

1

2

3

4 Identifying bias in models that detect vocal fold paralysis from audio
5 recordings using explainable machine learning and clinician ratings

6 Daniel M. Low ^{1,2}, Vishwanatha Rao ^{3,4}, Gregory Randolph ^{4,5}, Phillip C. Song ^{4,5 *},
7 Satrajit S. Ghosh ^{1,2,5 *}

8

9

10

11 ¹ Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, Boston,

12 MA, USA

13 ² McGovern Institute for Brain Research, MIT, Cambridge, MA, USA

14 ³ Department of Biomedical Engineering, Columbia University, New York, NY, USA

15 ⁴ Department of Otolaryngology–Head and Neck Surgery, Massachusetts Eye and Ear Infirmary,

16 Boston, MA, USA

17 ⁵ Department of Otolaryngology–Head and Neck Surgery, Harvard Medical School, Boston, MA,

18 USA

19 * Equal contribution

20 Corresponding author

21 Correspondence can be addressed to Daniel M. Low, Office: 46-4033F, 43 Vassar St,

22 Cambridge, MA 02139, USA. E-mail: dlow@mit.edu.

23

24

25

26

Abstract

27 **Introduction.** Detecting voice disorders from voice recordings could allow for frequent, remote,
28 and low-cost screening before costly clinical visits and a more invasive laryngoscopy
29 examination. Our goals were to detect unilateral vocal fold paralysis (UVFP) from voice
30 recordings using machine learning, to identify which acoustic variables were important for
31 prediction to increase trust, and to determine model performance relative to clinician
32 performance.

33 **Methods.** Patients with confirmed UVFP through endoscopic examination (N=77) and controls
34 with normal voices matched for age and sex (N=77) were included. Voice samples were elicited
35 by reading the Rainbow Passage and sustaining phonation of the vowel "a". Four machine
36 learning models of differing complexity were used. SHapley Additive exPlanations (SHAP) was
37 used to identify important features.

38 **Results.** The highest median bootstrapped ROC AUC score was 0.87 and beat clinician's
39 performance (range: 0.74 – 0.81) based on the recordings. Recording durations were different
40 between UVFP recordings and controls due to how that data was originally processed when
41 storing, which we can show can classify both groups. And counterintuitively, many UVFP
42 recordings had higher intensity than controls, when UVFP patients tend to have weaker voices,
43 revealing a dataset-specific bias which we mitigate in an additional analysis.

44 **Conclusion.** We demonstrate that recording biases in audio duration and intensity created
45 dataset-specific differences between patients and controls, which models used to improve
46 classification. Furthermore, clinician's ratings provide further evidence that patients were

47 over-projecting their voices and being recorded at a higher amplitude signal than controls.

48 Interestingly, after matching audio duration and removing variables associated with intensity in

49 order to mitigate the biases, the models were able to achieve a similar high performance. We

50 provide a set of recommendations to avoid bias when building and evaluating machine learning

51 models for screening in laryngology.

52 **Keywords:** vocal fold paralysis, acoustic analysis, voice, speech, explainability, interpretability,

53 machine learning, bias

54

55

56

57

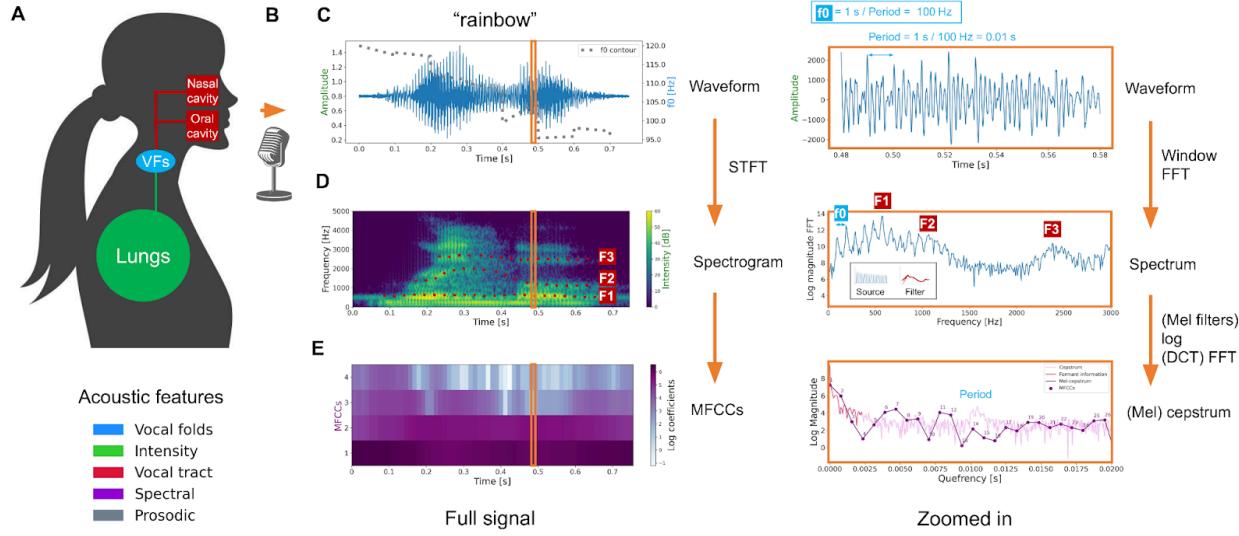
58 INTRODUCTION

59 Voice recordings provide a rich source of information related to vocal tract physiology
60 and human physical and mental health. Given advances in smartphones and
61 wearables, these recordings can be made anytime and anywhere. Thus the search for
62 disorder-specific acoustic biomarkers has been gaining momentum. Voice biomarkers
63 have been reported for detecting Parkinson's disease (1) as well as psychiatric
64 disorders including depression, schizophrenia, and bipolar disorder (for a systematic
65 review, see Low et al, 2020 (2)). Given our scientific understanding of the complexity of
66 speech production, multiple acoustic features have been devised for use in machine
67 learning models. In Figure 1, we describe a schematic of speech production and the
68 process of extracting certain acoustic features from an audio signal (see also Quatieri,
69 2008 (3)), which is an important part of explaining how pathophysiology would affect
70 acoustic features that are used in machine learning classifiers. Panel (A) depicts speech
71 as the result of the neural coordination of three subsystems: the respiratory system
72 (lungs), the laryngeal system (vocal folds), and the resonatory system of the vocal tract
73 (pharynx, oral cavity, nasal cavity, articulators, and subglottal effects). Speech
74 production requires air flow from the lungs to generate sound sources that are filtered
75 by the vocal tract. Panel (B) captures the fact that environmental, microphone, and
76 digital sampling characteristics (e.g., background noise, microphone gain, sampling
77 rate) can affect acoustic features. Panel (C) shows the waveform of the audio signal,
78 representing areas of compression (positive amplitude; higher air pressure) and

79 rarefaction (negative amplitude; lower air pressure). Higher amplitudes can lead to
80 higher perceived loudness. Prosodic features arise from changes over longer segments
81 of time, which is perceived in the rhythm, stress, and intonation of speech. A segment of
82 the waveform is shown in the right panel, indicating a periodic signal from the vocal
83 folds. Panel (D) shows that for a given time window, a spectrum (right panel) can be
84 obtained through a fast Fourier transform (FFT) which represents the magnitude of the
85 frequencies in the signal with peaks (formants F1–F3) due to vocal tract filtering of the
86 source signal produced by the vocal folds. The spectrogram (left panel) is a
87 representation of the spectrum as it varies over time and can be obtained through a
88 short-term Fourier transform (STFT). The approximate location of the F0 and first
89 formants are displayed. Finally, (E) It is possible to separate source and filter
90 components by computing the inverse FFT of the log of the magnitude of the spectrum,
91 called the cepstrum (right panel). The peak in the cepstrum reflects the periodic glottal
92 fold vibration while lower quefrency components reflect properties of the resonatory
93 subsystem. For speech recognition, Mel filters are applied to the spectrum to better
94 approximate human hearing. A conversion of the Mel-spectrum to a cepstrum using a
95 Discrete Cosine Transform (DCT) generates mel-frequency cepstral coefficients
96 (MFCCs). Similar to the cepstrum, lower MFCCs track vocal-tract filter information.

97

98



99

100 **Figure 1. Schematic of speech production and the process of extracting certain acoustic features**
101 **from an audio signal.**

102 (A) Speech production, (B) recording characteristics, (C) waveform of audio signal with fundamental
103 frequency (f0), (D) spectrogram with formants F1-F3 and intensity, (E) mel-frequency cepstral coefficients
104 (MFCCs). Full description in the main text.

105 Furthermore, while machine learning (ML) can be a powerful and successful approach

106 for diagnostics, they are often treated as "black-boxes". It can be difficult to determine

107 how the model is making a decision, that is, how it is combining input features from a

108 given patient to generate a prediction. This is particularly worrisome given ML

109 algorithms can detect and associate unintended or clinically irrelevant relationships and

110 introduce bias that may be difficult to anticipate. Explainable ML refers to a series of

111 methods and quantitative analyses for uncovering and "explaining" the rationale behind

112 the decision made by complex algorithms, which is particularly critical in the high-stake

113 decisions of medicine to increase trust among clinicians and patients (4).

114 There are many challenges for applying acoustic analysis to detect specific disorders.

115 Voice characteristics are highly varied and change over time. Laryngeal pathology, age,

¹¹⁶gender, size, weight, general state of health, smoking/vaping, and medications can
¹¹⁷impact vocal acoustic characteristics. Diseases in the larynx and phonatory system (i.e.,
¹¹⁸larynx, resonating structures, lungs) and/or neurological system, will also affect voice.
¹¹⁹Compensatory production strategies and environmental conditions can also change the
¹²⁰vocal signal. Furthermore, because hoarseness is such a frequent occurrence and
¹²¹specialty voice centers are rare, vocal fold disorders are often undiagnosed,
¹²²under-reported, or misdiagnosed (5).

¹²³We chose vocal fold paralysis as the study cohort for several reasons. First, it is
¹²⁴clinically important. UVFP can have detrimental effects on voice and quality of life with
¹²⁵resultant morbidity related to respiration, swallowing and aspiration (6). Vocal fold
¹²⁶paralysis may occur due to iatrogenic injury, malignancy, idiopathic, and neurological
¹²⁷disease (7). Overall, surgical iatrogenic injury accounts for 46% of all UVFP in adults
¹²⁸and thyroid and parathyroid surgeries are responsible for 32% of postsurgical UVFP (8).

¹²⁹There is a significant need for a screening tool for the diagnosis and tracking of UVFP
¹³⁰because of the high impact of this condition on productivity and quality of life. Screening
¹³¹could be done remotely and frequently, especially when surgical specialists and
¹³²laryngeal exams are not readily accessible due to geographical, financial, and other
¹³³barriers (9). Using an explainable ML model as a screening tool for UVFP can provide
¹³⁴greater clarity as to who most needs laryngoscopy and provides insight in the key voice
¹³⁵characteristics related to the pathophysiology (10–14). The costs associated with UVFP
¹³⁶not only relate to patient morbidity and diminished quality of life but also to the economic
¹³⁷burden placed on our healthcare system. Greater lengths of hospitalization and

138 increased hospital costs have been associated with postsurgical VFP (15,16). Access to
139 specialists for diagnosis is limited and early detection and management of UVFP appear
140 to improve length of stay and surgical outcomes (17). Special consideration should be
141 given to what the model can actually classify: a model that generalizes well in
142 classifying UVFP from controls may not be able to screen for UVFP out of other voice
143 disorders, but could be used to monitor UVFP patients remotely and affordably during
144 treatment or detect risk for UVFP when it is the most likely cause such as dysphonia
145 after thyroid surgery.

146 Furthermore, UVFP is an ideal model for demonstrating the explainability of ML. UVFP
147 occurs when the mobility of a single vocal fold is impaired as a consequence of
148 neurological injury and diagnosis is consistently verified through routine laryngoscopy;
149 therefore, ground truth labels are available. Second, the clinical signs of UVFP are
150 well-described. These characteristics include a weak, breathy voice quality, early vocal
151 fatigue, reduced cough strength, and aspiration with thin liquids (18,19). Therefore, the
152 acoustic differences between UVFP patients and healthy controls can be interpreted
153 with regards to perceptual symptoms and a well-understood pathophysiology. In
154 contrast, explaining important variables to predict a disorder which is hard to diagnose
155 (e.g., has low inter-rater reliability) and has an unclear pathophysiology would ironically
156 result in a poor explanation, because it would be puzzling how or even if the disorder
157 could modulate the important acoustic variables. Of course, machine learning models
158 can also offer novel explanations into a disorder by characterizing novel characteristics.
159 However, if these models use high-dimensional feature vectors, they are more likely to

¹⁶⁰overfit when using small datasets (20,21), which should lead to more skepticism of
¹⁶¹these novel explanations.

¹⁶²There have been several studies detecting unilateral vocal fold paralysis (UVFP) using
¹⁶³machine learning (22–30); however, most have included the disorder among a set of
¹⁶⁴voice disorders to be predicted. Limitations of these prior studies could be seen to fall
¹⁶⁵into one of following types: not reporting the performance when classifying the subset of
¹⁶⁶participants with UVFP out of the participants with dysphonia they were trying to detect;
¹⁶⁷small sample sizes given most studies contained 10 participants with UVFP or fewer
¹⁶⁸with one study containing 50 participants (31); a lack of algorithmic explanations: they
¹⁶⁹either do not report on the relative importance of each acoustic variable; use input data
¹⁷⁰such as a spectrogram in a black-box deep learning model which could make attempts
¹⁷¹at algorithmic explanations on images such as saliency maps more opaque than results
¹⁷²from feature importance of handcrafted features; use a black-box model such as neural
¹⁷³network without attempting to explain its predictions with deep learning explainability
¹⁷⁴methods (32); use a single type of model which may pick up on certain types of patterns
¹⁷⁵but miss others leading to incomplete conclusions on feature importance; use only a few
¹⁷⁶features which may impede better predictive performance by not capturing certain
¹⁷⁷relevant information; and/or not publicly share models or data to help test their
¹⁷⁸generalizability to new data.

¹⁷⁹The objectives of our study were: to detect UVFP using ML; to evaluate the
¹⁸⁰effectiveness of different models in differentiating the acoustic signals between patients

181 with UVFP and patients with normal functioning vocal folds (i.e., controls); to explain
182 which features are most important to the diagnostic models and examine the
183 pathophysiological relevance; and to compare performance to human clinicians
184 evaluating audio recordings. To achieve these objectives, we evaluated four different
185 classes of machine learning algorithms to assess classification performance, obtained
186 the minimal set of features necessary for detection, and identified the most important
187 acoustic features for model construction after removing redundant features. Ultimately,
188 we wanted to see if the most important features identified by the machine learning
189 models matched clinically-known relevant acoustic changes.

190

191 MATERIALS AND METHODS

192 This study was approved by the Institutional Review Board at Massachusetts Eye and
193 Ear Infirmary and Partners Healthcare (IRB 2019002711).

194 Participants and voice samples

195 Through retrospective chart analysis from 2009 to 2019, a total of 1043 patient charts
196 were reviewed from a tertiary care laryngology practice who underwent endoscopic
197 evaluation and voice testing. Of those, 53 patients with confirmed UVFP were identified.
198 They had documented vocal fold paralysis by endoscopic examination and had
199 undergone acoustic analysis as part of routine clinical care. Each patient had four

200 acoustic recordings. These included three sustained vocalizations of the "a" vowel
201 sound (a in the International Phonetic Alphabet) and a reading of the introductory
202 paragraph of the rainbow passage (33). The acoustic recordings were all taken in an
203 acoustically shielded room. For each of these 53 patients, a board-certified
204 otolaryngologist reviewed their clinical history, video laryngoscopy as well as their audio
205 samples to confirm that they were correctly classified to have UVFP. Voice samples
206 from an additional 24 patients were collected prospectively using a mobile software,
207 OperaVOX™ on an iPad, who were being treated for UVFP. These patients also had
208 the same four acoustic recordings as the patients from retrospective chart review. This
209 combination of data collection yielded a total of 77 UVFP patients for analysis, of which
210 48 had left UVFP and 29 right UVFP.

211 All of the patients were then matched with control samples from a database of patients
212 without UVFP who had also undergone acoustic analysis. Each control was the same
213 sex and had the same smoking status as the UVFP patient and within three years of
214 age, and had documented laryngeal examinations that verified the absence of vocal fold
215 mucosal pathology. The controls were excluded if they had established laryngeal
216 surgery, vocal fold lesions, radiation, head and neck cancer, or neurological disease.
217 The controls had recorded the same four acoustic recordings as the retrospectively
218 gathered UVFP group. A board-certified otolaryngologist confirmed that the voice
219 recordings and video laryngoscopies of these controls matched normal expectancies.
220 The reading samples were divided in thirds to match the amount of vowel production
221 samples, resulting in 6 samples for most participants. Reading recordings were not

222 available for three patients and three patient vowel samples were removed due to
223 containing multiple vowel productions or a cough. The final dataset that was analyzed is
224 described in Table 1. Reading+vowel refers to including all samples (i.e., ~6 samples)
225 from the same participant with the goal of either obtaining higher performance or
226 discovering features that show variation in relation to diagnosis consistently across
227 tasks. Mean (SD) audio lengths were 6.81s (5.47) for reading samples and 3.95s (1.00)
228 for vowel samples. The audio samples were processed using OpenSmile with the
229 eGeMAPS configuration file (article (34), source code (35)) which applies different
230 summarization statistics to the time series depending on the feature resulting in 88
231 features per sample covering information related to the vocal folds (F0, jitter, shimmer),
232 intensity (loudness, HNR), vocal tract (F1–3 frequency, bandwidth, amplitude), spectral
233 balance (alpha ratio, Hammamberg index, spectral slope, MFCC 1–4, spectral flux), and
234 prosody (voice and unvoiced segments, loudness peaks per second). See section
235 "eGeMAPS features" in Sup. Mat. for full list.

236 **Table 1. Sample sizes and demographic information**

	UVFP	Controls	Total
N	77	77	154
Mean age (SD)	56.4 (18.7)	56.6 (18.8)	56.5 (18.7)
Sex (F/M)	39/38	39/38	78/76
Reading	222	231	453
Vowel	227	231	458
Reading+vowel (total)	449	462	911

237 SD: standard deviation; F: female; M: male.

238 Machine learning models of increasing complexity

239 With the goal of classifying voices recording into either UVFP or controls, we used four

240 machine learning algorithms of increasing complexity from the *scikit-learn* (v0.21.3)

241 using the *pydra-ml* (v0.3.1) toolbox (36) (default parameters were used unless

242 otherwise specified). By complexity we mean models are more complex if they are

243 harder to simulate, that is, harder to take the input data and model parameters and step

244 through every calculation required to produce a prediction in a reasonable time which

245 increases with the amount of parameters and interactions (37).

246 (1) Logistic Regression: a simple linear model that is constrained to use few features

247 due to an L1 penalty making it the simplest model ("liblinear" solver was used which is

248 ideal for smaller datasets).

249 (2) Stochastic Gradient Descent (SGD) Classifier: we used a log loss which implements

250 a logistic regression; therefore, it is also a linear model but tends to use more features

251 due to an elastic net penalty, making it slightly more complex (the max_iter parameter

252 was set to 5000 and early_stopping was set to True).

253 (3) Random Forest: it is an algorithm that uses simpler decision trees (i.e., weak

254 learners) on feature subsets "but then takes the majority of the votes of the decision

255 trees' predictions to create a stronger learner, making it harder to interpret which

256 features are important across trees.

257 (4) Multi-Layer Perceptron: it is a neural network classifier which incorporates, in our
258 case, 100 instances of perceptrons (artificial neurons), which are connected to each
259 input feature through weights with a ReLU activation function to capture nonlinear
260 relationships in the data. It is not possible to know exactly how the hundreds of internal
261 weights interact to determine feature importance, making the model difficult to interpret
262 directly from its parameters (the max_iter parameter was set to 1000; alpha or the L2
263 penalty parameter was set to 1).

264 To generate independent test and train data splits, a bootstrapped group shuffle split
265 sampling scheme was used. Bootstrapping is more optimal than cross-validation on
266 smaller datasets and provides a measure of uncertainty through a confidence interval
267 (38). For each iteration of bootstrapping, a random selection of 20% of the participants,
268 balanced between the two groups, was used to create a held-out test set. The
269 remaining 80% of participants were used for training. This process was repeated 50
270 times, and the four classifiers were fitted and tested for each test/train split.. We used
271 the default of 50 bootstrapping splits from pydra-ml to reduce computational time.

272 Median ROC AUC stabilized to larger spit values at around 40 splits for logistic
273 regression models across tasks (see Sup. Mat. Figure S1) while reducing runtime
274 compared to larger split values. The Area Under the Receiver Operating Characteristic
275 Curve (ROC AUC; perfect classification = 1; chance = 0.5) was computed to evaluate
276 the performance of the models on each bootstrapping iteration, resulting in a distribution
277 of 50 ROC AUC scores for each classifier. To ensure results were not due to choosing
278 scikit-learn's hyperparameter default settings, hyperparameter tuning was performed on

279 the main models using all features and achieved similar performance to non-fine-tuned
280 models (see Sup. Mat. Table S1). The focus of our study is identifying bias and not
281 achieving –in our case– a small increment in performance; therefore, given the large
282 number of models, analyses, and bootstrapping samples in our study which focuses on
283 identifying bias, we chose default parameters given the small changes in performance
284 we observed with hyperparameter tuning. Additionally, for each iteration, each classifier
285 was trained with randomized patient/control labelings to generate a null distribution of
286 ROC AUC scores (i.e., a permutation test). Each model's performance was statistically
287 compared to their null model's distribution using an empirical p-value, a common and
288 effective measure for evaluating classifier performance (see Definition 1 in (39)). The
289 significance level was set to alpha = 0.05.

290 **Assessing feature importance**

291 Kernel SHAP (SHapley Additive exPlanations) was used to determine which acoustic
292 features were most important for each model to detect UVFP. This method is model
293 agnostic in that it can take any trained target model (even “black box” neural networks)
294 and compute feature importance (40). It does so by performing regression with L1
295 penalty between different sets of input features and a single prediction made by the
296 target model. It then uses the coefficients of the additional regression model as a
297 measure of feature importance for a single prediction. We took the average of the
298 absolute SHAP values across all test predictions (positive and negative values are both
299 important for classification). We then weighted the average values by the model's

300 median performance since an important feature for a bad model could be a less
301 important feature for a good model and vice versa. Since we trained each model 50
302 times (i.e., one for each bootstrapping split), we computed the mean SHAP values
303 across splits for each model. This pipeline (i.e., machine learning models, bootstrapping
304 scheme, SHAP analysis) was done using *pydra-ml*.

305 Reducing collinearity to do explainability analysis using

306 Independence Factor

307 Highly correlated features (i.e., collinearity) can influence model generation and
308 interpretation. Two models may obtain similar performance while using different features
309 or placing different weights on the same features (i.e., underspecification (20,41)). This
310 makes it difficult to compare algorithmic explanations across models. For instance,
311 mean F1 frequency may be less important to a given model because the model uses
312 mean F2 frequency which happens to capture very similar information in a particular
313 dataset (i.e., has a high correlation), whereas a different model may use F1 instead of
314 F2 or use both but assign less importance to each and still obtain the same
315 performance. To enforce models to use the same features that capture very similar
316 information and be able to compare feature importance across models, we kept a single
317 feature out of the sets of features that share similar information above a given threshold.

318 We used a custom algorithm we call Independence Factor whereby for each
319 feature in alphabetical (i.e., arbitrary) order, we removed features that show strong

320 dependence above a given threshold. The step was repeated for remaining features.

321 We use distance correlation from the Python *dcor* package (v0.4) because, unlike

322 Pearson *r* or Spearman *rho*, it can capture non-monotonic relationships (42,43). We

323 have included several examples of non-monotonic associations between variables in

324 our dataset that would be captured better by *dcor* (see Sup. Mat. Figure S2). We used

325 the following threshold values for the distance correlation [1.0, 0.9, 0.8, 0.7, 0.6, 0.5,

326 0.4, 0.3, 0.2] to compute the Independence Factor, which removed increasingly more

327 features (i.e., 1.0 keeps all features and 0.2 removes features that have a distance

328 correlation above 0.2). We chose the feature size which contains at least one model

329 that scores within three percentage points of the performance using all features, with

330 the goal of obtaining a more parsimonious model for subsequent explanation while

331 maintaining high accuracy. Thus, removing redundant features makes the models

332 easier to interpret for clinical relevance. To visualize the original redundancy across

333 features, we computed clustermaps using *seaborn* package (v0.10.1) performing

334 hierarchical clustering with the average-linkage method and Euclidean distance. This

335 was performed on the pairwise distance correlation, computed separately on data from

336 UVFP, controls, UVFP+controls and on reading, vowel, and reading+vowel.

337

338 **Performance using most important and least important features**

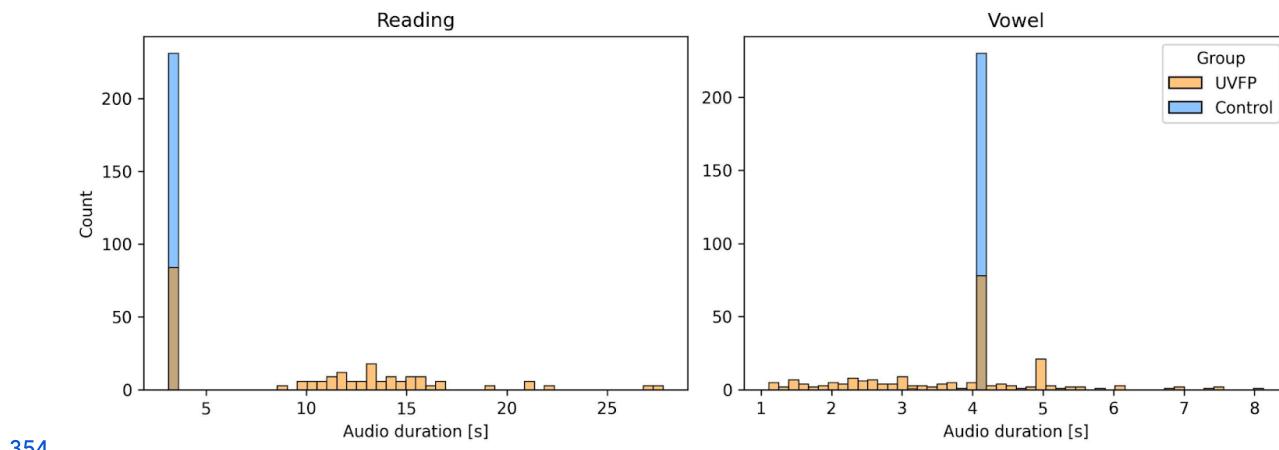
339 Studies tend to report and describe the top N features out of M features, but it is not

340 clear what performance the model would obtain when using only those top N features;

341 perhaps it would perform substantially worse than the full model. We will report
342 performance using only top 5 features as well as performance without top 5 features to
343 provide a more realistic evaluation of their importance.

344 Performance using audio duration

345 Figure 2 indicates clear differences in the distributions of audio recording duration between
346 UVFP patients and controls. This is due to how recordings were processed and saved and not
347 necessarily due to an intrinsic property of UVFP (e.g., slower speech), which reveals a bias that
348 models can leverage but is not expected to generalize well under different audio processing
349 procedures. Therefore, we examine whether audio duration alone could perform well in
350 classification of UVFP. The mean (and standard deviation) for the audio duration for reading
351 task is 3.5 s (0.00 s) for the controls and 10.25 s (6.17 s) for the UVFP patients and the audio
352 duration for sustained vowel task is 4.11 s (0.07 s) for the controls and 3.74 s (1.3 s) for the
353 UVFP patients.



355 **Figure 2. Distribution of audio duration for reading and vowel tasks split by group reveals**

356 a dataset bias. The mode of the audio durations for the controls is 3.5 s for reading samples
357 and 4.11 s for vowel samples.

358

359

360

361 Performance using cepstral peak prominence

362 To evaluate whether results are sensitive to choice of features, we use a different set of
363 features derived from cepstral peak prominence (CPP) given it has been shown to be a
364 good measure of breathiness and dysphonia (44,45). We match the summary statistics
365 across the audio recording that eGeMAPS uses: CPP mean, CPP coefficient of
366 variation (standard deviation normalized by the mean), CPP 20th percentile and CPP
367 80th percentile. We use our custom Python implementation which matches MatLab's
368 COVAREP output (46).

369 Clinician ratings

370 In order to corroborate whether there are unintended recording differences between
371 UVFP patients and controls that may lead to bias, one otorhinolaryngologist and two
372 speech-language pathologists rated each audio recording of the reading task (one per
373 participant, not split in three) for the following variables (and possible responses), in
374 order: background noise (None, Some, High); UVFP (yes, no), CAPE-V severity (0 to

375 100), CAPE-V roughness (0 to 100), CAPE-V breathiness (0 to 100), CAPE-V strain (0
376 to 100), CAPE-V pitch (0 to 100), CAPE-V loudness (0 to 100; estimated loudness as if
377 the rater were in the recording room), recording loudness (low, medium, high; loudness
378 of the recording). Inter-rater agreement was assessed using intra-class correlation for
379 all numerical variables and Light's k for the binary presence of UVFP (47) using the R
380 package *irr* (v0.84.1) (48). The entire reading task was provided instead of the task split
381 in three to make assignment easier for clinicians. The reading task was chosen over the
382 sustained vowel because we expected it to be easier for clinicians to detect UVFP.

383 **RESULTS**

384 **Performance of models using acoustic features**

385 In Table 2, we report performance for models using all features, models after removing
386 redundant features, models using only top 5 features (to understand their unique role in
387 performance), models using all 88 features without 5 features (to understand whether
388 the top 5 features are necessary for high performance), models using audio duration
389 length, and models using a different feature set based on CPP. Performance was found
390 to be high across most models except CPP-based models. Some of the models just
391 using audio duration length were able to achieve close to the highest performance,
392 which reflects the expected effect of the difference in the dataset. Given dependent
393 features provide similar information (see Supplementary Figures S1, S2, S3, S4, S5,
394 S6, S7, S8, and S9) and distort feature importance analyses, we then tested

395 performance after removing redundant features using the Independence Factor method
396 previously described. Supplementary Figure S12 shows performance for different
397 feature set sizes with increasing amounts of redundant features. From this analysis, we
398 selected the feature-set size that resulted in best performance using the least amount of
399 features for subsequent analyses: 39 features (reading), 13 (vowel), 19
400 (reading+vowel). After removing related features (i.e., reducing collinearity) from the
401 original 88 features, similar performance was obtained (median ROC AUC = 0.84–0.87)
402 using fewer features. Supplementary Materials "Feature selection" section describes an
403 analysis of how this method compares to removing features across each train set (see
404 Sup. Mat. Table S1).

405

406
407
408

409 **Table 2. Model performance**

	Features	LogisticRegression	MLP	RandomForest	SGDClassifier
Reading	88	.87 (.78–.93; .50)	.87 (.80–.93; .50)	.87 (.76–.91; .49)	.83 (.76–.89; .50)
Vowel	88	.84 (.77–.89; .50)	.86 (.79–.91; .50)	.86 (.79–.91; .51)	.80 (.72–.87; .50)
Reading+Vowel	88	.84 (.76–.91; .50)	.86 (.74–.92; .48)	.85 (.77–.92; .49)	.79 (.72–.86; .51)
Reading	39	.84 (.76–.92; .50)	.83 (.76–.91; .50)	.87 (.77–.91; .51)	.78 (.71–.86; .51)
Vowel	13	.80 (.70–.90; .50)	.81 (.74–.91; .50)	.84 (.75–.90; .52)	.74 (.58–.87; .51)
Reading+Vowel	19	.79 (.70–.84; .50)	.82 (.75–.88; .51)	.84 (.77–.91; .51)	.70 (.61–.77; .52)
Reading	Top 5	.81 (.73–.89; .50)	.86 (.78–.92; .47)	.85 (.77–.90; .50)	.75 (.56–.87; .57)
Vowel	Top 5	.78 (.67–.87; .50)	.82 (.74–.92; .53)	.81 (.72–.87; .50)	.72 (.57–.82; .49)
Reading+Vowel	Top 5	.80 (.70–.86; .50)	.82 (.74–.88; .50)	.81 (.74–.89; .53)	.72 (.55–.83; .52)
Reading	88 - Top 5	.85 (.76–.92; .50)	.87 (.77–.92; .49)	.85 (.77–.90; .52)	.82 (.71–.89; .51)
Vowel	88 - Top 5	.84 (.75–.93; .50)	.86 (.72–.93; .51)	.84 (.74–.94; .52)	.80 (.70–.90; .48)
Reading+Vowel	88 - Top 5	.84 (.74–.89; .50)	.85 (.76–.91; .50)	.85 (.76–.91; .50)	.79 (.71–.87; .50)
Reading	Duration 1	.81 (.73–.88; .50)	.81 (.73–.88; .50)	.85 (.77–.93; .50)	.76 (.50–.88; .50)
Vowel	Duration 1	.70 (.61–.77; .50)	.80 (.70–.91; .51)	.86 (.76–.94; .52)	.50 (.31–.68; .51)
Reading+Vowel	Duration 1	.70 (.64–.76; .50)	.76 (.67–.84; .50)	.86 (.73–.92; .50)	.64 (.45–.70; .50)
Reading	CPP 4	.76 (.64–.84; .50)	.76 (.64–.84; .46)	.71 (.64–.78; .55)	.74 (.60–.84; .50)
Vowel	CPP 4	.82 (.73–.90; .50)	.82 (.71–.90; .53)	.77 (.65–.85; .50)	.77 (.40–.86; .49)
Reading+Vowel	CPP 4	.72 (.65–.80; .50)	.74 (.68–.84; .53)	.72 (.65–.78; .50)	.68 (.44–.78; .49)

410 Performance of models using either all 88 features, non-redundant features (39, 13, 19), top five most
 411 important features, all 88 features minus top 5 most important features using eGeMAPS features. We
 412 then compared this to using just audio duration as well as a different feature set based on CPP. Median
 413 ROC AUC score from 50 bootstrapping splits (90% confidence interval; median score of null model
 414 trained on permuted labels which should be at .50 if at chance). For full distributions of scores see Figure

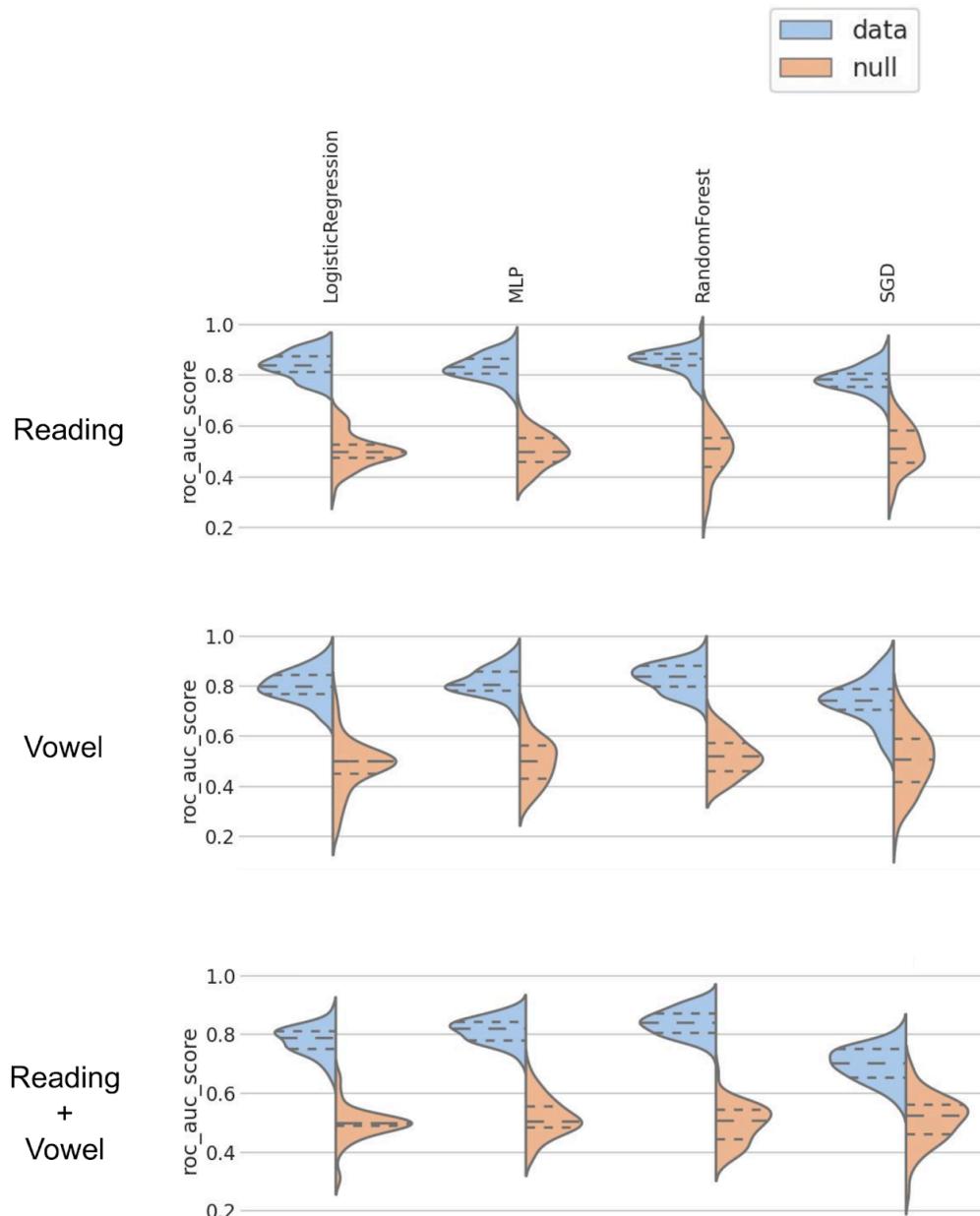
415 S11 in Supplementary Materials. Removing features is a post-hoc analysis because features were
416 selected based on observing performance on the test sets, and therefore performance might be slightly
417 overly optimistic and would need to be tested on an independent test set for further validation. MLP:
418 Multi-Layer Perceptron; SGD: Stochastic Gradient Descent Classifier; CPP: Cepstral Peak Prominence.

419 The bootstrapped ROC AUC distributions and permutation tests for the reduced
420 (parsimonious) models using the non-redundant feature set are shown in Figure 3.

421 Models distribution were all significantly different than their null distribution after
422 correcting for multiple comparisons using a Benjamini-Hochberg procedure.

423

424



425

426 **Figure 3. Model performance comparison using a permutation test using non-redundant features.**

427 Scores from models trained on true labels (blue) and trained on permuted labels (orange) over

428 bootstrapping splits.

429

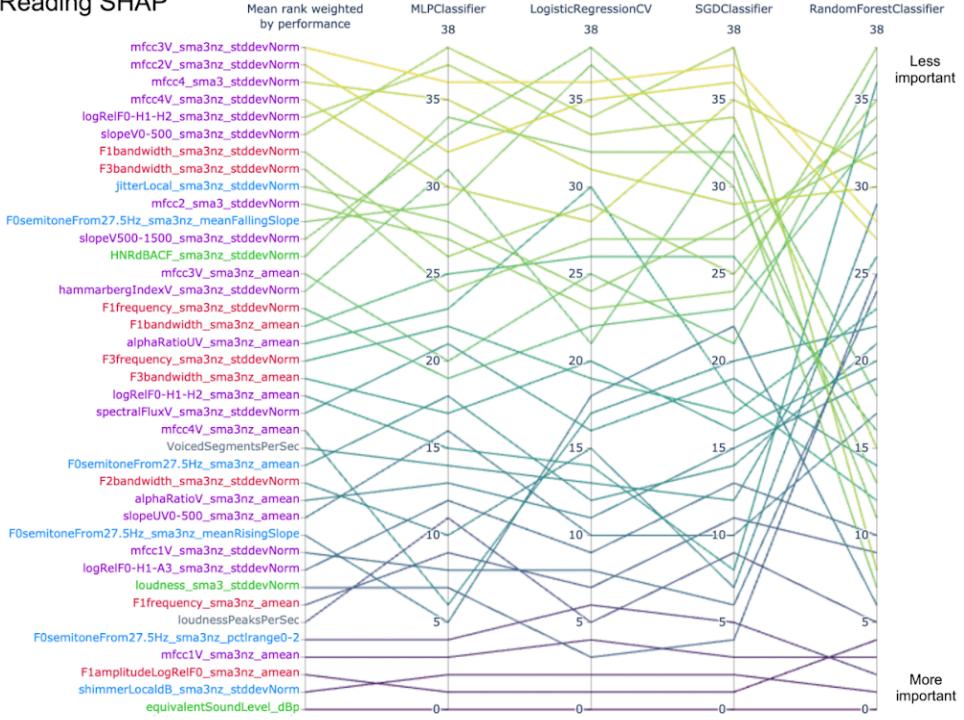
430 Given 24 UVFP patients were recorded with a different device, an iPad, we trained
431 models without their samples to make sure these differences in recordings were not
432 driving performance. There was a small drop in performance, which could be due to a
433 bias (the full, original model using information of the recording device), but could also be
434 due to removing training samples. The drop in performance is not large enough to
435 suspect that differences in recording are driving the full original model's performance
436 (see Sup. Mat. Table S2, Table S3, and analysis in Supplementary section
437 "Performance removing participants that used other recording system").

438 **Assessing feature importance**

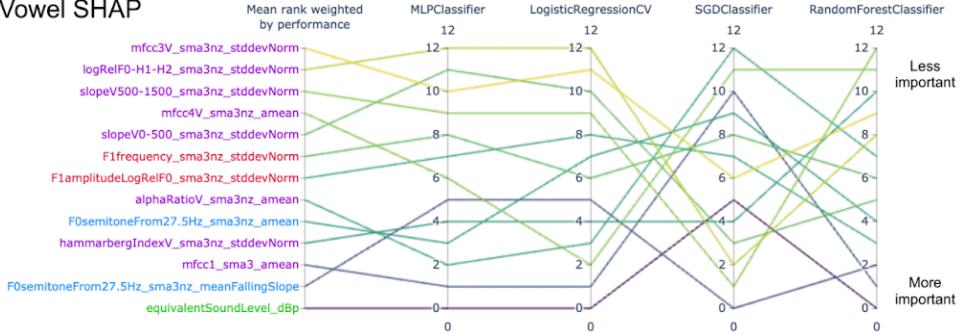
439 Figure 4 reports feature importance using SHAP for all models. For the reading-based
440 models, all models tend to use the same top 5 features except SGD, which also has the
441 lowest performance. For further description of features and the chosen classification of
442 features, see Eyben et al. (2015) (34) and Low et al. (2020) (2). When reviewing
443 important features, it is key to note that any of the features with which it is codependent
444 or associated could be a reasonable important feature (see clusters of redundant
445 features in Supplementary Figures S3-S11). The variance on feature importance rank is
446 evidence that models can use different feature information and still obtain similar high
447 –although not perfect– performance. We further display the distribution of each top
448 feature and its individual performance in Figure 5, which shows that no single feature is
449 enough to dissociate groups with high performance. This figure also revealed the bias:
450 the intensity-related feature equivalent sound level was counterintuitively higher for

451 UVFP patients than controls. Figure 6 reports similarity between top 5 features and all
452 original 88 eGeMAPS features. Features that have a high dcor or distance correlation
453 (i.e., cluster) with top 5 features were not used in models to avoid redundancy, but still
454 share similar information and can therefore be considered important features as well.
455 Hierarchically-clustered heatmaps for other data types (vowel, reading, both) and
456 groups (UVFP patients, controls, both) are displayed in Supplementary Figures S1, S2,
457 S3, S4, S5, S6, S7, S8, and S9. Clustering tends to reflect pre-defined features types
458 such as those reflecting patterns from vocal folds, intensity, vocal tract, spectral
459 analyses, and prosody.

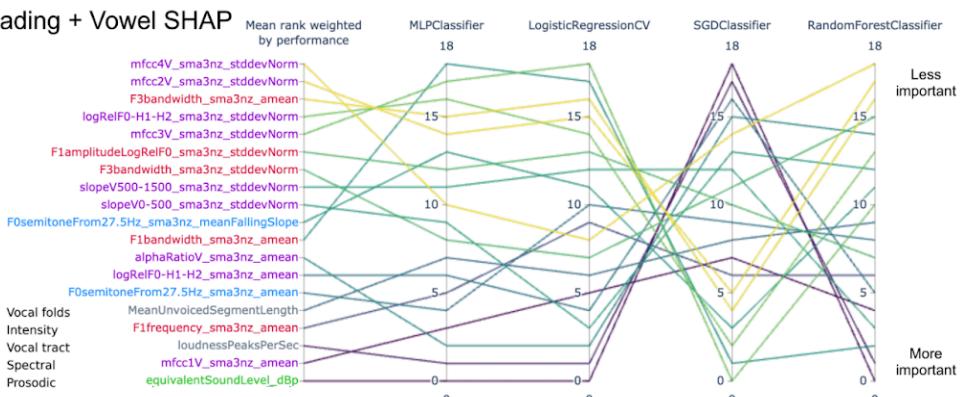
Reading SHAP



Vowel SHAP

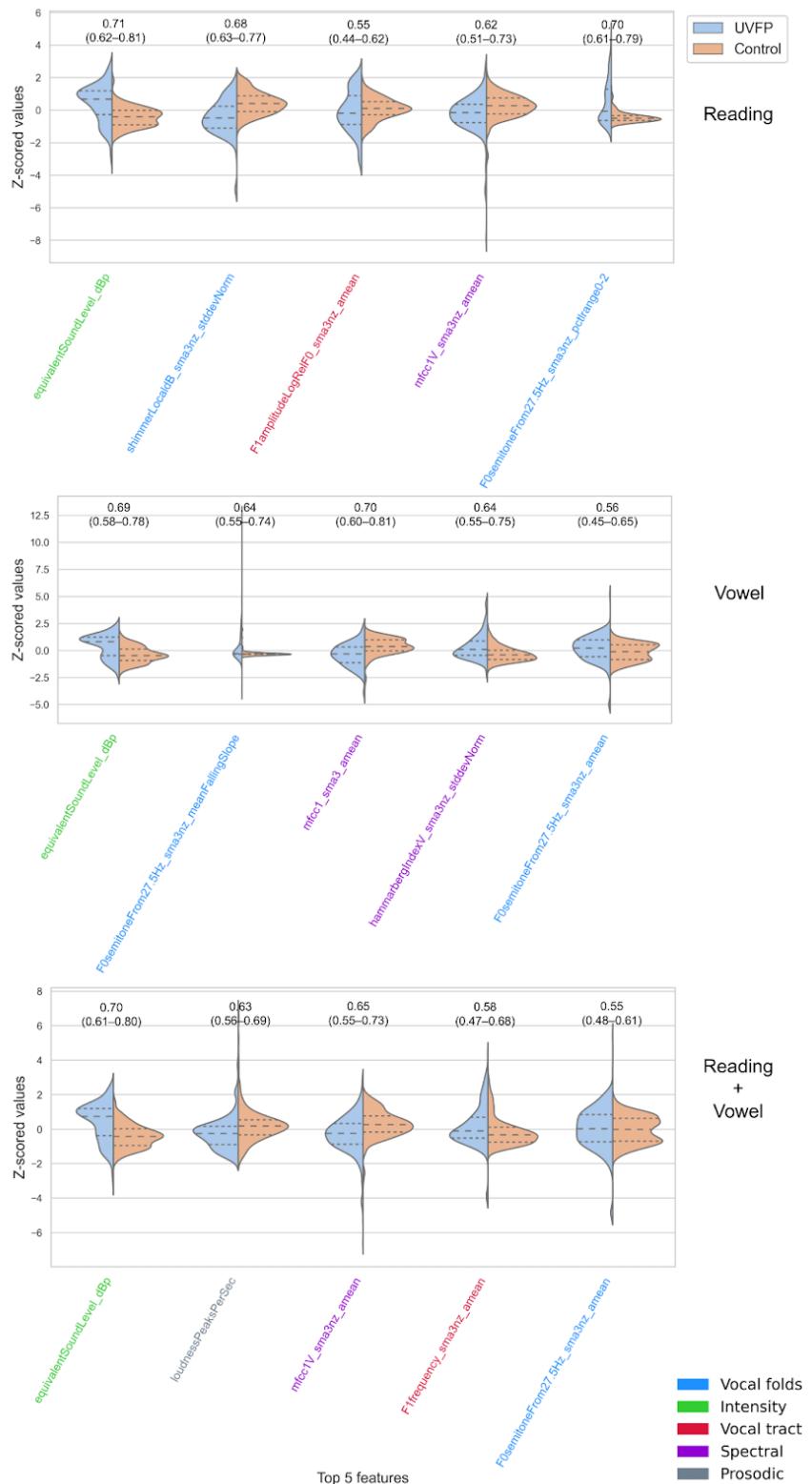


Reading + Vowel SHAP



460

461 **Figure 4. Feature importance parallel coordinate plot.** Rank reads from bottom (most important) to top
 462 (least important). Mean rank is weighted by performance of each model to avoid a lower performing model
 463 biasing the mean rank.

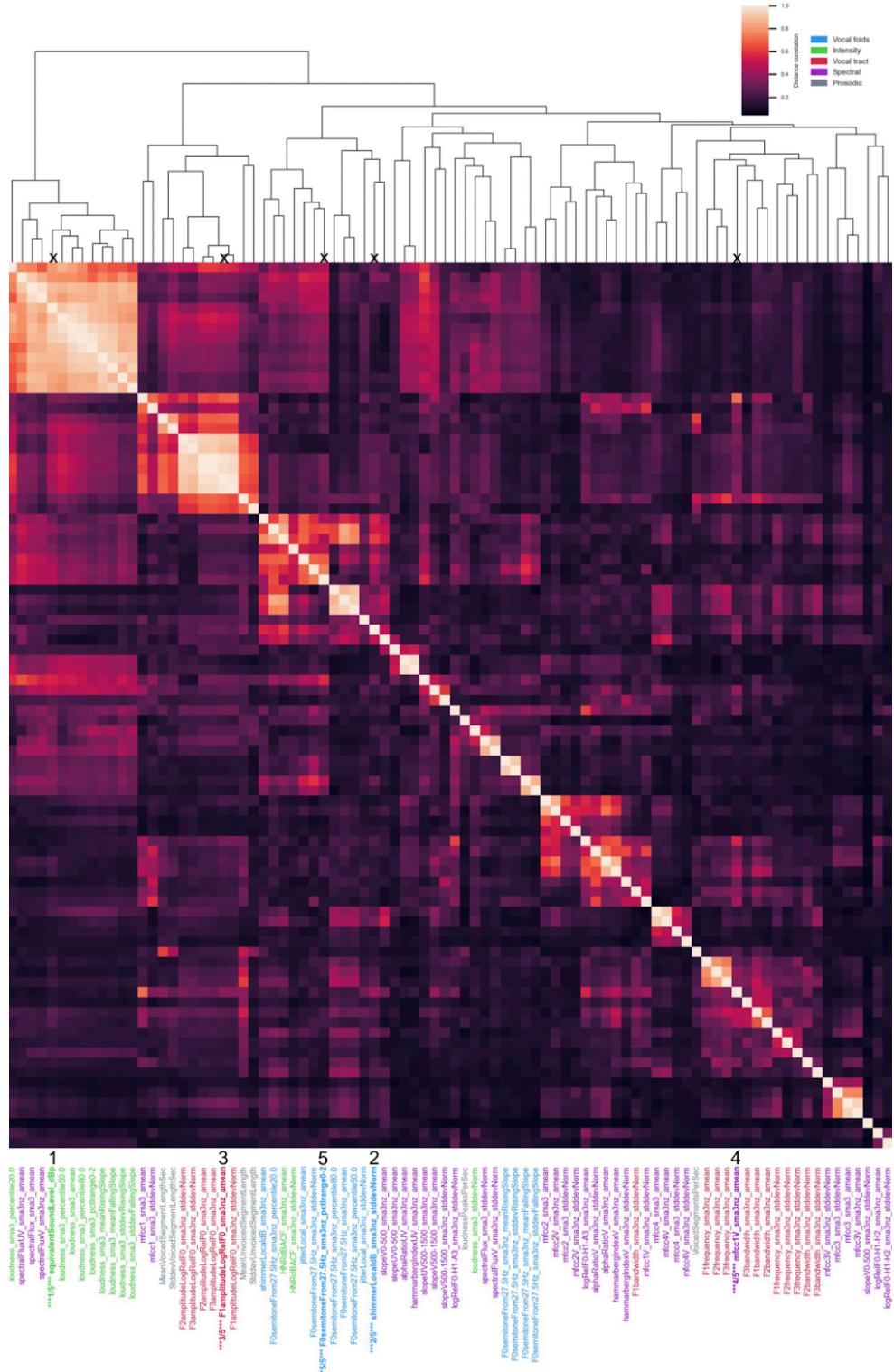


464

465 **Figure 5. Distributions for top 5 features and corresponding performance for single features.** Logistic

466 Regression with L1 penalty was used. No single feature is enough to dissociate groups with high

467 performance. Null models' median performance was 0.5.



468

469 **Figure 6. Feature redundancy with top 5 features highlighted.** Top 5 features are highlighted in bold and
470 their rank is displayed. Squares are clusters of redundant features. Computed with all participants on the
471 reading task.

472 Clinician ratings

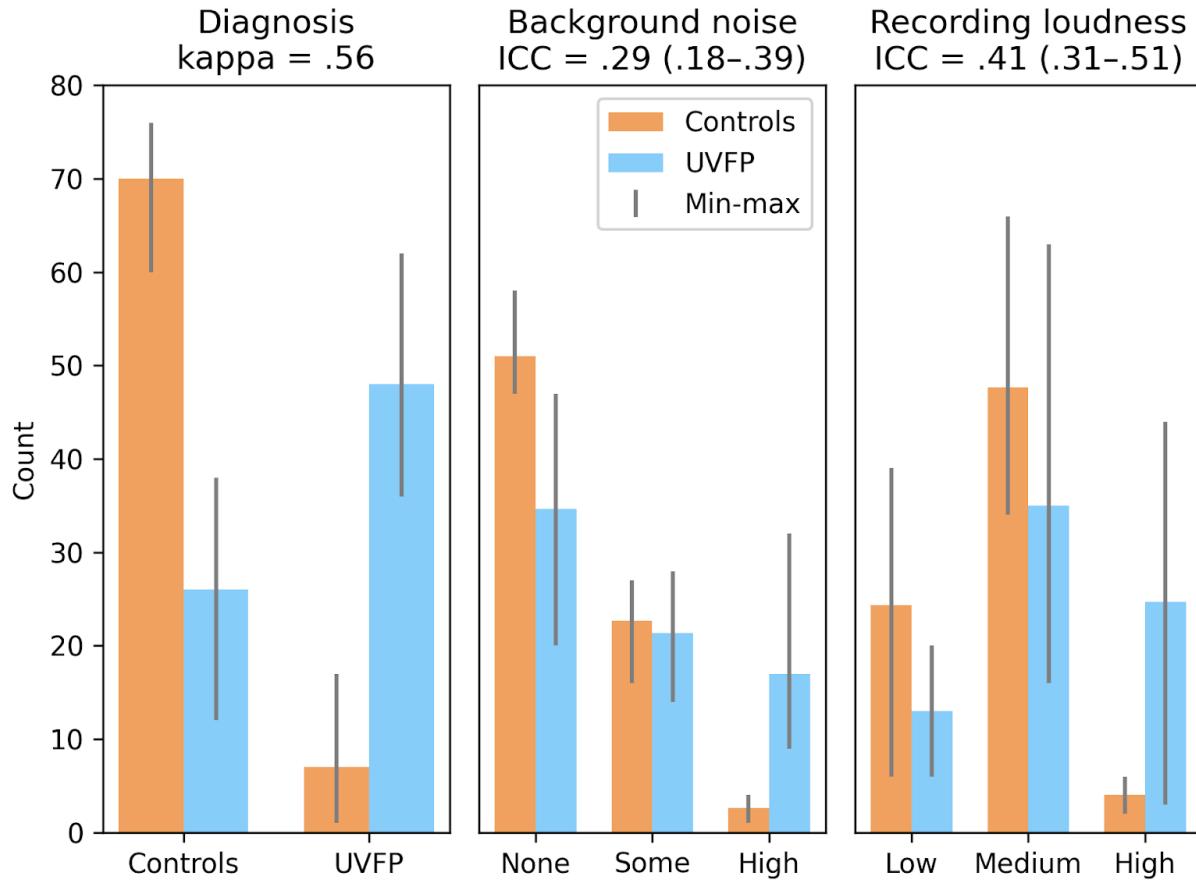
473 The median ROC AUC for humans was 0.78 (min. = 0.74 to max. = 0.81) meaning the
474 machine learning models performed better than the highest performing clinician on the
475 limited available data, that is, the audio samples of the reading task. Interestingly, using
476 the average clinician's CAPE-V ratings within machine learning models was able to obtain
477 a maximum median ROC AUC of 0.84 (0.72–0.92) with the Random Forest model (Table
478 3). Using clinicians' perceptual ratings of background noise and recording loudness
479 achieved a maximum median ROC AUC of 0.77 (.63–.87).

480 **Table 3. Performance using clinician ratings as variables for machine learning models**

	Features	LogisticRegression	MLP	RandomForest	SGD
CAPE-V	6	.80 (.69–.88; .50)	.81 (.71–.90; .50)	.84 (.72–.92; .49)	.77 (.45–.92; .51)
Noise+ loudness	2	.76 (.59–.86; .50)	.77 (.63–.87; .50)	.73 (.62–.83; .52)	.64 (.45–.78; .50)

481 Median ROC AUC score from 50 bootstrapping splits (90% confidence interval; median score of null model
482 trained on permuted labels which should be at .50 if at chance).

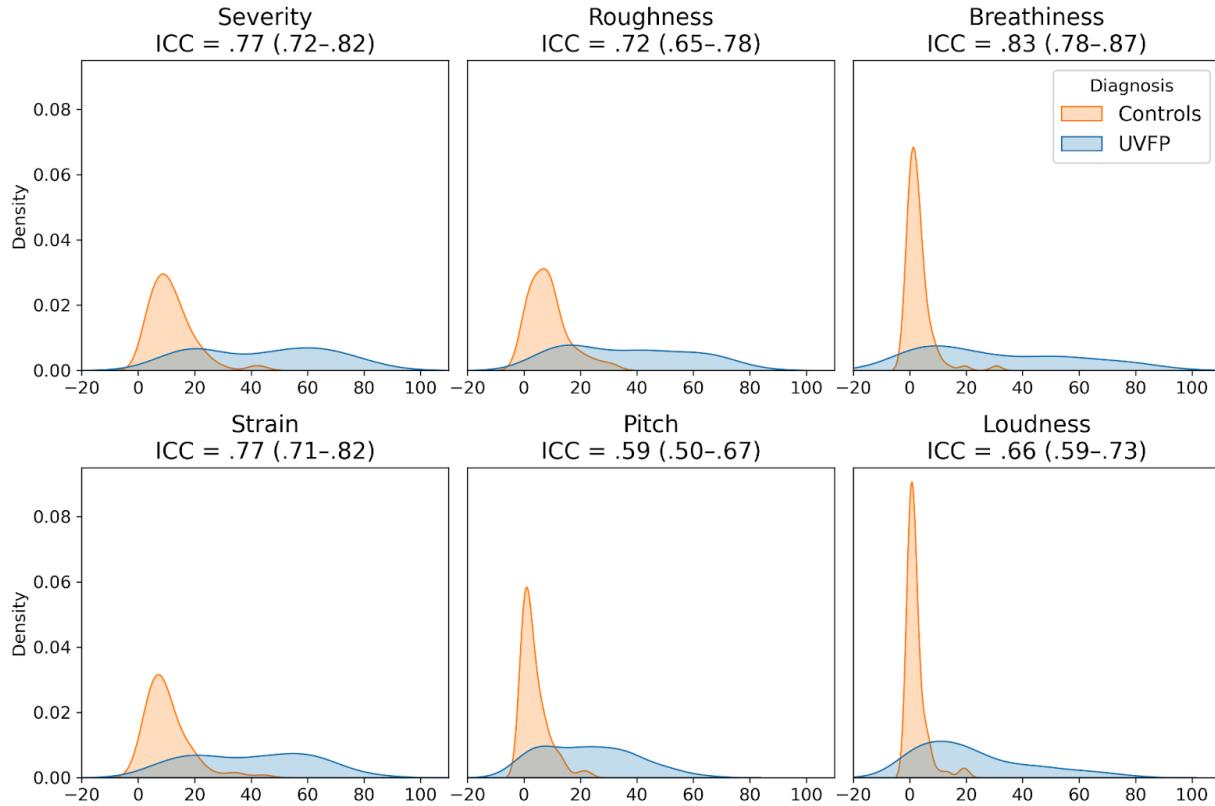
483 In Figures 6 and 7 we report the inter-rater reliability (Fleiss's kappa and ICC) along with
484 the distribution of the ratings. Common cutoffs for inter-rater agreement are poor for values
485 less than .40, fair for values between .40 and .59, good for values between .60 and .74,
486 and excellent for values between .75 and 1.0 (49). Background noise had poor reliability
487 across rater, UVFP and recording loudness had fair reliability (see Figure 7) and
488 CAPE-V-inspired ratings scored good to excellent except for pitch which was fair (see
489 Figure 8).



490

491 **Figure 7. Descriptive statistics and inter-rater reliability of clinician ratings for unilateral vocal fold**
492 **paralysis (UVFP), background noise, and recording loudness indicating likely bias.** Controls and UVFP
493 are ground truth diagnosis from the full clinical interview. Ratings are on brief reading samples. Bars indicate
494 maximum and minimum count across the three raters. The disproportionate amount of UVFP samples rated
495 as having high background noise and high loudness indicates likely bias, where the gain might have been
496 raised for some UVFP patients and they may have phonated more intensely. kappa: Light's kappa; ICC:
497 intra-class correlation coefficient.

498



500 **Figure 8. How clinicians rate the audio recordings of read speech: descriptive statistics and**
501 **inter-rater reliability of average clinician ratings.** The average across raters was taken for each recording.
502 ICC: intra-class correlation coefficient.

503

504 Bias mitigation: matching audio duration and removing features associated to
505 intensity

506 We trimmed the longer UVFP samples so they were matched to control samples (all samples were
507 the same duration), removing the audio duration difference. Vowel samples could not be matched
508 by trimming as some UVFP samples were shorter and some were longer than control samples;
509 therefore we demonstrate an attempt at bias mitigation only with reading samples. In Table 4, we

510 show results on these samples after additionally removing all intensity features as well as variables
511 that have a distance correlation (dcor) with any of them ≥ 0.3 and 0.4 based on the reading
512 samples. Models have comparable performance to models with the original duration and
513 intensity-related biases. See section "Biased features" and Table S4 in Sup. Mat. for a list of the 44
514 features associated with audio duration and the 14 intensity related features. For distance
515 correlations between audio duration and features, see Sup. Mat. Table S6.

516

517 **Table 4. Performance keeping features least associated with intensity features on samples**
518 **of equal audio length after trimming.**

	Features	LogisticRegression	MLP	RandomForest	SGD
dcor<0.4	44	.88 (.80–.92; .50)	.87 (.81–.92; .47)	.87 (.78–.93; .45)	.83 (.76–.90; .48)
dcor<0.3	20	.84 (.78–.89; .50)	.83 (.76–.9; .49)	.85 (.78–.91; .53)	.79 (.66–.87; .51)

519 Median ROC AUC score from 50 bootstrapping splits (90% confidence interval; median score of null model
520 trained on permuted labels which should be at .50 if at chance).

521

522 Discussion

523 This study achieves high performance in detecting UVFP from healthy voices using a few
524 seconds of audio recordings and surpassing clinician evaluations even after mitigating the
525 biases we found in the dataset. As a result of performing the explainability analysis, we
526 discovered a likely bias: intensity features were higher for UVFP patients than controls on
527 average (Figure 5) when UVFP patients should have weaker voices. There are two likely
528 causes. A first cause is that the software that had been used prompted users to speak

529 louder if they had a weak voice in order to achieve an audible recording. A second cause

530 was supported by clinicians' ratings: clinicians rated UVFP patients as having louder

531 recordings and more background noise than controls on average –when they should have

532 similar levels–, which are proxies for microphone gain having been increased. This would

533 have helped models improve performance using characteristics stemming from the

534 recording idiosyncrasies instead of from pathophysiology. However, we removed features

535 correlating with the clearly biased features and still achieved high performance.

536 Our study expands on prior studies which have used pre-existing commercial databases,

537 smaller sample sizes, fewer features, and/or methods for model evaluation that can be

538 biased in small datasets given the test sets may not be representative (for a discussion on

539 bootstrapping for clinical datasets, see Figure 6⁽²⁾). Critically, we provide a roadmap for

540 evaluating models more thoroughly including quantitatively explaining models and

541 checking the robustness of the models to different choices of speech-eliciting tasks,

542 algorithms, and feature sets. All of this should increase trust when no bias is found and

543 when explanations are robust across models and make sense to experts. Such a model

544 could fulfill several clinical needs: (1) postoperative screening for thyroid surgery-related

545 UVFP since after thyroid surgery, UVFP is common, occurring in up to 5 to 10% of cases²⁷.

546 Furthermore, laryngoscopy is not readily available to all postoperative populations and

547 symptomatic changes are notoriously variable. An ML-based screening could help identify

548 patients needing further workup and treatment, and earlier diagnosis is essential to

549 optimize long-term outcomes^{28,29}. (2) Monitoring voice during speech therapy and after

550 surgical treatment for confirmed UVFP to measure when and if the patient's voice is

551 approximating a healthy voice. (3) Preoperative screening prior to surgeries that are at
552 high risk for developing UVFP such as thyroid, head and neck, cardiac, thoracic,
553 esophageal, and cervical spine operations.

554 In Table 5 we summarize several key recommendations to avoid bias when building and
555 explaining machine learning tools for laryngology, although more could be added, and we
556 expand upon how we dealt with some of these steps in the following sections.

557

558 **Table 5. Recommendations to avoid bias for explainable machine learning models that use**
559 **audio recordings for screening in laryngology**

Recommendations	Description
Before data collection	<ul style="list-style-type: none">- Pre-register hypotheses as to which variables should be important for predicting the target group to question effects that are not anticipated by theory (50)
During recording	<ul style="list-style-type: none">- In a <i>controlled</i> recording setting: models could use any unintended differences between groups to improve classification (demonstrated in our study); therefore, it is important to make sure microphone gain, background noise, instructions are consistent across participants and reflect how recordings will be done once deployed.- In a <i>remote</i> setting: it is desirable that models work on people's mobile devices outside the clinic. Since we cannot fully control the recording procedure, we should make sure there are no biases affecting one group more than another, test pilot instructions, and collect much more data to weaken the effect of individual recording idiosyncrasies.- Perform pilot studies to do an initial quality control- Collect representative samples so models generalize to different protected groups (e.g., ages, genders, races) or provide appropriate warnings (51).- Providing instructions so participants do not overproject their voice and control recording procedure so a minimum loudness threshold is not needed (as demonstrated in our study)
Preprocessing and exploratory data analysis	<ul style="list-style-type: none">- Quality control: remove non-natural outliers due to measurement errors, wrong data collection, or wrong data entry (e.g., fixing mislabeled files, unexpected silent recordings, recordings with extreme much background noise)(52)- Avoid or be cautious with preprocessing steps that might reduce the properties associated with the disorder (e.g., denoising may remove breathiness information which may be useful for prediction).- Observe distribution of variables between groups (e.g., audio duration) to make sure there are no differences that are not intrinsic to the disorder. Extra inspection of the data should be taken with retrospective studies where recording protocols were not controlled as in our study.
During training and evaluation	<ul style="list-style-type: none">- Train multiple machine learning models of different complexity: two models may perform similarly but use input variables in different ways. If after training a model we only explain one of them, we might have biased conclusions of what variables characterize the disorder as we demonstrate.- Avoid overfitting (i.e., finding patterns that do not generalize to new samples). Simple held-out test sets (e.g., of 20%) may not be representative of the population or the dataset, and therefore resampling methods (k-fold cross-validation, bootstrapping) are better. If performing hyperparameter tuning, nested resampling is needed to avoid overfitting (2). Avoid feature selection and dimensionality reduction using information from the test set/s. (38,53)- Report performance on most and remaining important features as done in our study
During explainability analyses	<ul style="list-style-type: none">- Choosing one of the variables that are highly dependent due to collinearity (e.g., that correlate above 0.8 Spearman rho or dcor above a threshold that does not reduce performance as we did in this study) or due to multicollinearity (remove variables if variance inflation factor > 5 or 10) (54); grouping correlated variables using leave-one-feature-out (LOFO); obtaining one variable from the correlated variables through dimensionality reduction (without using the test set which could lead to overfitting).

	<ul style="list-style-type: none">- Make conclusions from the features that are robustly important <i>across</i> models; here we take the average importance rank weighted by model performance.- Evaluate potential bias: do important features match hypotheses? Do they dissociate groups in the expected direction? Do certain recording conditions perform better than others and were these done for only one group? Does the model work worse for certain races or age groups? Several metrics can evaluate this (e.g., see packages AIF360, fairlearn, and EqualityML).- Use expert ratings to evaluate any potential sources of bias as done in our study.- Understandability: are the explanations understandable for the engineer, the clinician, and/or the patient? (55)
If bias is detected	<ul style="list-style-type: none">- Use bias mitigation strategies either during pre-processing (removing variables generating the bias along with variables correlated with these ones), training (adversarial debiasing, prejudice remover), or evaluation (equalized odds, reject option classification) (56). See packages AIF360, fairlearn, and EqualityML.
After deployment	<ul style="list-style-type: none">- Continuous assessment: we need to review predictions and re-assess accuracy once deployed as new environments and populations could change performance (i.e., dataset shift (57)).

560

561 **Explaining acoustic features relevant to detecting vocal fold paralysis**

562 Objective acoustic measurement changes associated with vocal fold paralysis have been
563 described and these changes include reduced loudness and maximum phonation time,
564 higher perturbation measurements such as jitter and shimmer, and increased signal to
565 noise ratio (19,58,59); however these were univariate models, and we have demonstrated
566 that using single variables does not seem to provide high predictive performance. While
567 other multivariate machine learning models have been used, these used few features and
568 small or undefined samples and only report feature importance results for one model;
569 therefore it is not clear whether the important features reported would hold using larger
570 feature sets or how other models would perform. Using a much larger initial set of acoustic
571 features for analysis, we demonstrate that several machine learning algorithms of
572 increasing complexity (using more parameters) identify vocal fold paralysis from healthy

573 voices. We also report that these models can use different features to achieve similar
574 performance. Different models emphasize different features not simply because of its
575 relevance to a disorder, but because of the mathematics associated with the model (e.g.,
576 containing different degrees of interaction effects, regularization, or propensity to
577 underfitting or overfitting) (60). The variability of the ranking of features used by our
578 individual models also illustrates the potential danger of using the single highest
579 performing model, which is commonly seen in published literature.

580 Instead of simply reporting the important features from the highest performing model, we
581 analyzed the models to find common features. The most important features across models
582 were somewhat associated with intensity features (Sup. Mat. Table S5); therefore, even if
583 not strongly associated with intensity features, they could be important due to a
584 combination of intrinsic differences between UVFP and controls for which we provide
585 hypotheses or because of how intensity influences them; a new unbiased dataset would be
586 needed to confirm this. These top features were: intensity, especially equivalent sound
587 pressure level which was redundant with multiple loudness features and seems to be due
588 to some patients trying to use more breath for projection or being recorded with a higher
589 microphone gain; Mel Frequency Cepstral Coefficients (especially the first coefficient,
590 which captures spectral envelope or slope, which has been shown to be important for
591 predicting UVFP ((29)); mean F0 semitones given F0 originates from vocal-fold oscillation,
592 a vocal-fold paralysis is expected to alter F0, and has been shown to help predict
593 pathological speech including UVFP (28);, mean F1 amplitude and frequency, influenced
594 by how the vocal tract filters F0 and the shape of the glottal pulse which would be affected

595 by UVFP voiced and unvoiced segments (prosodic and speech articulation features which
596 may be altered due to changes in the periodicity of F0), and CPP features (which indicate
597 voice quality degradations that could include more breathiness, a typical feature of UVFP
598 (61)). Shimmer variability was important just for reading, and it captures variability in glottal
599 pulses and pressure patterns which ultimately affect F0 and has been found to be
600 significantly different between UVFP and a control group (62). When we removed the top 5
601 features from the full feature set, performance is practically equivalent to using 88 features,
602 as expected, since there are features that are redundant with the top 5 features. Therefore,
603 it is not that only these 5 specific features drive performance, but rather the information
604 they contain, which in this dataset is also captured by other features as shown in Figure 6.

605 These acoustic features would corroborate our clinical understanding of glottal
606 incompetence from UVFP and with common patient complaints of reduced loudness, vocal
607 instability, hoarseness, and rough voice; however, they could also be important due to their
608 associations with intensity features. Uncovering and understanding the basic mechanisms
609 and features that models use to generate predictions and outcomes are important as these
610 tools become part of the clinical decision making process.

611 **Identifying and addressing bias**

612 Equivalent Sound Level was higher in UVFP patients than controls. This is counter-intuitive
613 because UVFP patients are known to have softer voices as already described; however,
614 clinicians rated most UVFP samples as being louder than controls. The bias discovered
615 was likely due to increasing the gain on the microphone for some UVFP patients, which

616 would explain the increased background noise in UVFP patients' recordings. A second
617 source of bias may have occurred from requesting UVFP patients to speak louder in order
618 to meet the minimum intensity threshold on the recording softwares Computerized Speech
619 Lab™ and OperaVOX, or patients could have tried this on their own knowing they were
620 being recorded. This behavioral compensation is likely to occur in biomarker research
621 when the participant has a soft voice, especially in retrospective studies like ours where
622 the study goal is not known at the time of recording or when certain software properties
623 lead individuals with weak voices to speak louder. Even though the current models perform
624 better than the clinicians, a systematic comparison would require more clinician and model
625 assessments across datasets. It is likely a model trained on a single dataset might learn
626 intrinsic characteristics of that dataset that do not generalize as well as clinical expertise
627 might.

628 Having said this, this line of research would help us understand the extent to which UVFP
629 detection is generalizable from acoustic data alone. Finding an objective measure of
630 hoarseness is important given a "normal voice" is a fundamentally subjective classification
631 that is not well defined (63,64) and varies with training (65,66), which may result in low
632 reliability of evaluation of disordered voices among clinical rating scales (67).

633 As a post hoc analysis, we address bias by trying to mitigate its effect: we removed
634 variables associated with intensity variables on samples matched on audio duration. After
635 removing these features, the models were able to obtain similar performance using a very
636 different set of features. It is possible that these remaining features better reflect

637 pathophysiology or that the features extracted are still influenced by intensity, but further

638 studies should address their generalizability or their relation to intensity variation.

639 **Evaluating the sensitivity to tasks, model complexity, and features used**

640 In addition to getting a better understanding of features, we explored performance in the

641 context of different vocal tasks. Participants carried out two different tasks to elicit voice,

642 *reading*, which captures more complex speech dynamics, and *sustaining vowels*, which is

643 a simpler measure of vocalization and the respiratory subsystem. Overall, these dynamics

644 from the speech task may have improved model performance as was observed.

645 Comparing simpler and more complex models is important because simpler models such

646 as Logistic Regression could be preferred because they tend to generalize better given

647 they are less at risk for overfitting the training set and they are more interpretable and thus

648 biases can be assessed more directly (68).

649 By removing redundant features, we can concentrate on finding the most useful features

650 for further analysis. Performance decreased only slightly while we made models more

651 parsimonious and explainable. This approach is key given the curse of dimensionality in

652 machine learning that may make models unnecessarily complex and harder to generalize

653 (20).

654 Often studies will report the top N features but not how predictive they are in isolation. In

655 our study we ran models on the top 5 features together (Table 2). The lower performance

656 of these top 5 features relative to a richer feature set helps demonstrate that model

657 performance is dependent on interactions across multiple additional features (with the
658 exception of samples from the reading task which obtained an AUC of 0.86 using just the 5
659 features). We also ran models without top 5 features to demonstrate that leaving features
660 that are redundant with these top features results in almost equivalent high performance to
661 using all 88 features since the redundant features share information. Furthermore, when
662 training models on the individual features from within these top 5 one at a time, the
663 performance was reduced considerably with scores from 0.55 to 0.71. This indicates the
664 need for these models to combine multiple features to achieve high performance and any
665 model evaluation should not focus on only the common or top features without testing their
666 predictive performance.

667 **Limitations and future directions**

668 We cannot determine how the bias will affect the model's performance on future samples,
669 but it will likely underperform in samples where length was not different between groups,
670 where gain cannot be changed, and where participants are instructed to not overproject
671 their voice; however, it is possible the model could underperform for other reasons
672 including dataset shift (e.g., the distribution of voice characteristics or demographics is
673 different in a new sample).

674 The classification using just duration itself varied across models and clinicians who
675 listened to the reading passage in its entirety did not achieve as good a classification as
676 the best performing models. Duration itself was not included as a feature in the
677 eGeMaps-based models and has a complex effect on both machines and humans. For

678 example, duration could have affected eGeMAPS features (e.g., introduce more variability
679 to the functionals that are computed over sliding time windows) and duration of vowels
680 varied extensively across the UVFP group thus cannot itself be tied to underlying
681 pathophysiology. Therefore, important future work should analyze how duration may affect
682 these features, should address the intrinsic variability in durations of UVFP patients in
683 responding to speech items, and should incorporate models of production that include a
684 consideration of respiratory capabilities, articulation changes, and vocal fold
685 pathophysiology.

686 It is not clear whether these models could detect UVFP from other voice disorders or just
687 healthier voices; however, a model that generalizes well in classifying UVFP from controls
688 could be used to monitor UVFP patients remotely and affordably during treatment or detect
689 risk for UVFP when it is the most likely cause (e.g., dysphonia after thyroid surgery).
690 Larger sample sizes with curated examinations can help increase diverse representation
691 across voice quality and thereby potentially reduce bias in classifier performance. We did
692 not analyze potential racial bias given this data was not extracted from the chart review.
693 Our choice of a standardized feature set worked well in this setting, but may fail to work for
694 differential voice disorder diagnosis or when generalizing to larger datasets, which may
695 bring in additional sources of variance unaccounted for in this dataset. With the availability
696 of more data, additional features could be extracted that better capture changes in
697 coordination (e.g., XCORR (69)).
698 Furthermore, while our feature importance evaluation method, SHAP, shows a certain
699 amount of robustness across models, alternative model-agnostic feature-importance

700 methods (e.g., LOFO, permutation importance) as well as model-specific methods
701 (coefficient values for linear models, mean decrease in impurity for Random Forest) could
702 be compared. Model understandability –how easily are the explanations understood by a
703 speech scientist or a clinician– could be assessed rigorously (55).
704 Finally, debiasing the models by removing features correlated with the biased ones was
705 attempted although it is not clear how exactly intensity may influence certain features; we
706 assume if intensity is influencing a variable, it generally should create some considerable
707 association which we discarded using dcor. Therefore, the effect of the bias can be
708 assessed by testing the model's generalizability to new unbiased datasets. Therefore, we
709 are not promoting our final debiased models as completely unbiased or ready to use, it is
710 possible our debiasing strategies are only partially effective, additional biases remain,
711 and/or additional ways of debiasing have not been considered.
712 We tested how well a model using only the top 5 features performed independently of the
713 model with all features; it is possible to also test how well the incremental set of top
714 features performs (1st, 1st and 2nd, 1st–3rd, etc.), which would be useful in order to
715 compare different models' performance as a function of which features are being used.

716 Conclusion

717 Using one of the largest UVFP datasets to date, our study demonstrates the importance of
718 checking for biases using explainable machine learning and clinician perceptual ratings. In
719 order to first explain models, we tackle collinearity (i.e., redundant or highly correlated
720 independent variables), which biases feature importance, using a custom method called

721 Independence Factor that selects one out of a set of associated features without losing
722 predictive performance. We then compare how results change across different
723 speech-eliciting tasks, training algorithms, features, features set sizes, and highest and
724 lowest performing features to better understand the process that models use to predict
725 vocal changes associated with laryngeal disease, since analyzing a single model will result
726 in a biased view of how predictions are achieved. During this process, we discovered there
727 was a difference in audio duration between groups clearly not related to intrinsic
728 differences in UVFP speech rate, but in cropping all control recordings to a certain length
729 during audio storage. We also discovered that sound equivalent level was
730 counterintuitively higher in UVFP patients, a likely bias resulting from the weak or
731 underprojected voice that characterizes many UVFP patients: patients were prompted by
732 the recording software to speak louder and the microphone gain was likely raised
733 selectively for these patients with weaker voices, possibly generating higher background
734 noise which was detected through clinician's ratings; therefore the models picked up on
735 the acoustic correlates of this increased intensity, which would impede generalization
736 under different recording procedures and natural audio durations. This is more likely to
737 occur in laryngology datasets when patients have a softer voice.

738 Interestingly, we found that matching audio duration between groups and removing all
739 variables that were clearly related to intensity (e.g., bias mitigation) resulted in similar high
740 performance. In this case, the model may be using information more related to
741 pathophysiology, which would need to be further confirmed by future unbiased samples.
742 Machine learning models tended to surpass clinician's evaluation of the same audio

743 recordings. Interestingly, using clinician's voice quality ratings on the recordings in machine
744 learning models performed better than their binary evaluation on whether recordings
745 contained a sample of UVFP voice or not.

746 We hope to promote moving beyond using a single model and only reporting top features
747 to a better explanation of how these models work as well as being able to understand
748 variance across modeling and evaluation choices. We believe these are all aspects of
749 machine learning that clinicians need to understand prior to using such applications.

750 With these considerations along with the recommendations we make, machine learning
751 applications could aid in laryngology screening, allowing for the potential development of
752 in-home screening assessments and continuous pre- and post-treatment monitoring.

753 Acknowledgments

754 We would like to thank Cody Sullivan and Carolyn Hsu for their help in rating the audio
755 samples and thank Daryush Mehta, Robert Hillman, and John Guttag for their feedback on
756 an earlier version of this study. DML was supported by a National Institute on Deafness
757 and Other Communication Disorders T32 training grant [5T32DC000038-28], a RallyPoint
758 Fellowship, and an Amelia Peabody Professional Development Award. The work was
759 supported by a gift to the McGovern Institute for Brain Research at MIT. SSG was partially
760 supported by National Institutes of Health grants for the development of pydra-ml [R01
761 EB020740], for reproducible practices [P41 EB019936], and the Bridge2AI voice data
762 generation project [1OT2OD032720-01]. The authors declare that there is no conflict of

763 interest.

764

765 Data Availability Statement

766 All data and code are available through Github (<https://github.com/danielmlow/vfp>) and

767 Zenodo (<https://doi.org/10.5281/zenodo.5009208>) including a tutorial to test our models on

768 your own data (https://github.com/danielmlow/vfp/blob/main/vfp_detector.ipynb).

769

770 Author Contributions

771 Daniel M. Low: Data curation, Methodology, Formal analysis, Software, Writing - Original

772 Draft; Vishwanatha Rao: Data Curation, Formal analysis, Writing - Original Draft; Gregory

773 Randolph: Writing - Review & Editing; Philip C. Song: Conceptualization, Methodology,

774 Writing - Original Draft, Supervision, Data curation; Satrajit S. Ghosh: Conceptualization,

775 Methodology, Writing - Original Draft, Supervision, Software

776

777 References

778 1. Wroge TJ, Özkanca Y, Demiroglu C, Si D. Parkinson's disease diagnosis using machine learning and
779 voice. 2018 IEEE signal [Internet]. 2018.

780 2. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A
781 systematic review. Laryngoscope Investig Otolaryngol. 2020 Feb;5(1):96–116.

782 3. Quatieri TF. Discrete-Time Speech Signal Processing: Principles and Practice. Pearson Education;
783 2008. 816 p.

784 4. Molnar C. Interpretable Machine Learning. Lulu.com; 2019. 319 p.

785 5. Stachler RJ, Francis DO, Schwartz SR, Damask CC, Digoy GP, Krouse HJ, et al. Clinical practice
786 guideline: Hoarseness (dysphonia) (update). Otolaryngol Head Neck Surg. 2018
787 Mar;158(1_suppl):S1–42.

- 788 6. Brunner E, Friedrich G, Kiesler K, Chibidziura-Priesching J, Gugatschka M. Subjective breathing
789 impairment in unilateral vocal fold paralysis. *Folia Phoniatr Logop.* 2011;63(3):142–6.
- 790 7. Spataro EA, Grindler DJ, Paniello RC. Etiology and Time to Presentation of Unilateral Vocal Fold
791 Paralysis. *Otolaryngol Head Neck Surg.* 2014 Aug;151(2):286–93.
- 792 8. Sritharan N, Chase M, Kamani D. The vagus nerve, recurrent laryngeal nerve, and external branch of
793 the superior laryngeal nerve have unique latencies allowing for intraoperative documentation of The
794 [Internet]. 2015.
- 795 9. Randolph GW, Kamani D. The importance of preoperative laryngoscopy in patients undergoing
796 thyroidectomy: voice, vocal cord function, and the preoperative detection of invasive thyroid malignancy.
797 *Surgery.* 2006 Mar;139(3):357–62.
- 798 10. Colton RH, Paseman A, Kelley RT, Stepp D, Casper JK. Spectral moment analysis of unilateral vocal
799 fold paralysis. *J Voice.* 2011 May;25(3):330–6.
- 800 11. Balasubramanium RK, Bhat JS, Fahim S 3rd, Raju R 3rd. Cepstral analysis of voice in unilateral
801 adductor vocal fold palsy. *J Voice.* 2011 May;25(3):326–9.
- 802 12. Little M, Costello D, Harries M. Objective dysphonia quantification in vocal fold paralysis: comparing
803 nonlinear with classical measures. *Nature Precedings.* 2009 Apr 21;1–1.
- 804 13. Bielamowicz S, Stager SV. Diagnosis of unilateral recurrent laryngeal nerve paralysis: laryngeal
805 electromyography, subjective rating scales, acoustic and aerodynamic measures. *Laryngoscope.* 2006
806 Mar;116(3):359–64.
- 807 14. Hartl DAM, Hans S, Vaissière J, Brasnu DAMF. Objective acoustic and aerodynamic measures of
808 breathiness in paralytic dysphonia. *Eur Arch Otorhinolaryngol.* 2003 Apr;260(4):175–82.
- 809 15. Francis DO, Pearce EC, Ni S, Garrett CG, Penson DF. Epidemiology of vocal fold paralyses after total
810 thyroidectomy for well-differentiated thyroid cancer in a Medicare population. *Otolaryngol Head Neck
811 Surg.* 2014 Apr;150(4):548–57.
- 812 16. Jeannon JP, Orabi AA, Bruch GA, Abdalsalam HA, Simo R. Diagnosis of recurrent laryngeal nerve palsy
813 after thyroidectomy: a systematic review. *Int J Clin Pract.* 2009 Apr;63(4):624–9.
- 814 17. Bhattacharyya N, Kotz T, Shapiro J. Dysphagia and aspiration with unilateral vocal cord immobility:
815 incidence, characterization, and response to surgical treatment. *Ann Otol Rhinol Laryngol.* 2002
816 Aug;111(8):672–9.
- 817 18. Pinho CMR, Jesus LMT, Barney A. Aerodynamic measures of speech in unilateral vocal fold paralysis
818 (UVFP) patients. *Logoped Phoniatr Vocol.* 2013 Apr;38(1):19–34.
- 819 19. Hartl DM, Crevier-Buchman L, Vaissière J, Brasnu DF. Phonetic effects of paralytic dysphonia. *Ann Otol
820 Rhinol Laryngol.* 2005 Oct;114(10):792–8.
- 821 20. Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. Digital medicine
822 and the curse of dimensionality. *NPJ Digital Medicine.* 2021 Dec;4(1):s41746–021.
- 823 21. Rusz J, Švihlík J, Krýže P, Novotný M, Tykalová T. Reproducibility of Voice Analysis with Machine
824 Learning. *Mov Disord.* 2021 May;36(5):1282–3.
- 825 22. Schönweiler R, Hess M, Wübbelt P, Ptok M. Novel approach to acoustical voice analysis using artificial
826 neural networks. *J Assoc Res Otolaryngol.* 2000 Dec;1(4):270–82.

- 827 23. Godino-Llorente JI, Gómez-Vilda P. Automatic detection of voice impairments by means of short-term
828 cepstral parameters and neural network based detectors. *IEEE Trans Biomed Eng.* 2004
829 Feb;51(2):380–4.
- 830 24. Fraile R, Saenz-Lechon N, Godino-Llorente JI, Osma-Ruiz V, Fredouille C. Automatic detection of
831 laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient
832 parameters and differentiation of patients by sex. *Folia Phoniatr Logop.* 2009;61(3):146–52.
- 833 25. Voigt D, Döllinger M, Yang A, Eysholdt U, Lohscheller J. Automatic diagnosis of vocal fold paresis by
834 employing phonovibrogram features and machine learning methods. *Comput Methods Programs
835 Biomed.* 2010 Sep;99(3):275–88.
- 836 26. Lopes LW, Batista Simões L, Delfino da Silva J, da Silva Evangelista D, da Nóbrega E Uguilino AC,
837 Oliveira Costa Silva P, et al. Accuracy of Acoustic Analysis Measurements in the Evaluation of Patients
838 With Different Laryngeal Diagnoses. *J Voice.* 2017 May;31(3):382.e15–382.e26.
- 839 27. Powell ME, Rodriguez Cancio M, Young D, Nock W, Abdelmessih B, Zeller A, et al. Decoding phonation
840 with artificial intelligence (DeP AI): Proof of concept. *Laryngoscope Investig Otolaryngol.* 2019
841 Jun;4(3):328–34.
- 842 28. Dibazar AA, Narayanan S, Berger TW. Feature analysis for automatic detection of pathological speech.
843 In: Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the
844 Biomedical Engineering Society] [Engineering in Medicine and Biology. 2002. p. 182–3 vol.1.
- 845 29. Seedat N, Aharonson V, Hamzany Y. Automated and interpretable m-health discrimination of vocal cord
846 pathology enabled by machine learning. In: 2020 IEEE Asia-Pacific Conference on Computer Science
847 and Data Engineering (CSDE). 2020. p. 1–6.
- 848 30. Mittal V, Sharma RK. Deep Learning Approach for Voice Pathology Detection and Classification. *IJHISI.*
849 2021 Oct 1;16(4):1–30.
- 850 31. Hu HC, Chang SY, Wang CH, Li KJ, Cho HY, Chen YT, et al. Deep Learning Application for Vocal Fold
851 Disease Prediction Through Voice Recognition: Preliminary Development Study. *J Med Internet Res.*
852 2021 Jun 8;23(6):e25247.
- 853 32. Ras G, Xie N, van Gerven M, Doran D. Explainable Deep Learning: A Field Guide for the Uninitiated.
854 *jair.* 2022 Jan 25;73:329–96.
- 855 33. Fairbanks G. *Voice and Articulation Drillbook.* Harper; 1960. 196 p.
- 856 34. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The Geneva Minimalistic
857 Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on
858 Affective Computing.* 2016 Apr;7(2):190–202.
- 859 35. audEERING GmbH. *openSMILE (Version 2.3) [Internet].* 2017. Available from:
860 <https://github.com/naxingyu/opensmile/blob/3a0968e7b36c1b730a4ffd2977031091ee9abf7f/config/gemaps/eGeMAPSv01a.conf>
- 862 36. Satrajit S Ghosh, Daniel M Low, Hoda Rajaei et al. Pydra-ML doi:10.5281/ZENODO.4170850 [Internet].
863 Available from: <https://github.com/nipype/pydra-ml>
- 864 37. Lipton ZC. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is
865 both important and slippery. *Queueing Syst.* 2018 Jun 1;16(3):31–57.
- 866 38. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning [Internet].

- 867 arXiv [cs.LG]. 2018. Available from: <http://arxiv.org/abs/1811.12808>
- 868 39. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. In: 2009 Ninth IEEE
869 International Conference on Data Mining. IEEE; 2009. p. 1833–63.
- 870 40. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. arXiv [cs.AI]. 2017.
871 Available from: <http://arxiv.org/abs/1705.07874>
- 872 41. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification presents
873 challenges for credibility in modern machine learning. *J Mach Learn Res.* 2022 Jan 1;23(1):10237–97.
- 874 42. de Siqueira Santos S, Takahashi DY, Nakata A, Fujita A. A comparative study of statistical methods
875 used to identify dependencies between gene expression signals. *Brief Bioinform.* 2014
876 Nov;15(6):906–18.
- 877 43. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances.
878 2007.
- 879 44. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous
880 speech. *J Speech Hear Res.* 1996 Apr;39(2):311–21.
- 881 45. Murton O, Hillman R, Mehta D. Cepstral Peak Prominence Values for Clinical Voice Evaluation. *Am J
882 Speech Lang Pathol.* 2020 Aug 4;29(3):1596–607.
- 883 46. G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer. COVAREP—A collaborative voice analysis
884 repository for speech technologies. *Proc IEEE Int Conf Acoust Speech Signal Process* [Internet]. 2014
885 [cited 2023 Oct 21].
- 886 47. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor
887 Quant Methods Psychol.* 2012;8(1):23–34.
- 888 48. Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W. Package “irr.” Various coefficients of
889 interrater reliability and agreement. 2012;22:1–32.
- 890 49. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized
891 assessment instruments in psychology. *Psychol Assess.* 1994 Dec;6(4):284–90.
- 892 50. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution [Internet]. Vol. 115,
893 Proceedings of the National Academy of Sciences. 2018. p. 2600–6. Available from:
894 <http://dx.doi.org/10.1073/pnas.1708274114>
- 895 51. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine
896 Learning. *ACM Comput Surv.* 2021 Jul 13;54(6):1–35.
- 897 52. Osborne JW, Overbay A. The power of outliers (and why researchers should ALWAYS check for them).
898 *Practical Assessment, Research, and Evaluation.* 2019;9(1):6.
- 899 53. Kapoor S, Cantrell E, Peng K, Pham TH, Bail CA, Gundersen OE, et al. REFORMS: Reporting
900 Standards for Machine Learning Based Science [Internet]. arXiv [cs.LG]. 2023. Available from:
901 <http://arxiv.org/abs/2308.07832>
- 902 54. Thompson CG, Kim RS, Aloe AM, Becker BJ. Extracting the Variance Inflation Factor and Other
903 Multicollinearity Diagnostics from Typical Regression Results. *Basic Appl Soc Psych.* 2017 Mar
904 4;39(2):81–90.
- 905 55. Zhou Y, Ribeiro MT, Shah J. ExSum: From Local Explanations to Model Understanding [Internet]. arXiv

- 906 [cs.CL]. 2022. Available from: <http://arxiv.org/abs/2205.00130>
- 907 56. Hort M, Chen Z, Zhang JM, Harman M, Sarro F. Bias Mitigation for Machine Learning Classifiers: A
908 Comprehensive Survey. ACM J Responsib Comput [Internet]. 2023 Nov 1; Available from:
909 <https://doi.org/10.1145/3631326>
- 910 57. Dockès J, Varoquaux G, Poline JB. Preventing dataset shift from breaking machine-learning biomarkers.
911 Gigascience [Internet]. 2021 Sep 28;10(9).
- 912 58. Ramig LA, Scherer RC, Titze IR, Ringel SP. Acoustic analysis of voices of patients with neurologic
913 disease: rationale and preliminary data. Ann Otol Rhinol Laryngol. 1988 Mar-Apr;97(2 Pt 1):164–72.
- 914 59. Morsomme D, Jamart J, Wéry C, Giovanni A, Remacle M. Comparison between the GIRBAS Scale and
915 the Acoustic and Aerodynamic Measures Provided by EVA for the Assessment of Dysphonia following
916 Unilateral Vocal Fold Paralysis. Folia Phoniatr Logop. 2001 Nov-Dec;53(6):317–25.
- 917 60. Kriegeskorte N, Douglas PK. Interpreting encoding and decoding models. Curr Opin Neurobiol. 2019
918 Apr;55:167–79.
- 919 61. Hartl DM, Hans S, Vaissière J, Riquet M, Brasnu DF. Objective voice quality analysis before and after
920 onset of unilateral vocal fold paralysis. J Voice. 2001 Sep;15(3):351–61.
- 921 62. Ma Y, Xu X, Hou G, Zhou L, Zhuang P. Acoustic analysis in patients with unilateral arytenoid dislocation
922 and unilateral vocal fold paralysis. Lin Chung Er Bi Yan Hou Jing Wai Ke Za Zhi. 2016
923 Feb;30(4):268–71.
- 924 63. Misono S. The Voice and the Larynx in Older Adults: What's Normal, and Who Decides? JAMA
925 Otolaryngol Head Neck Surg. 2018 Jul 1;144(7):572–3.
- 926 64. Eadie T, Sroka A, Wright DR, Merati A. Does knowledge of medical diagnosis bias auditory-perceptual
927 judgments of dysphonia? J Voice. 2011 Jul;25(4):420–9.
- 928 65. Helou LB, Solomon NP, Henry LR, Coppit GL, Howard RS, Stojadinovic A. The role of listener
929 experience on Consensus Auditory-perceptual Evaluation of Voice (CAPE-V) ratings of
930 postthyroidectomy voice. Am J Speech Lang Pathol. 2010 Aug;19(3):248–58.
- 931 66. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of
932 dysphonic voice. J Voice. 2006 Dec;20(4):527–44.
- 933 67. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based
934 (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. J
935 Voice. 2007 Sep;21(5):576–90.
- 936 68. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use
937 Interpretable Models Instead. Nat Mach Intell. 2019 May;1(5):206–15.
- 938 69. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and Facial Biomarkers of
939 Depression based on Motor Incoordination and Timing. In: Proceedings of the 4th International
940 Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: Association for Computing
941 Machinery; 2014. p. 65–72. (AVEC '14).
- 942