 7$)	
Perturbation	Jitter (J-local, J-RAP, J-PPQ5); Shimmer (S-local, S-APQ3, S-APQ5)	Praat; pitch search 50–500 Hz	value	
Periodicity	Harmonics-to-noise ratio (HNR)	Praat harmonicity (cc)	mean	[12]
Time-domain	Skewness (SKEW)	raw waveform amplitude distribution	value	
Time-domain	Shannon entropy (SE)	$\alpha = 1$; raw waveform	value	

encodings; iii) elimination of highly correlated features, using Spearman correlation with an absolute cutoff of 0.9; iv) Min-max normalization of all remaining predictors to the $[0, 1]$ range. This preprocessing provides a uniform, fully numeric, non-redundant, and scale-consistent feature representation across all 120 tasks, enabling fair comparison among classifiers.

3.4 Algorithms

This study evaluates four off-the-shelf classification algorithms spanning ensemble learning, boosting, margin-based optimization, and distance-based methods. All models were trained using the default hyperparameter configurations provided by the `mlr3verse` framework. Their main characteristics are described as follows:

3.4.1 Random Forest. Random Forest is an ensemble method composed of multiple uncorrelated decision trees trained over bootstrap samples of the data [13] . At each split, only a random subset of features is considered, which lowers tree correlation and reduces variance without substantially increasing bias. This model is inherently capable of capturing nonlinear relationships and higher-order interactions among acoustic features. It is also robust to multicollinearity, noisy predictors, and high-dimensional inputs.

3.4.2 XGBoost. XGBoost is a gradient boosting algorithm that constructs trees sequentially, with each new tree modeling the pseudo-residuals of the previous ensemble [14] . Its formulation includes L1 and L2 regularization, shrinkage (learning rate), and stochastic subsampling of rows and columns, which collectively improve generalization and mitigate overfitting. Because of its ability to learn complex nonlinear relationships and feature interactions. Its regularization mechanisms also make it well suited for datasets with heterogeneous distributions and potentially redundant predictors.

3.4.3 Support Vector Machine (SVM). The SVM is a maximum-margin classifier that aims to find the hyperplane that best separates the classes while maximizing the geometric margin [15]. By leveraging kernel functions, SVM can implicitly map the input space into higher dimensions, enabling separation of otherwise nonlinearly separable patterns. This property makes SVM effective in high-dimensional acoustic representations, where clear class boundaries may emerge only after nonlinear transformations.

3.4.4 Generalized Linear Model (GLM). The GLM algorithm implements generalized linear models with elastic net regularization, combining both L1 and L2 penalties [16]. This approach performs variable selection (via L1) while controlling coefficient shrinkage (via L2), enabling stable estimation even in high-dimensional settings. In classification tasks, GLM fits a regularized logistic regression model capable of handling correlated predictors and mitigating overfitting.

3.5 Experimental Setup

All experiments were performed using stratified 10-fold cross-validation, preserving the proportion of healthy and pathological samples in each fold. To ensure full reproducibility, the entire resampling procedure was controlled by a global random seed (123).

The experimental pipeline was implemented in R, relying on the `mlr3verse` framework for task definition, model instantiation, resampling, and performance estimation. Auxiliary packages included `data.table` for high-efficiency data manipulation, `progressr` for progress monitoring, and `future` to enable parallel execution across folds and learners, thereby reducing total computation time. All classifiers produced probability-based predictions, which ensured consistency in the calculation of performance metrics.

In total, the study required 120 tasks × 6 algorithms × 10 folds = 7200 independent training–testing cycles, executed with parallelization through the `future` package. The benchmark was executed on a workstation equipped with 24 GB of RAM and 10 CPU threads. The full execution required approximately 3 hours.

Model performance was quantified using a set of complementary metrics balanced accuracy, F-score, area under the ROC curve (AUC), precision, and recall. All evaluation outputs, predictions, and metadata were stored for downstream analysis.

4 RESULTS

Figure 4 summarizes the classification performance, expressed in terms of AUC, across all combinations of vowels, pitch conditions, speaker sex, feature sets, and learning algorithms. Subplots are organized accordingly to vowels and speaker sex options. The x-axis lists all variations of pitches, while the y-axis shows the AUC performance values obtained by each algorithm and feature set combination. Different algorithms are represented by different colored lines, while different feature sets are represented by different point shapes.

First, pitch condition has a clear influence on performance. When all pitch types are pooled together ("all"), the models tend to achieve higher AUC values compared to single-pitch subsets. This was expected, since combining multiple recordings increases the amount of acoustic information available to the classifiers. However, this improvement comes at a usability cost: requiring the patient to produce four recordings instead of one increases acquisition time and may reduce the practicality of real-world screening systems.

A similar trend appears when comparing single-vowel inputs with the "all-vowels" condition. Although combining */a/*, */i/*, and */u/* slightly increases performance, the gain is modest; for several scenarios, using a single vowel results in AUC values comparable to the multi-vowel condition. This suggests that vowel choice alone does not significantly alter the discriminability of pathological versus healthy phonation in this dataset.

Regarding single-pitch scenarios, models trained on low pitch (and in some cases on low-high-low intonation) tend to present better performance than those trained only on high pitch. The high-pitch condition consistently produces the lowest AUC values, especially in the male-only subset. This suggests that high-pitch phonation may compromise the reliability of certain acoustic features, particularly for male speakers. Moreover, pitch interacts with vowel and sex: certain combinations show clear dependencies. For example, the vowel */a/* at normal pitch produces strong performance for female speakers, while the corresponding male subset is more sensitive to pitch variation.

Regarding speaker sex, differences between male-only, female-only, and mixed-sex datasets are generally small. This suggests that the feature representations are robust with respect to gender distribution, although isolated scenarios show slight sex-specific effects—mostly when pitch interacts with vowel choice.

4.1 Best setups overall

We applied the Friedman test [17], with significance levels at $\alpha = 0.05$ to evaluate the statistical significance of the experimental results. The null hypothesis states that all groups of settings are

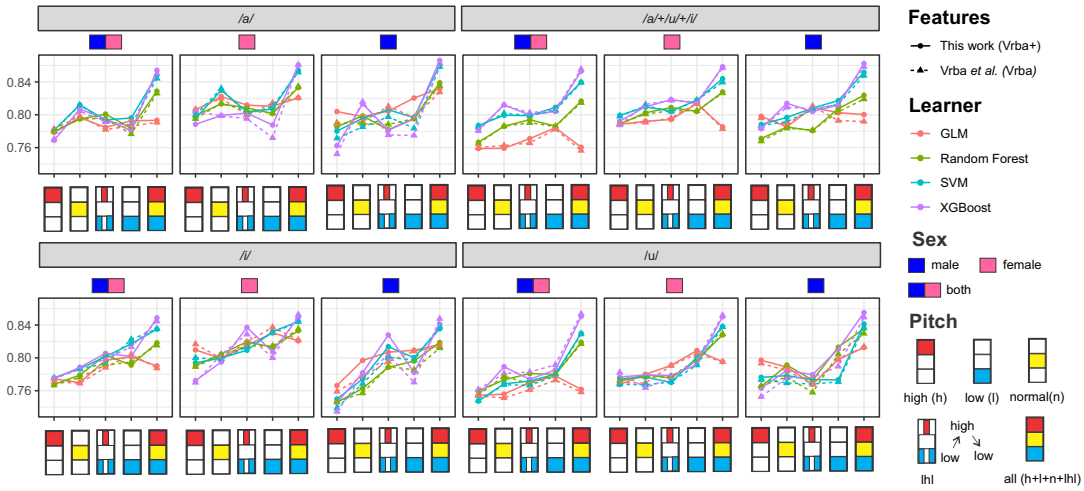


Figure 4: Performance results (AUC) for each combination of sex, pitch, vowel, feature set, and classification method.

equivalent in terms of predictive AUC performance. If the null hypothesis was rejected, the Nemenyi post-hoc test was applied, stating that the performances of two different groups of techniques are significantly different if the corresponding average ranks differ by at least a CD value. Figure 5 presents CD diagrams for all the scenarios: varying algorithms, feature sets, vowels, and pitches.

The diagram shows that XGBoost and SVM consistently achieve the best ranks and are not statistically different from each other. This reinforces that non-linear classifiers with good capacity to model complex decision boundaries are particularly suitable for the highly overlapping distributions.

The same analysis indicates that our extended Vrba+ feature set is statistically superior to the original Vrba baseline, even though the absolute AUC gains are modest. The improvement is systematic across tasks, showing that the additional features contribute to the prediction performance.

Regarding the vocal task, the ranks for vowels reveal that using only the sustained vowel */a/* is as effective as using all three vowels together. This means that asking the patient to produce a single */a/* token is sufficient to reach our best-performing configurations. This simplifies the protocol, reduces recording and processing time, and improves usability in realistic scenarios, such as telemedicine or mobile applications.

Pitch conditions have a more significant impact. Combining all four pitches yields the best overall rank, as expected when the model can integrate information from neutral, low, high, and low–high–low frequencies. However, this requires four recordings per subject. When only one pitch can be acquired, low and low–high–low emerge as the most reliable options, while high pitch is consistently the worst, especially for male speakers. This pattern suggests that low phonation preserves important acoustic evidence of pathology.

Sex also plays a relevant role: female-only models obtain the best ranks, followed by male-only and, lastly, mixed-sex models. Training a single model on data from both males and females forces the classifier to learn a larger variability in fundamental frequency and formant structure, which appears to negatively impact performance. In practice, our results suggest that sex-specific models are a suitable approach.

4.2 Best models for each vowel

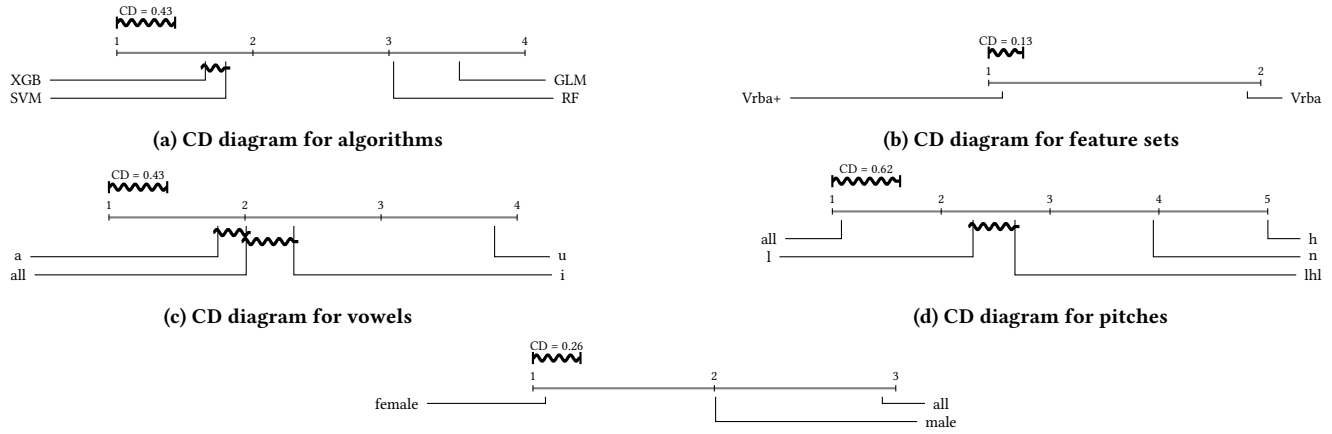
Table 2 summarizes the best-performing configuration for each vowel-related subtask. For */a/*, */u/*, and the all-vowels condition, the top setup corresponds to the extended Vrba+ feature set combined with all pitch conditions, restricted to male speakers and classified by XGBoost, reaching AUC values between 0.855 and 0.866. For the */i/* configuration, the best model also uses all pitch conditions but relies on the original Vrba feature set and is trained on female speakers, achieving an AUC of 0.852 with XGBoost.

All best setups considered all pitch conditions, confirming that combining neutral, low, high, and low–high–low records provides richer acoustic information than any single pitch alone. Furthermore, the prevalence of Vrba+ in three out of four subtasks shows that the additional spectral and perturbation descriptors bring consistent gains, while the exception in the */i/*. Also, all top models are sex-specific, indicating that separating male and female speakers simplifies the decision boundary and improves performance. Finally, XGBoost appears in every best configuration, confirming its robustness across vowels and subsets and supporting it as a strong default choice for deploying binary voice-pathology screening models based on handcrafted acoustic features.

The confusion matrices of the best XGBoost models in Figure 6 show the distribution of predictions in relation to the true label. In

Table 2: Best setups for each vowel subtask

Vowel	Features	Pitch	Sex	Algo	AUC
a	Vrba+	all	male	XGB	0.866
all	Vrba+	all	male	XGB	0.862
u	Vrba+	all	male	XGB	0.855
i	Vrba	all	female	XGB	0.852



(a) CD diagram for algorithms

(b) CD diagram for feature sets

(c) CD diagram for vowels

(d) CD diagram for pitches

(e) CD diagram for sex

Figure 5: Comparison of the AUC values of the ML algorithms, feature sets, vowels, and pitches according to the Friedman-Nemenyi test. Groups that are not significantly different are connected.

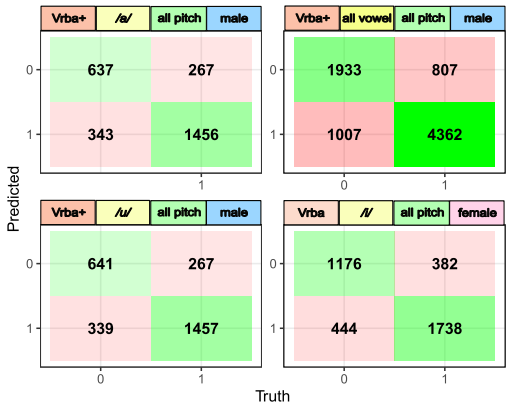


Figure 6: Confusion matrices for the best-performing XGBoost models in each vowel configuration.

the Vrba+ male tasks (*/a/*, */u/*, and the all-vowels), the dominant error mode is false positives, i.e., healthy samples (0) being predicted as pathological (1). In contrast, pathological samples (1) are more often mapped to the correct class, indicating a conservative decision-making tendency compatible with screening-oriented use cases, where flagging potentially altered phonation is preferable to missing pathology. The confusion patterns for */a/* and */u/* are nearly identical, reinforcing that vowel choice has limited impact on the decision behavior once pitch variability is aggregated. For the /i/ configuration (female subset), the matrix shows a comparatively more balanced confusion pattern, suggesting a cleaner separation between classes at the operating point induced by the default threshold. Finally, because the subtasks are not perfectly class-balanced and differ in sample size, we interpret Figure 6 primarily in terms of error directionality and consistency across configurations, while AUC remains the main criterion for ranking performance. Finally, model choice plays a considerable role. XGBoost and SVM models consistently achieve the highest AUC values across most conditions.

4.3 Unveiling feature importance

The feature-importance analysis in Figure 7 helps to understand why Vrba+ performed better than the baseline. Each plot depicts the relative feature importance (mean gain) assigned by XGBoost to the input variables in the top-performing task for each vowel configuration (*/a/*, */i/*, */u/*, and all vowels combined). Among the additional descriptors, the spectral bandwidth mean (BW) appears systematically among the top features in the best XGBoost models. This feature measures the average spread of spectral energy around the centroid, thereby capturing how concentrated the harmonic and noise components are over frequency. Pathological voices tend to exhibit broadened spectra due to irregular vibration, making spectral bandwidth an indicator of deviation from normal phonation. Jitter and shimmer-based measures also rank highly, confirming that short-term cycle-to-cycle instability is related to pathology. Therefore, Figure 7 indicates that Vrba+ improves over the baseline not by adding many redundant variables, but by introducing a small set of complementary features.

It is important to note that these results may be influenced by dataset-specific characteristics, which may limit external validity when transferring the models to other clinical populations or recording conditions. In particular, variations in microphone setup, environment, or pathology distribution could affect the stability of the reported acoustic patterns. Moreover, all models were trained with default hyperparameters. Although this choice ensures a fair comparison between methods, further tuning could improve the absolute performance values.

5 CONCLUSIONS

This work investigated how vowel choice, pitch condition, speaker sex, and feature design affect the performance of machine-learning models for voice pathology detection on the SVD corpus. Overall,

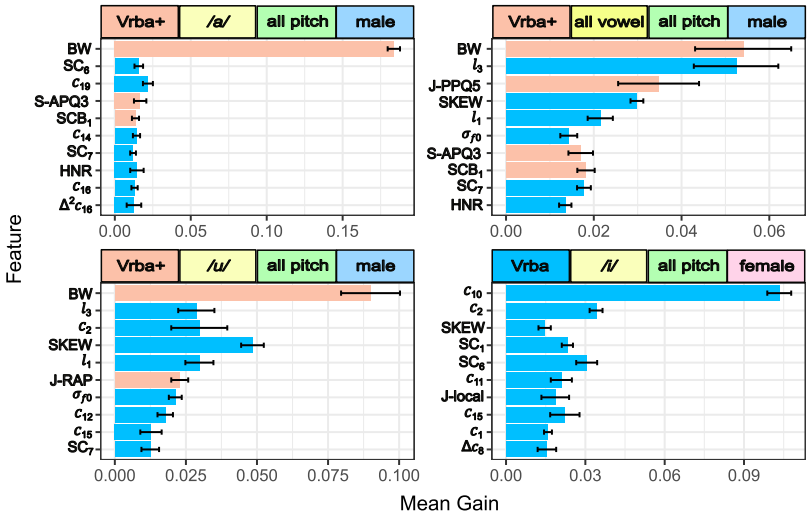


Figure 7: XGBoost relative feature importance (Mean Gain) in the top-tasks for each specific vowel configuration.

the AUC values we obtained (median around 0.80, with the best results reaching 0.866) are comparable to those of recent ML-based studies using the SVD database, which also rely on handcrafted acoustic features.

Our study offers a reproducible benchmark on SVD with systematically varied vowels, pitch conditions, and sex, and shows that classical, interpretable acoustic features combined with off-the-shelf ML algorithms can deliver competitive performance and straightforward design guidelines for practical screening tools. Nonetheless, the work is limited to a single corpus, binary healthy/pathological labels, and sustained vowels only, and does not include a direct comparison with end-to-end deep learning models.

Possible extensions for a long-form study include: improving learning tasks preprocessing, performing one-hot encoding for categorical variables; hyperparameter tuning for the most promising algorithms; feature selection procedures for the best learning tasks; and considering data balancing techniques, such as SMOTE. Moreover, the use of comprehensive time-series feature libraries may enable a more complete search for discriminative descriptors. Additionally, evaluating deep features extracted from state-of-the-art neural architectures is a natural possibility, yet it has not been explored for this dataset. Furthermore, future research should extend this analysis to multiclass diagnoses, additional datasets and speech tasks.

REFERENCES

[1] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2: 100074, 2022. doi: 10.1016/j.cmpbup.2022.100074.

[2] Nurul Mu'azzah Abdul Latiff, Fahad Taha Al-Dhief, Nurul Fariesya Suhaila Md Sazihan, Marina Mat Baki, Nik Noordini Nik Abd. Malik, Musatafa Abbas Abbood Albadr, and Ali Hashim Abbas. Voice pathology detection using machine learning algorithms based on different voice databases. *Results in Engineering*, 25:103937, 2025. doi: 10.1016/j.rineng.2025.103937.

[3] Rab Nawaz Bashir, Muhammad Ali Shahid, Tahir Rashid, Muhammad Faheem, Taoufik Saidani, Oumaima Saidani, and Amjad Rehman Khan. Voice pathology identification using mel spectrogram features and deep learning. *Signal, Image and Video Processing*, 19:909, 2025. doi: 10.1007/s11760-025-04527-4.

[4] Aymen Ksibi, Nasser A. Hakami, Naif Alturki, Mohammed M. Asiri, Mohammed Zakariah, and Mohamed Ayadi. Voice pathology detection using a two-level classifier based on combined cnn–rnn architecture. *Sustainability*, 15(4):3204, 2023. doi: 10.3390/su15043204.

[5] Sozan Abdullah Mahmood. Multi-dimensional features extraction for voice pathology detection based on deep learning methods. *Journal of Voice*, 2025. doi: 10.1016/j.jvoice.2024.12.048. Accepted December 30, 2024.

[6] Mark Huckvale and Catinca Buciuleac. Automated detection of voice disorder in the saarbrücken voice database: Effects of pathology subset and audio materials. In *Proc. Interspeech 2021*, pages 1559–1563, 2021. doi: 10.21437/Interspeech.2021-1507. URL https://discovery.ucl.ac.uk/id/eprint/10139814/.

[7] Jan Vrba, Jakub Steinbach, Tomáš Jirsa, Laura Verde, Roberta De Fazio, Yuwen Zeng, Kei Ichiji, Lukáš Hájek, Zuzana Sedláková, Zuzana Urbániová, Martin Chovanec, Jan Mareš, and Noriyasu Homma. Reproducible machine learning-based voice pathology detection: Introducing the pitch difference feature. *Journal of Voice*, 2025. ISSN 0892-1997. doi: 10.1016/j.jvoice.2025.03.028. Online ahead of print.

[8] Manfred Pützer and William J. Barry. Instrumental dimensioning of normal and pathological phonation using acoustic measurements. *Clinical Linguistics & Phonetics*, 22(6):407–420, 2008. doi: 10.1080/02699200701830869.

[9] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. IEEE ICASSP*, pages 659–663, 2014. doi: 10.1109/ICASSP.2014.6853678. URL https://www.eecs.qmul.ac.uk/~simond/pub/2014/MauchDixon-PYIN-ICASSP2014.pdf.

[10] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980. doi: 10.1109/TASSP.1980.1163420.

[11] Orlik Murton, Robert E. Hillman, and Daryush D. Mehta. Cepstral peak prominence values for clinical voice assessment. *American Journal of Speech-Language Pathology*, 29(3):1596–1607, 2020. doi: 10.1044/2020_AJSLP-20-00001.

[12] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, Univ. of Amsterdam*, 17:97–110, 1993.

[13] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

[14] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018.

[16] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

[17] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.