Automatic Classification of Strain in the Singing Voice Using Machine Learning☆

*Yuanyuan Liu, †Mittapalle Kiran Reddy, ‡Madhu Keerthana Yagnavajjula, §Okko Räsänen, ¶Paavo Alku, *Tero Ikävalko, **Tua Hakanpää, ‖Aleksi Öyry, and *Anne-Maria Laukkanen, *§*Tampere,* ¶‖*Espoo,* **Turku, Finland,* †*Raichur, and* ‡*Kharagpur, India*

Summary: Objectives. Classifying strain in the singing voice can help protect professional singers from vocal overuse and support singing training. This study investigates whether machine learning can automatically classify singing voices into two levels of perceived strain. The singing samples represent two genres: classical and contemporary commercial music (CCM).

Methods. A total of 324 singing voice samples from 15 professional normophonic singers (nine female, six male) were analyzed. Nine singers were classical, and six were CCM singers. The samples consisted of syllable strings produced at three to six pitches and three loudness levels. Based on expert auditory-perceptual ratings, the samples were categorized into two strain levels: *normal-mild* and *moderate-severe*. Three acoustic feature sets (mel-frequency cepstral coefficients (MFCCs), the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), and wavelet scattering features) were compared using two classifier models [support vector machine (SVM) and multilayer perceptron (MLP)]. Feature selection was performed using recursive feature elimination, and the Mann-Whitney *U* test was used to assess the discriminative power of the selected features.

Results. The highest classification accuracy of 86.1% was achieved using a subset of wavelet scattering features with the MLP classifier. A comparison between individual features showed that the first MFCC coefficient, representing spectral tilt, exhibited the greatest between-class separation.

Conclusion. This study demonstrates that machine learning models utilizing selected acoustic features can classify perceptual strain of singing voices automatically with high accuracy. These preliminary findings highlight the potential for larger studies involving more diverse singer groups across different genres.

Key Words: Auditive-perceptual evaluation—Support vector machine—Multiple layer perceptron—Fisher vector—Wavelet scattering coefficients—Mel-frequency cepstral coefficients.

INTRODUCTION

Background

Voice production involves a complex interplay of physiological processes primarily within the respiratory system, larynx, and vocal tract.[1] To evaluate voice quality, various auditory-perceptual scales have been developed, primarily to identify vocal abnormalities that may signal organic or functional voice disorders. Among these, GRBAS and CAPE-V[2] are two of the most widely used clinical tools for assessing characteristics such as roughness, breathiness, and strain. Strain is described as a psychoacoustic impression of hyperfunctional phonation.[3]

Vocal hyperfunction is defined as excessive or imbalanced muscle activity[4] and is often considered the primary cause of vocal fold traumas, such as polyps and nodules, due to the excessive tissue loading it induces.[5] Strain may be further conceptualized as the auditory perception of effort in voicing,[6,7] whereas vocal effort refers to the self-perceived exertion during voicing.[8] Some researchers use the term strain synonymously with vocal overloading.[9,10]

Impact stress, the pressure per unit area during vocal fold collision, is regarded as the most damaging mechanical factor in voice production, acting perpendicular to the vocal fold tissue fibers.[11] Impact stress increases with fundamental frequency (F0), intensity, and adduction.[12] Singing involves much wider F0 and intensity ranges than speech, and different singing styles vary in phonation types along a continuum from breathy (low adduction) to pressed (high adduction).[13,14] Sonninen et al identified imbalanced muscle function related to register control as a vocal loading factor.[15] They viewed "open singing"—singing at high pitches without adequately reducing the vibrating mass of the vocal folds through thyroarytenoid muscle activity—as particularly hazardous for vocal health. Their X-ray studies revealed that "open singing" involves greater vocal fold strain (ie, increased fold length per pitch rise) compared to "covered singing," a technique typical of classically trained singers.[16] A loud, yell-like singing style

Accepted for publication March 25, 2025.

☆ The study was supported by The Academy of Finland (grant number 356528 awarded for the project "Vocal efficiency and economy in loud classical Operatic and Contemporary Commercial Music (CCM) singing styles compared to loud speech."

From the *Speech and Voice Research Laboratory, Tampere University, Tampere 33100, Finland; †Department of Computer Science and Engineering, Indian Institute of Information Technology Raichur, Raichur 584135, Karanataka, India; ‡Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India; §Unit of Computing Sciences, Tampere University, Tampere 33720, Finland; ¶Department of Information and Communications Engineering, Aalto University, Espoo 02150, Finland; **Faculty of Education, University of Turku, Turku 20014, Finland; and the ‖Aalto Acoustics Lab, Aalto University, Espoo 02150, Finland.

Address correspondence and reprint requests to: Yuanyuan Liu, Speech and Voice Research Laboratory, Tampere University, Tampere 33100, Finland. E-mail: yuanyuan.liu@tuni.fi

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

© 2025 The Authors. Published by Elsevier Inc. on behalf of The Voice Foundation. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.1016/j.jvoice.2025.03.040

known as belting[13,14] exemplifies "open singing." A review in[17] discussed the pathology of singing voices from both scientific and artistic perspectives, concluding that vocal problems among singers often stem from incorrect technique or vocal abuse. Even minor voice pathologies can severely impact professional singers, with strained singing manifesting as hoarseness, reduced vocal range, timbre changes, fatigue, and throat pain. In summary, strain in the singing voice is likely greater and poses a higher risk of vocal fold damage than strain in speaking voices. However, research on singing voice strain remains sparse.

Belting has been characterized as employing a more strained or pressed voice quality than neutral or classical singing.[18] This difference appears linked to higher subglottic pressure and specific voice source characteristics, such as a longer glottal closed time, shorter glottal closing phase, and smaller spectral tilt.[19] Belting is often associated with voice disturbances,[20] although singers and singing teachers distinguish between healthy and unhealthy production methods for belting.[21] This raises an intriguing question: can automatically extracted acoustic features indicate whether a singing voice is produced with excessive strain, irrespective of pitch, loudness, or genre-related timbre differences? Early recognition and management of voice strain could help singers prevent further damage, maintain vocal health, and improve their technique. Automatic classification of singing voice strain (eg, mild vs. severe) could significantly benefit singers by optimizing performance and facilitating effective learning.

According to a comprehensive review,[22] voice research has primarily focused on breathiness and roughness. Meanwhile, voice quality evaluation has predominantly studied healthy and pathological speaking voices,[23] leaving the singing voice less explored. The following section presents a brief review of acoustic features and machine learning techniques employed in the automatic assessment of strain, breathiness, and other voice quality dimensions in speaking voices.

Previous work

Automatic voice analysis was investigated in[24] using a multiple linear regression model to predict auditory-perceptual breathiness ratings. The study analyzed over 900 voice samples, comprising voiced segments from continuous speech and sustained vowel /a/ samples produced by both healthy and dysphonic speakers. A total of 28 acoustic features were compared. Statistical analysis revealed that cepstral peak prominence (CPP) had the highest correlation with breathiness ratings, followed by glottal-to-noise excitation (GNE). To predict breathiness ratings, a multiple linear regression model was developed using nine features: high-frequency noise level, amplitude difference between the first and second harmonics (H1-H2), CPP, harmonics-to-noise ratio (HNR), period standard deviation, GNE, jitter, and shimmer. The model demonstrated a strong correlation between its predictions and the expert-assigned ground truth breathiness ratings.

A small-scale study in[22] analyzed 28 dysphonic vowel samples, revealing a strong correlation between perceptual ratings of voice strain and acoustic features such as CPP, sharpness, and several spectral measurements. Similarly,[25] explored the acoustic correlates of self-perceived and auditorily perceived effort (ie, strain) in untrained speakers. Their findings indicated that the most reliable acoustic correlates were sound pressure level (SPL), low-to-high spectral ratio, and HNR.

In,[26] spectral and cepstral features were extracted from sustained vowels and continuous speech produced by 23 dysphonic speakers, whose dysphonia was primarily characterized by strained voice quality, and 23 healthy speakers. The results indicated that the acoustic features demonstrated moderate to high correlations with perceptual ratings of strain severity in the voice samples. The study concluded that the spectral and cepstral features effectively differentiated between strained and normal voices.

In,[27] an automatic voice assessment system was developed to classify pathological and healthy voices into four levels of grade, one of the five perceptual items in the GRBAS scale. A total of 65 acoustic features were utilized, including traditional methods such as mel-frequency cepstral coefficients (MFCCs) and GNE, as well as nonlinear dynamical analyses. To reduce feature redundancy, four feature selection techniques were applied. Multiclass classification was performed using support vector machines (SVM) and extreme learning machine (ELM) classifiers, yielding moderate correlation with expert grade ratings.

The same task was investigated in,[28] which employed 44 acoustic features (including MFCCs, CPPs, and long-term average spectrum) and a deep belief network to classify sustained vowels and running speech samples into four voice severity categories as defined by GRBAS. The results demonstrated a moderate correlation between acoustic features and overall dysphonia severity.

In,[29] linear regression was used to explore the relationship between auditory-perceptual ratings of voice quality (breathiness, roughness, and strain) and acoustic features such as HNR, CPP, and low-to-high spectral ratio (L/H ratio). The voice samples consisted of sustained vowels and passage readings from patients with muscle tension voice disorders and healthy speakers. The study found that the L/H ratio effectively predicted strain.

Recent machine learning studies have also utilized raw voice waveforms as classifier inputs, bypassing the need for handcrafted acoustic features. A notable example is,[30] where a one-dimensional convolutional neural network was employed to predict GRBAS scores from pathological voice samples. Among the GRBAS items, strain prediction achieved the highest accuracy and F1-score.

Goals

Signal processing and ML techniques have been successfully applied to the automatic assessment of perceptual ratings in the speaking voice. This study explores whether

these techniques can also be used to classify the level of strain in the singing voice. Specifically, the study aims to:

1) Determine whether automatic ML classifiers can distinguish between singing voices with normal to mild strain (*normal-mild*) and those with moderate to severe strain (*moderate-severe*).
2) Apply a feature selection algorithm to identify the optimal subset of acoustic features for classification and compare the distributions of individual features across the two strain classes. The results demonstrate that strain levels in singing voices can be automatically classified with relatively high accuracy.

The paper is structured as follows: Section 2 describes the singing voice dataset used in this study. Section 3 outlines the methods. Experimental results are presented in Section 4. Section 5 analyzes the acoustic features investigated. Discussions are provided in Section 6, and conclusions are drawn in Section 7.

DATA

Voice recording

The singing voice samples used in this study were sourced from the Multiple Modality Singing Voice Database (MMSVD), which was collected at the Speech and Voice Research Laboratory of Tampere University. MMSVD contains multi-channel recordings of singing, including acoustic voice, oral air pressure, oral flow, and electroglottography (EGG) signals. These signals were recorded from 15 trained singers, of whom nine were classically trained (two males, seven females) and six were contemporary commercial music (CCM) singers (four males, two females). The participants' ages ranged from 25 to 77 years (mean = 49.7 years, SD = 14.9 years), and all were vocally healthy according to their own reports and their Voice Handicap Index sum scores. The singers were recruited via social media, and all participated voluntarily, providing written consent.

The singers sang a syllable string consisting of five utterances of one consonant-vowel syllable (/pa/, /pe/, and /po/), for example, a singing syllable string of "pa pa pa pa pa". Each string was repeated twice in three different pitches, one octave apart, covering a total pitch range of two octaves. The pitches used by each singer are shown in Table 1. The F0 range of the samples was 186-816 Hz for the females and 94-656 Hz for the males. The singing tasks were first performed at the lowest perceivable loudness (phonation threshold) that the singer could produce. Thereafter the singers produced the tasks in medium loudness and in loud stage voice (mezzoforte and forte in music terms). The singers were instructed to evenly stress each of the five repeated syllables and to keep the tempo between 120 and 180 syllables per minute. A reference tempo was provided by one of the authors using a metronome.

During the singing task, the singers wore the Glottal Enterprise MA-1 (large) flow mask over their mouth and nose to record aerodynamic data for other studies. The acoustic voice was recorded using a Brüel and Kjaer 4188 microphone connected to a Brüel and Kjaer Mediator 2238 sound level meter. The microphone was positioned 30 cm from the singer's lips. To measure SPL, a calibration tone generated by a Brüel and Kjaer 4230 calibrator was recorded. The signals were recorded using KAY CSL MODEL 4500 software and converted to wav format for later analysis. The recordings were made with a sampling frequency of 44.1 kHz and a resolution of 16 bits. For the current study, all recorded voice signals were downsampled to 22.05 kHz.

Perceptual rating of the level of strain

The level of strain of each recorded syllable string was perceptually assessed by two raters. Both raters were

TABLE 1.
Participants and Their Singing Pitches

Singer	Pitch 1	Pitch 2	Pitch 3	Pitch 4	Pitch 5	Pitch 6
CCM M1	C#4	G4				
CCM M2	G#2	D3	G#3	D4	A4	
CCM M3	A2	A3	A4			
CCM M4	A#2	E3	A#3	E4	A#4	E5
CCM F1	G3	G4	G5			
CCM F2	G3	G4	G5			
Classical M1	A2	D#3	A3	D#4	A4	
Classical M2	G2	G3	D#4	F#4		
Classical F1	G3	G4	G5			
Classical F2	G3	G4	G5			
Classical F3	G3	G4	G5			
Classical F4	G3	G4	G5			
Classical F5	G3	G4	G5			
Classical F6	G3	G4	G5			
Classical F7	G3	G4	G5			

CCM refers to a singer with a contemporary commercial music background and Classical refers to a singer with a background in classical music.

university-educated vocologists with PhDs in vocology and a strong background in singing. One rater was a professional CCM singer and singing teacher, while the other one had extensive training in classical singing and worked as a professional voice trainer. The selection criteria for the raters included: (a) extensive training in voice science and significant experience in perceptual analysis, (b) a background in singing, and (c) representation of different musical genres (classical operatic and CCM - pop/jazz, rhythm music). The raters evaluated the strain level of each recorded syllable string using a discrete scale from 0 to 5, where 0 = no strain at all, and 1 = mild, 2 = rather mild, 3 = moderate, 4 = severe, or 5 = extreme strain. Each rater scored the voices twice in two separate sessions on different days. Both the intra-rater reliability (consistency between the first and second rating for each rater) and the inter-rater reliability between the raters was studied with an intraclass correlation coefficient (ICC). Intra-rater reliability analysis yielded ICC 0.895 for one rater (95% Confidence Interval (CI) 0.869-0.915, $P < 0.001$) and ICC 0.891 for the other (CI 0.864-0.912, $P < 0.001$). For average consistency between the raters the ICC was 0.715 (CI 0.646-0.771, $P < 0.001$). The final strain score for each sample was calculated as the mean of the ratings from both raters across the two sessions. The distribution of mean strain scores is illustrated in Figure 1.

For the purposes of this study, the voice samples were divided into two classes based on their mean strain scores: *normal-mild*, syllable strings with mean strain score between 0.0 and 2.0 (exclusive); *moderate-severe*, syllable strings with mean strain scores between 2.0 and 5.0 (inclusive). The *normal-mild* class contained 169 syllable strings from 14 singers (95 by female singers and 74 by male singers). The *moderate-severe* class included 155 syllable strings from 15 singers (67 by female singers and 88 by male singers). The duration of all the 324 voices ranged from 1.4 to 6.2 seconds, with an average of 2.9 seconds.

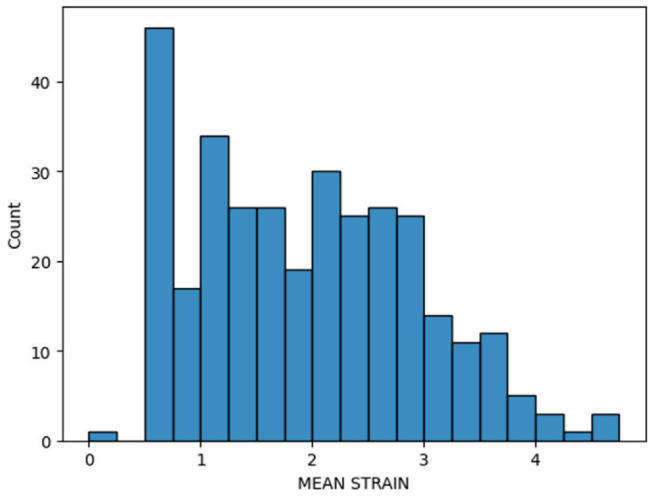


FIGURE 1. Histogram of the mean strain scores for all singing voice samples.

METHODS

To investigate the automatic classification of strain in the singing voices, automatic ML systems were developed following a conventional pipeline architecture, consisting of three main components: *feature extraction*, *classifier*, and *feature selection*. Each component plays a critical role in ensuring the system's effectiveness. This section provides a detailed description of these three components.

Feature extraction

Three acoustic feature sets were compared in the study, as presented in Sections 3.1.1, 3.1.2, and 3.1.3. All features were initially computed on a frame-wise basis and subsequently aggregated into fixed-length feature vectors for each syllable string using temporal statistical functionals. In addition to commonly used temporal functionals (eg, mean, standard deviation), frame-wise features were also combined into syllable-wise feature vectors through the Fisher vector encoding,[31] as described in Section 3.1.4.

MFCCs

As discusses in Section 1.2, MFCCs have been widely utilized in numerous studies on speech and voice research, making them one of the feature sets selected for comparison in this study. MFCCs provide a compressed, perceptually-driven representation of the short-term power spectrum of sound,[32] as they are derived using the mel scale, which models human auditory perception by emphasizing lower frequencies where the ear is more sensitive for frequency discrimination. In this work, the first 13 MFCCs were extracted, and are denoted by MFCC0, MFCC1, …, MFCC12, respectively. The MFCC0 represents the overall energy of the signal, and MFCC1 corresponds to spectral tilt (balance of energy between low and high frequencies). The other coefficients (MFCC1-MFCC12) capture different aspects of the sound's timbre and quality, which are essential for distinguishing between various types of voices or speech sounds. In this work, 13 MFCCs were computed using a frame length of 25 ms, a hop length of 10 ms, and a Hann window, following the MFCC implementation from the librosa library.[33] The Hann window was applied during the short-time spectral analysis to minimize spectral leakage, which helps maintain frequency resolution by reducing the effect of abrupt signal truncation. The frame-level MFCCs were smoothed by a moving average filter with a window size of three frames. From the smoothed frame-level MFCCs, temporal statistical functionals (mean, standard deviation, skewness, and kurtosis) were computed, resulting in a 52-dimensional vector per syllable string. Since the mean is a temporal statistical functional commonly used in previous studies,[27,34] MFCC features were also computed using the mean as the sole functional, producing in a 13-dimensional feature vector per syllable string. This MFCC feature is referred to as MFCC_mean in this study.

Extended Geneva minimalistic acoustic parameter set (eGeMAPS)

The second feature set used in this study was the eGeMAPS.[35] The eGeMAPS set consists of a variety of time-domain and frequency-domain parameters, such as F0, jitter, shimmer, HNR, formants, and spectral slope, which are considered robust indicators of voice perturbations. These parameters, referred to as 25 low-level descriptors (LLDs), are computed frame-wise in eGeMAPS using a frame length of 25 ms and a hop length of 10 ms. To derive fixed-length feature vectors for each syllable string, the frame-wise computed parameters were aggregated using various temporal statistical functionals, including arithmetic mean, coefficient of variation, and percentiles. This process resulted in an 88-dimensional feature vector per syllable string. In the present work, the eGeMAPS features were extracted using the Python package *openSMILE*,[36] a widely used toolkit for speech analysis. OpenSMILE provides a standardized approach to extract prosodic (eg, pitch, intensity, duration) and spectral (eg, formants) characteristics of speech and voice signals, which are crucial, for example, in voice quality assessment.

Wavelet scattering coefficients (WSCs)

The third feature type selected in this study was based on the wavelet scattering network.[37] These features are referred to as the WSCs. The wavelet scattering network provides a time-frequency representation that is invariant to translation, stable under time-warping deformations, and minimizes within-class variations while preserving discriminability across classes.[38] The network processes input signals through stages, each consists of a cascade of wavelet transform, complex modulus, and low-pass filtering operations.[38] The scattering coefficients produced at each stage exhibit local stability and translation invariance, which are crucial for classification tasks.[38] The use of scattering coefficients has demonstrated high performance in various applications, such as phoneme recognition and the detection and classification of pathological speech.[38–40] WSCs are particularly useful for capturing fine-grained pitch variations and subtle articulation differences, making them well-suited for analyzing singing voices and voice disorders.

In principle, the scattering network can include multiple stages. However, a two-stage scattering network is typically sufficient for most audio applications, as higher-order scattering coefficients generally contain negligible energy.[38,39,41] Therefore, a two-stage scattering network was employed in this study. The two key parameters of the network are: (i) the number of wavelets per octave in the filterbanks used for computing the wavelet transform at each stage, and (ii) the invariance scale. Scattering coefficients were extracted using an invariance scale of 150 ms, with four filters per octave in the first stage and one filter per octave in the second stage. Based on empirical testing, the invariance scale was set to 150 ms, which provided the best classification performance in our study. The scattering features were computed using Matlab. The scattering network generated an $N \times P$ feature matrix, where P represents the number of time windows (or frames) and N is the fixed number of scattering coefficients per time window. The scattering coefficients were log-transformed and averaged along the time axis to produce a 158-dimensional feature vector for each syllable string.

Fisher vector encoding

The Fisher vector (FV) encoding is a powerful technique widely used in image and audio research to transform variable-length feature sets into fixed-length vectors, enabling their application in standard ML models. By modeling the distribution of local descriptors (such as MFCCs) using a Gaussian Mixture Model (GMM), the FV encoding captures higher-order statistics, including the mean and covariance deviations of the descriptors from the GMM components. This results in a rich representation that effectively encodes the underlying structure and variability within the data. FV encoding has been successfully applied to various tasks, such as action and event recognition,[42] emotion recognition,[43] speaker verification,[44] and speech recognition.[45]

In this work, we explored FV encoding as an alternative to the statistical temporal functionals to aggregate frame-wise MFCCs, eGeMAPS LLDs, and WSCs into syllable-level vectors, as described in Section 3.1.1, 3.1.2, and 3.1.3. The FV encoding was computed by partitioning the singing voice data into training and testing sets using a leave-one-singer-out cross validation (LOSO-CV) procedure. In LOSO-CV approach, the syllable strings of 14 singers were used as the training set, while the samples of the remaining singer were served as the test set. This process was repeated until every singer had been tested. This LOSO-CV procedure was consistently applied for the feature selection (Section 3.2) and classification experiments (Sections 3.3 and 4).

To compute the FV encoding, a GMM with two Gaussians was first trained. Then FVs were then derived form the GMM for both the training and test sets. A classifier for strain level prediction was trained with the training set FVs and evaluated on the test set FVs. This procedure was repeated for each singer, and the overall classification accuracy was calculated by aggregating the predictions from all 15 singers along with their corresponding strain level labels. The dimensionality of the FV-encoded feature vectors per syllable string was 54 for MFCCs, 102 for eGeMAPS, and 634 for WSCs when the FV encoding was applied to frame-level features.

Feature selection

To identify the most effective features for the classification task, feature selection was performed using the recursive feature elimination (RFE) algorithm.[46] RFE begins with the complete feature set and iteratively removes the least

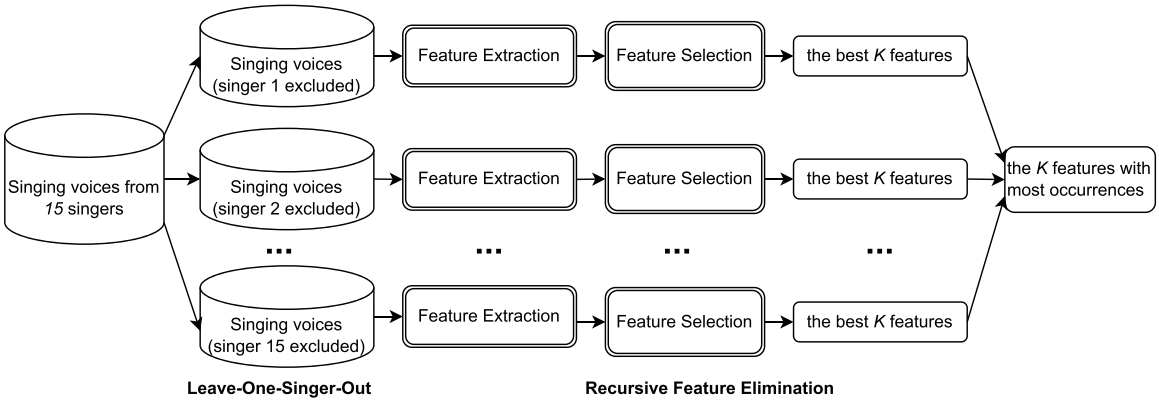


FIGURE 2. A schematic view of the process to select K best features using Leave-One-Singer-Out (LOSO) and Recursive Feature Elimination (RFE).

important features until the desired number of features is obtained. In this process, a Ridge classifier[47] was employed to determine feature importance.

The Ridge classifier applies Ridge regression by first converting binary class labels to $\{-1, 1\}$, treating the task as a regression problem. The model optimizes its coefficients to minimize a penalized residual sum of squares between the observed targets and the predicted targets from a linear approximation.[48] The resulting optimal model coefficients are then used to eliminate the least important features. The RFE implementation from the Python library *scikit-learn* was used in this study.

Figure 2 illustrates the RFE-based feature selection process conducted in this study. RFE was performed by varying the number of selected features (K) from 2 to $D-1$, where D is the full dimensionality of the feature set (eg, $D = 52$ for the MFCCs). Using the LOSO-CV approach described in Section 3.1.4, K features were first selected for each of the 15 singers through RFE, resulting in 15 subsets with the dimensionality of K. From these subsets, the K features with the highest occurrences were finally chosen. The procedure was repeated for each of the four full feature sets described in Sections 3.1.1–3.1.3. The resulting four feature sets with dimensionality of K were used for classifier training and testing with the LOSO-CV approach described in Section 3.1.4.

Classifiers

This study compared two widely-used and powerful ML models: SVMs and multilayer perceptrons (MLPs). Both classifiers were evaluated by partitioning the singing voice data into training and testing sets using the LOSO-CV approach described in Section 3.1.4. Performance is reported as the overall accuracy computed over the voices of all 15 singers.

SVM

SVMs have been proven to be highly effective for binary classification, particularly when high-dimensional feature representations are used in conjunction with small training datasets, such as in the automatic assessment of pathological voices.[49] In this study, SVMs with a nonlinear radial basis function (RBF) kernel were employed. The RBF kernel is defined by the formula $K(x, x') = exp(-\gamma\|x - x'\|^2)$, where x and x' are two samples in dataset X. The kernel coefficient γ was set as $\gamma = 1/(d^*\sigma_X^2)$, where d represents the feature dimensionality and σ_X^2 is the variance of dataset X. The regularization parameter C was set as 1.0.

MLP

MLPs are another classifier type widely used in voice and speech research. For example, a one-hidden-layer MLP was used to detect pathological voices in.[50] In,[51] two-hidden-layer MLPs were trained with wavelet packet transform features successfully detected voice disorders characterized by roughness, breathiness, and strain, separately, which achieved good classification accuracies. In this study, MLPs with one hidden layer were utilized. The number of hidden neurons was set to one less than twice the input dimensionality. For example, for the eGeMAPS festure set, the number of hidden neurons was calculated as: $2 \times 88 - 1 = 175$. Tanh was used as the activation function, as it allows for faster learning by centering values around zero and helps capture both positive and negative variations in singing voice data. The models were trained until numerical convergence was detected (no significant change of training loss) or until a predefined maximum number of epochs was reached. The optimizer used was Adam with its default learning rate of 0.001, as provided by *scikit-learn*. Adam is well-suited for training deep learning models as it adapts learning rates for different parameters, improving stability and convergence.

CLASSIFICATION RESULTS

Classification accuracy for the RFE-selected feature sets is presented in Tables 2 and 3 for the SVM and MLP classifier, respectively. Both tables also show the dimensionality (ie, parameter K) of the optimal subset selected by the

TABLE 2.
Classification Accuracy (%) for the RFE-Selected Feature Sets Using the SVM Classifier

Feature	Dimensionality	Accuracy
MFCC_mean	5	82.4
MFCCs	14	84.3
eGeMAPS	79	82.7
WSCs	107	85.5

TABLE 3.
Classification Accuracy (%) for the RFE-Selected Feature Sets Using the MLP Classifier

Feature	Dimensionality	Accuracy
MFCC_mean	3	84.3
MFCCs	23	84.6
eGeMAPS	19	84.6
WSCs	21	**86.1**

RFE procedure for each of the four feature sets. The results demonstrate that the MLP classifier constantly achieved higher accuracy across all four feature sets. The highest accuracy 86.1% was obtained using the MLP classifier with the RFE-selected WSCs feature set. Similarly, the WSCs feature set yielded the best accuracy (of 85.5%) for the SVM classifier. As noted in Section 3.1.4, the FV encodings varied across the LOSO-CV folds, rendering the RFE-based feature selection method inapplicable for the FV-encoded features.

Confusion matrices for the classifications performed by the best classifier (ie, MLP) are shown for the RFE-selected MFCC_mean, MFCCs, eGeMAPS, and WSCs features in Tables 4, 5, 6, and 7, respectively. The confusion matrices reveal that the superior accuracy provided by the WSCs features is primarily attributed to the lower number of misclassifications in the *moderate-severe* class (14 in Table 7).

Table 8 presents the classification accuracy of the two classifiers when using the original full feature sets. A comparison of these results with the accuracy values reported in Tables 2 and 3 reveals that the RFE-based feature selection approach significantly improved classification performance across all feature sets and for both classifiers. For instance, the classification accuracy of the MFCC features improved from 80.6% to 84.3% with SVM and from 77.2% to 84.6% with MLP when RFE-based feature selection was applied. For reference, the results using the FV encoding as an alternative to traditional temporal functionals are also introduced in Table 8. The FV encoding did not demonstrate any clear advantages over traditional temporal functionals. Furthermore, increasing the number of GMM components for FV computation did not yield any performance improvements.

TABLE 4.
Confusion Matrix for the RFE-Selected MFCC_Mean Feature Set Using the MLP Classifier (Accuracy = 84.3%)

Predicted True	normal-mild	moderate-severe
normal-mild	139	30
moderate-severe	21	134

TABLE 5.
Confusion Matrix for the RFE-Selected MFCCs Feature Set Using the MLP Classifier (Accuracy = 84.6%)

Predicted True	normal-mild	moderate-severe
normal-mild	142	27
moderate-severe	23	132

TABLE 6.
Confusion Matrix for the RFE-Selected eGeMAPS Feature Set Using the MLP Classifier (Accuracy = 84.6%)

Predicted True	normal-mild	moderate-severe
normal-mild	140	29
moderate-severe	21	134

TABLE 7.
Confusion Matrix for the RFE-Selected WSCs Feature Set Using the MLP Classifier (Accuracy = 86.1%).

Predicted True	normal-mild	moderate-severe
normal-mild	138	31
moderate-severe	14	141

TABLE 8.
Classification Accuracies (%) of Both Classifiers Using the Original Full Feature Sets Without the RFE Feature Selection

Feature	Dimensionality	SVM	MLP
MFCC_mean	13	81.5	75.0
MFCCs	52	80.6	77.2
eGeMAPS	88	76.9	81.5
WSCs	158	84.3	75.9
FV_MFCC	54	80.2	75.3
FV_eGeMAPS(LLD)	102	77.2	77.8
FV_WSC	634	79.6	78.7

RANKING OF INDIVIDUAL FEATURES

The results presented in Section 4 demonstrate that the RFE-based feature selection improved the accuracy of the automatic strain level classification systems compared to

using the full feature sets. This section provides further analysis of the selected features by studying their rankings within each of the four full sets. Additionally, class separability achieved by the highest-ranking individual feature of each set is visualized by feature value histograms and quantified by statistical tests.

Feature rankings for each feature set were determined using the RFE procedure and the LOSO-CV approach described in Section 3.2. The ranking procedure consisted of the following steps:

- **Initial ranking per fold:** For a feature set with a dimensionality of D (eg, $D = 88$ for the eGeMAPS set), each feature was assigned a rank, represented by an integer between 1 and D, in each LOSO-CV fold using the RFE procedure.
- **Averaging across folds:** The ranks obtained in all the 15 LOSO-CV folds were averaged to compute a real-valued mean rank for each feature in the set.
- **Final ranking:** Based on the mean rank, the features were arranged in a descending order. The obtained final rankings are presented in Tables 9, 10, 11, 12 for the RFE-selected MFCC_mean, MFCCs, eGeMAPS, and WSCs features, respectively.

Additionally, statistical analysis of feature values between the two strain classes was conducted using the Mann-Whitney U test (also known as the Wilcoxon rank-sum test).[52] The null hypothesis was rejected at $P < 0.05$. For each feature, the Mann-Whitney U test provided three values: effect size (r), significance (p), and z-score (z). These parameters are given in the third, fourth, and fifth columns of Tables 9, 10, 11, 12. It is worth noting that features can also be ranked based on the effect size (r) from the Mann-Whitney U test, with larger r values indicating higher feature importance.

Figure 3 presents the histograms of the top-ranking individual feature of the MFCC_mean, MFCCs, eGeMAPS, and WSCs feature sets, along with their respective effect size (r) and P value given by the Mann-Whitney U test. These histograms show that the best-performing features in the MFCC_mean, MFCCs, and eGeMAPS sets demonstrate clear separation between the two strain classes. Additionally, these features exhibit high r values (ranging from 0.550 to 0.610) with statistically significant differences between classes.

In contrast, the best ranking WSC feature shows considerable overlap between the two classes, a low r value (0.070), and no statistically significant difference. Poor class separability given by the best-ranked WSC feature is contrary to what might be expected by the results reported in Section 4, which show that the WSC features showed the highest accuracy in the automatic classification of strain. This suggests that the information regarding strain is distributed across many WSC feature dimensions that jointly contribute to accurate classification.

TABLE 9.
The Three Best Features From the MFCC_mean Set, as Identified by RFE With the MLP Classifier

Feature	Mean Rank	r	P	z
MFCC2_mean	1.0	0.551	< 0.001	9.916
MFCC4_mean	2.0	0.475	< 0.001	8.544
MFCC6_mean	3.2	0.260	< 0.001	4.682

The 2nd column shows the mean feauture rank across all folds in the LOSO-CV, and the remaining columns denote the effect size (r), P value (P), and z score (z) from the Mann-Whitney U test when comparing the features across the classes.

TABLE 10.
The 23 Selected Features of the MFCC Set, Computed by the RFE Procedure

Feature	Mean Rank	r	P	z
MFCC1_mean	2.3	0.562	< 0.001	10.116
MFCC4_mean	3.3	0.475	< 0.001	8.544
MFCC0_var	3.5	0.356	< 0.001	6.406
MFCC0_kurt	6.1	0.037	0.509	−0.661
MFCC2_mean	6.3	0.551	< 0.001	9.916
MFCC1_kurt	7.5	0.101	0.070	−1.811
MFCC1_var	9.1	0.312	< 0.001	−5.621
MFCC3_skew	9.3	0.399	< 0.001	−7.180
MFCC0_skew	11.5	0.148	0.008	2.664
MFCC9_skew	11.5	0.288	< 0.001	5.179
MFCC8_skew	13.1	0.153	0.006	2.751
MFCC9_kurt	13.1	0.117	0.036	2.097
MFCC1_skew	16.0	0.431	< 0.001	−7.752
MFCC6_mean	17.5	0.260	< 0.001	4.682
MFCC0_mean	18.9	0.316	< 0.001	−5.684
MFCC6_kurt	19.5	0.007	0.905	−0.121
MFCC2_skew	19.5	0.340	< 0.001	−6.122
MFCC8_kurt	21.3	0.090	0.105	−1.620
MFCC11_var	21.9	0.229	< 0.001	−4.116
MFCC4_var	24.1	0.023	0.682	0.410
MFCC7_mean	25.1	0.051	0.357	0.922
MFCC10_kurt	25.5	0.217	< 0.001	3.905
MFCC7_skew	29.2	0.097	0.080	1.749

These features achieved an accuracy of 84.6% with the MLP classifier. The 2nd column shows the mean rank across all folds in LOSO-CV. For the MFCCs, "mean," "var," "skew," "kurt" refer to the mean, variance, skewness, and kurtosis, respectively. In this table, r, P, z represent the effect size, significance, and z-score of U statistic of the Mann-Whitney U test for each feature, respectively, comparing the normal-mild and moderate-severe classes.

DISCUSSION

This study investigated the automatic classification of the level of strain in singing voices. Feature set dimensionalities were reduced using the RFE technique to identify the best-functioning feature sets. The highest classification accuracy (86.1%) was obtained using the WSC feature set, whose dimension was reduced with RFE, and using the MLP classifier. The other feature set. al.o resulted in high accuracy (≥84.3%) with the MLP. The RFE-based ranking between individual features revealed that the MFCC1 feature was

TABLE 11.
The 19 Selected Features of the eGeMAPS Set, Computed By the RFE Procedure

Feature	Mean Rank	*r*	*P*	*z*
mfcc1V_sma3nz_amean	2.7	0.610	< 0.001	10.986
F3amplitudeLogRelF0_sma3nz_stddevNorm	6.3	0.493	< 0.001	8.875
spectralFlux_sma3_amean	6.4	0.189	0.001	−3.407
spectralFluxV_sma3nz_amean	8.2	0.198	< 0.001	−3.559
loudness_sma3_meanRisingSlope	8.6	0.327	< 0.001	−5.892
equivalentSoundLevel_dBp	8.7	0.276	< 0.001	−4.968
loudness_sma3_percentile20.0	10.0	0.200	< 0.001	−3.607
MeanVoicedSegmentLengthSec	11.8	0.069	0.215	−1.240
mfcc3_sma3_amean	13.3	0.502	< 0.001	9.039
F2bandwidth_sma3nz_stddevNorm	14.4	0.197	< 0.001	−3.552
F2frequency_sma3nz_amean	15.1	0.283	< 0.001	−5.095
F1frequency_sma3nz_amean	16.9	0.306	< 0.001	−5.515
mfcc4V_sma3nz_amean	17.5	0.251	< 0.001	4.522
F2amplitudeLogRelF0_sma3nz_stddevNorm	18.7	0.473	< 0.001	8.511
F1frequency_sma3nz_stddevNorm	19.0	0.177	0.002	−3.181
mfcc3V_sma3nz_amean	19.1	0.479	< 0.001	8.630
slopeV0-500_sma3nz_amean	19.9	0.112	0.044	−2.014
F2frequency_sma3nz_stddevNorm	20.6	0.057	0.304	−1.028
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	23.4	0.173	0.002	−3.111

These features achieved an accuracy of 84.6% with the MLP classifier. The 2nd column shows the mean rank across all folds in LOSO-CV. For detailed descriptions of each parameter, the reader is referred to.[35] In this table, *r, P, z* represent the effect size, significance, and *z*-score of *U* statistic from the Mann-Whitney *U* test for each feature, respectively, comparing the normal-mild and moderate-severe classes.

TABLE 12.
The 21 Selected Features of the WSCs Set, Computed By the RFE Procedure

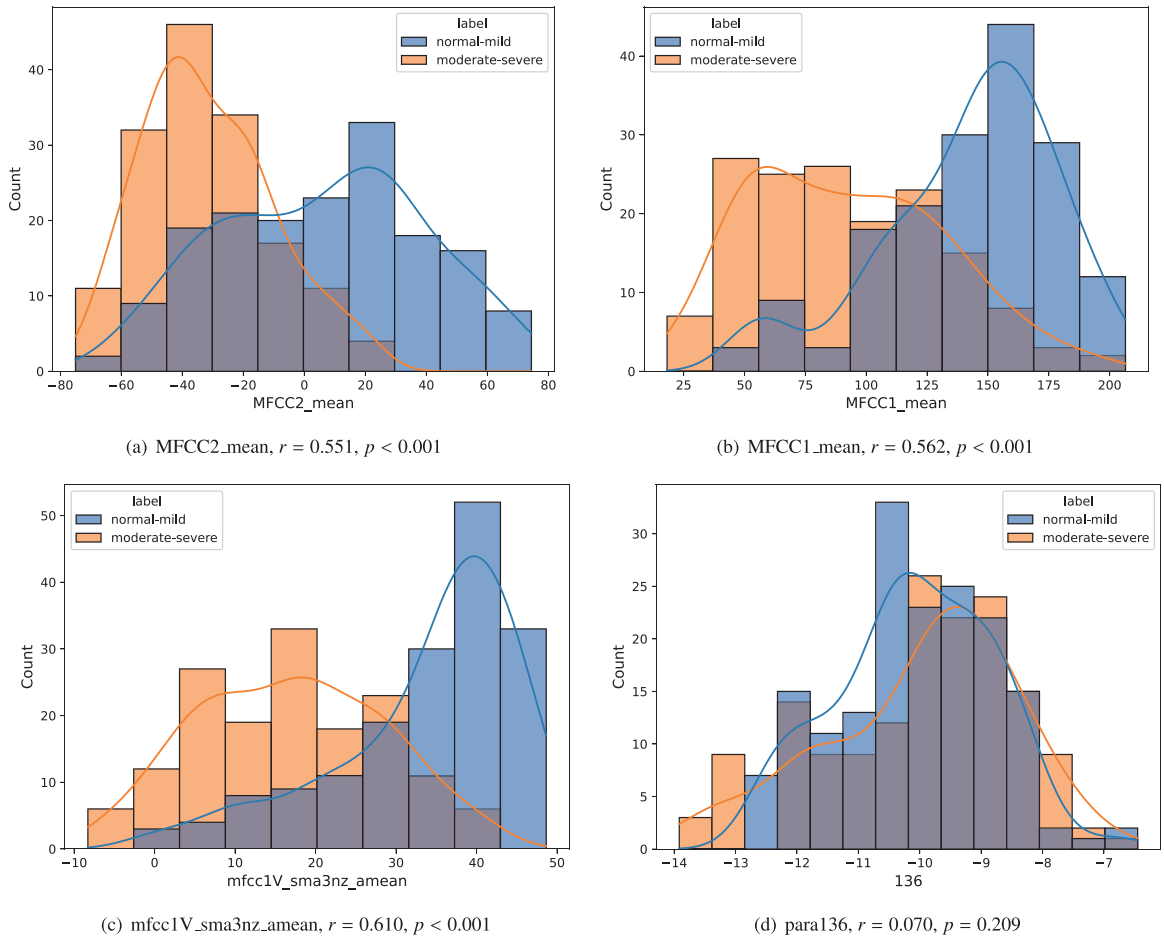
Feature	Mean Rank	*r*	*P*	*z*
para136	5.9	0.070	0.209	−1.258
para137	6.5	0.028	0.617	−0.500
para25	7.5	0.315	< 0.001	5.678
para157	8.5	0.257	< 0.001	4.630
para77	9.0	0.571	< 0.001	−10.287
para146	9.0	0.221	< 0.001	3.974
para20	9.5	0.246	< 0.001	4.429
para66	12.4	0.466	< 0.001	−8.384
para68	16.1	0.464	< 0.001	−8.353
para121	17.3	0.367	< 0.001	−6.615
para73	19.3	0.406	< 0.001	−7.302
para129	21.2	0.392	< 0.001	−7.055
para148	22.1	0.160	0.004	2.876
para128	27.6	0.380	< 0.001	−6.832
para38	29.2	0.354	< 0.001	−6.380
para44	32.5	0.446	< 0.001	−8.032
para40	39.2	0.439	< 0.001	−7.910
para75	39.6	0.462	< 0.001	−8.310
para19	41.1	0.264	< 0.001	4.752
para43	45.9	0.462	< 0.001	−8.312
para13	52.1	0.427	< 0.001	−7.691

These features achieved an accuracy of 86.1% with the MLP classifier. The 2nd column shows the mean rank across all folds in LOSO-CV. The 158 features of the WSC set are labeled as "para0 – para157." The *r, P,* and *z* represent the effect size, significance, and *z*-score of *U* statistic from the Mann-Whitney *U* test for each feature, respectively, when comparing the normal-mild and moderate-severe classes.

ranked as the best individual feature in two of the studied sets (the RFE-reduced MFCCs and eGeMAPS sets).

Wavelet scattering provides a rich time-frequency representation that preserves critical signal structures across multiple scales, making it robust to pitch and timbre variations in singing.[53] In comparison, MFCCs primarily capture the overall spectral shape, often at the expense of finer temporal details essential for identifying subtle vocal strain.[54] Designed for general perceptual audio analysis, MFCCs can overlook transient features critical in strain detection. Despite this limitation, the MFCC model demonstrated remarkable efficiency, achieving an accuracy of 84.3% using just three selected features. This indicates that MFCCs, while compact, can still effectively represent vocal strain characteristics when coupled with an appropriate feature selection algorithm like RFE. Similarly, while eGeMAPS encompasses a broad range of voice characteristics,[35] it appears slightly less effective than scattering coefficients in capturing intricate patterns of vocal strain. However, its performance (84.6% with 19 features) remains competitive and demonstrates the utility of feature sets that combine a diverse range of acoustic measures. The findings suggest that while wavelet scattering offers the most comprehensive representation for this classification task, MFCCs remain a powerful and efficient alternative, particularly in scenarios where simplicity and computational efficiency are prioritized.

The RFE algorithm improved performance across all feature sets by selecting the most important features. For



(a) MFCC2_mean, $r = 0.551$, $p < 0.001$

(b) MFCC1_mean, $r = 0.562$, $p < 0.001$

(c) mfcc1V_sma3nz_amean, $r = 0.610$, $p < 0.001$

(d) para136, $r = 0.070$, $p = 0.209$

FIGURE 3. Histograms of the best-ranked features from each feature sets for **(a)** MFCC_mean, **(b)** MFCCs, **(c)** eGeMAPS, and **(d)** WSCs. *r* represents the effect size ("class separation") and *P* the *P* value calculated by the Mann-Whitney *U* test for the feature under comparison between the *normal-mild* and *moderate-severe* classes. The solid curves denote smoothed distributions using kernel density estimation.

the WSC features, the best-performing model retained 21 features, while MFCCs and eGeMAPS models retained only 3 and 19 features, respectively. The larger number of relevant features for WSCs suggests that they encode complementary aspects of the voice that together characterize strain in voice, potentially requiring their non-linear combination for effective class separation (as learned by the SVM and MLP classifiers during training). Although the MFCC model achieved strong results with just three features, they likely fail to capture some of the nuances in voice encoded by the scattering coefficients.

Further insights into the acoustic characteristics of strained voices emerge from the feature distributions shown in Figure 3. MFCC1 plays a key role in identifying vocal strain levels, as shown in Figure 3(b)-(c), where lower MFCC1 values are associated with *moderate-severe* strain. Since MFCC1 reflects spectral tilt, or the balance between low and high-frequency energy, lower values suggest that strained voices have more high-frequency energy. This is in line with the results in,[55] where perception of synthetic stimuli was studied. An increase in the spectral center of gravity and an increase in the amplitude of higher harmonics and spectral slope have been related to the perception of strain.[56] A relatively strong energy concentration in the high-frequency range, as such, is a necessity for genre-typical voice timbre and well-projecting voice both in classical opera voice and in many CCM styles, particularly in the so called belting which was represented among the samples of the present study. In operatic voice, the energy concentration is found in the 2-4 kHz range (singer's formant cluster), depending on sex and voice type,[57,58] and in the 1-1.7 kHz range in belting.[14] Further studies should aim to develop genre-specific thresholds for excessive strain, which singers could use to adjust their technique, aiming for a more balanced energy distribution and reducing the risk of excessive vocal loading.

Additionally, in voices with moderate-severe strain, the spectral envelope often appears flatter, with less pronounced peaks and valleys. This results in lower MFCC2 values, as shown in Figure 3(a). The reduced curvature in the spectral envelope may indicate not only a gentle spectral slope but also formants which are of higher energy and located more close to each other. This in turn may be

related to the raised larynx and narrowed pharynx.[59] Higher F1 and F2 frequencies have been found to be related to the perception of pressedness (strain).[60] While a high laryngeal position and pharyngeal constriction is often seen in patients with muscle tension dysphonia,[61,62] these characteristics are also voluntarily used in belting.[63,64] Further studies should investigate closer the differences between dysphonic strain and strain-resembling characteristics exploited in artistic voice use.

Among the selected MFCC features listed in Tables 9 and 10, MFCC0_var, MFCC4_mean, and MFCC6_mean were also highly ranked. MFCC0 represents the log energy of the signal, indicating how strong or weak the sound is. By checking the feature distributions, we found that strained voices likely have a larger MFCC0_mean but a smaller MFCC0_var (variance), which indicates that singers (especially CCM singers) who sounded more strained tended to produce more consistent energy and have reduced intensity variation. In general, the higher-order MFCCs are not easy to interpret in terms of physical meaning. Like the selected MFCC4 and MFCC6 of this study, they capture finer details in the spectral envelope and variations, which may reflect subtle irregularities in vocal fold vibration and phonatory stability under strain. Together, these patterns underline how subtle spectral features can effectively differentiate strain levels when paired with robust classification methods.

Among the 19 selected eGeMAPS features, eight are related to F0 and the first three formants (F1-F3). As such F0 and its variations reflect laryngeal muscle tension and less phonatory stability, which can be indicators of vocal strain. Synthesized samples with higher pitch received higher ratings of pressedness (strain) in.[65] Samples with higher pitch sound brighter[66] which relates higher pitch with perception of strain due to stronger high-frequency harmonics. Furthermore, higher pitch requires larger subglottic pressure, which has been found to correlate with the perception of pressedness.[60] The amplitude, frequency, and bandwidth of formants can indicate altered vocal tract configurations and vocal fold vibrations due to muscle tension. The other selected features include four MFCCs, three loudness or sound level features, two spectral flux measurements, and one slope measure of the voiced segment. Spectral flux measures the rate of change in the spectrum of a voice signal over time, where a higher value indicates a more prominent change in the spectrum, often correlating with more dynamic or rapidly changing sounds. The slope feature measures the change in energy in the frequency range of 0-500 Hz between two adjacent time frames.

Among the various feature sets tested, the selected 21 WSCs in Table 12 demonstrated the most effective performance in classifying strain levels. However, establishing a clear and direct relationship between each individual WSC feature and the presence of vocal strain is challenging, requiring further investigation.

In terms of classifiers, MLP consistently outperformed SVM across all feature sets, highlighting its ability to model complex, nonlinear relationships in high-dimensional feature spaces, particularly for the WSC features. In all cases, the application of the RFE-based feature selection improved classification performance from using the original higher-dimensional feature sets with both classifiers. The system based on the WSC features and the MLP classifier achieved the highest accuracy (86.1%), highlighting the ability of this combination to provide a comprehensive representation of vocal strain in singing voices.

The findings discussed above emphasize the importance of selecting complementary feature sets and classifiers. The success show by the system, which combined the WSC features with the MLP classifier, suggests the promising potential for further research into vocal health monitoring and singing analysis.

While the current study offers promising results, several limitations should be acknowledged. The dataset was relatively small, with 324 samples, 15 singers, and two listeners, which limits generalizability. However, the data covered diverse vowels, pitches, intensities, and genres, providing a broad range of valuable acoustic variation. Additionally, the listeners brought a mix of vocological expertise and background knowledge of the two genres studied. The inter- and intra-rater reliabilities in the evaluation were satisfactory (Spearman's rho 0.60, $P < 0.001$ and rho 0.79, $P < 0.001$, respectively), which is noteworthy result in perceptual voice analysis,[67] especially in the evaluation of the midrange and in the perception of strain.[68,69] Given these factors, the findings from this study are promising.

Since the acoustic voice data used in this work were recorded while wearing an airflow mask, the classifier models trained on the current data may not generalize well to voice samples recorded without using a flow mask. However, the presence of mask in the recordings does not compromise the general findings presented in this study, as the effects of the mask are the same for all compared samples, and hence cannot result in higher-than-normal classification accuracy or differences in features across strain levels. In any case, further research is required to replicate the current experiment with singing samples recorded in the free field without using a flow mask.

Further studies should also aim to include a larger number of singers, a greater variety of perceptual ratings from expert listeners, and a more extensive dataset. One promising direction involves incorporating biodata, such as EGG signals and self-assessment of vocal ease and quality, as done in.[70,71] The current data originates from a project that also captured EGG and aerodynamic variables (eg, subglottic air pressure and airflow), along with participants' self-assessments, which will allow for expanding the sample size and improving data diversity for future machine learning applications. A vision for the future is that it would be possible to relate acoustic signal—as the easiest signal to collect—closely to the measurement of impact stress. This would allow automatic detection of vocal loading, thereby avoiding the inaccuracies and individual differences in the subjective perceptual voice assessment, as

caused by, eg, different voice timbre ideals in different song genres or differences between singing schools and individual teachers.

CONCLUSION

This study explored the automatic classification of vocal strain levels in singing voices using machine learning approaches. The analysis demonstrated that WSCs, combined with a MLPs classifier, achieved the highest accuracy (86.1%) in distinguishing *normal-mild* from *moderate-severe* strain levels. MFCCs and eGeMAPS also showed strong performance (84.6% each) when paired with MLP and feature selection. This highlights the potential of both traditional and advanced acoustic representations for vocal strain classification.

The findings also demonstrate the importance of selecting complementary feature-classifier combinations, with the WSC features and the MLP classifier emerging as particularly effective for capturing complex vocal patterns. At the same time, the strong performance of MFCCs with a small feature subset emphasizes the value of efficient feature sets in achieving competitive results.

These results lay a foundation for further research into automated vocal health monitoring. Expanding the dataset, refining annotation practices, and exploring multimodal approaches that integrate acoustic, physiological, and self-assessment data will be key to advancing the field and enabling practical applications of these methods in singing voice analysis.

Declaration of Competing Interests

The authors declare no conflicts of interest. This work is conducted solely for academic and research purposes and does not intend to compete with or undermine other works in the field.

References

1. Zhang Z. Mechanics of human voice production and control. *J Acoust Soc Am.* 2016;140:2614–2635. https://doi.org/10.1121/1.4964509. ([Online]. Available).
2. Nemr K, Simoes-Zenari M, Cordeiro GF, et al. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice.* 2012;26:812.e17–812.e22.
3. Hirano M. Clinical examination of voice. In: Arnold BWGE, Winckel F, eds. *Disorders of Human Communication 5.* Wien: Springer; 1981:81–84.
4. Hillman RE, Holmberg EB, Perkell JS, Walsh M, Vaughan C. Objective assessment of vocal hyperfunction: an experimental framework and initial results. *J Speech Lang Hear Res.* 1989;32:373–392.
5. Mehta DD, Van Stan JH, Zañartu M, et al. Using ambulatory voice monitoring to investigate common voice disorders: research update. *Front Bioeng Biotechnol.* 2015;3:1–14.
6. Groll MD, Hablani S, Stepp CE. The relationship between voice onset time and increase in vocal effort and fundamental frequency. *J Speech Lang Hear Res.* 2021;64:1197–1209.
7. Kempster GB, Gerratt BR, Abbott KV, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am Speech-Lang-Hear Assoc.* 2009;18:124–132.
8. Baldner EF, Doll E, Mersbergen MRVan. A review of measures of vocal effort with a preliminary study on the establishment of a vocal effort measure. *J Voice.* 2015;29:530–541.
9. Fuchs M, Meuret S, Geister D, et al. Empirical criteria for establishing a classification of singing activity in children and adolescents. *J Voice.* 2008;22:649–657.
10. Mezzedimi C, Spinosi MC, Massaro T, Ferretti F, Cambi J. Singing voice: acoustic parameters after vocal warm-up and cool-down. *Logop Phoniatr Vocol.* 2020;45:57–65.
11. Titze IR. Mechanical stress in phonation. *J Voice.* 1994;8:99–105.
12. Jiang JJ, Titze IR. Measurement of vocal fold intraglottal pressure and impact stress. *J Voice.* 1994;8:132–144.
13. Thalen M, Sundberg J. Describing different styles of singing: a comparison of a female singer's voice source in 'classical,' 'pop,' 'jazz,' and 'blues. *Logop Phoniatr Vocol.* 2001;26:82–93.
14. Sundberg J, Thalén M, Popeil L. Substyles of belting: phonatory and resonatory characteristics. *J Voice.* 2012;26:44–50.
15. Sonninen A, Damste P, Jol J, Fokkens J. On vocal strain. *Folia Phoniatr Logop.* 1972;24:321–336.
16. Sonninen A, Hurme P, Vilkman E. Roentgenological observations on vocal fold length-changes with special reference to register transition and open/covered voice. *Scand J Logop Phoniatr.* 1992;17:95–106.
17. García-López I, GavilánBouzas J. The singing voice. *Acta Otorrinolaringol (English Edition).* 2010;61:441–451.
18. Sundberg J, Thalén M. Respiratory and acoustical differences between belt and neutral style of singing. *J Voice.* 2015;29:418–425.
19. Sundberg J, Thalén M, Alku P, Vilkman E. Estimating perceived phonatory pressedness in singing from flow glottograms. *J Voice.* 2004;18:56–62.
20. Sundberg J, Gramming P, Lovetri J. Comparisons of pharynx, source, formant, and pressure characteristics in operatic and musical theatre singing. *J Voice.* 1993;7:301–310.
21. Stoney J. How all singers should think about belting. *Voice Council Magazine.* 2016.
22. Anand S, Kopf LM, Shrivastav R, Eddins DA. Objective indices of perceived vocal strain. *J Voice.* 2019;33:838–845.
23. Latoszek BBV, Maryn Y, Gerrits E, Bodt MDe. A meta-analysis: acoustic measurement of roughness and breathiness. *J Speech Lang Hear Res.* 2018;61:298–323.
24. Barsties B, Latoszek V, Maryn Y, Gerrits E, De Bodt M. The acoustic breathiness index (ABI): a multivariate acoustic model for breathiness. *J Voice.* 2017;31:511.e11–511.e27.
25. McKenna VS, Stepp CE. The relationship between acoustical and perceptual measures of vocal effort. *J Acoust Soc Am.* 2018;144:1643–1658.
26. Lowell SY, Kelley RT, Awan SN, Colton RH, Chan NH. Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality. *Ann Otol Rhinol Laryngol.* 2012;121:539–548.
27. Wang Z, Yu P, Yan N, Wang L, Ng ML. Automatic assessment of pathological voice quality using multi-dimensional acoustic analysis based on the GRBAS scale. *J Signal Process Syst.* 2016;82:241–251.
28. XieS, Yan N, YuP, *et al.* Deep neural networks for voice quality assessment based on the GRBAS scale, In the 17th Annual Conference of the International Speech Communication Association, 2016, 2656–2660.
29. Nguyen DD, Madill C. Auditory-perceptual parameters as predictors of voice acoustic measures. *J Voice.* 2023:1–11.
30. Fujimura S, Kojima T, Okanoue Y, et al. Classification of voice disorders using a one-dimensional convolutional neural network. *J Voice.* 2022;36:15–20.
31. Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: theory and practice. *Int J Comput Vision.* 2013;105:222–245.
32. Tiwari V. Mfcc and its applications in speaker recognition. *Int J Emerg Technol.* 2010;1:19–22.
33. McFee B, Raffel C, Liang D, *et al* Librosa: Audio and music signal analysis in python, In Procedings 14th Python in Science Conference, 2015, 18–24.

ARTICLE IN PRESS

34. Tirronen S, Kadiri SR, Alku P. The effect of the MFCC frame length in automatic voice pathology detection. *J Voice.* 2024;38:975–982.
35. Eyben F, Scherer KR, Schuller BW, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput.* 2016;7:190–202.
36. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor, In Procedings of the 21st ACM International Conference on Multimedia, 2013, 835–838.
37. Bruna J, Mallat S. Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell.* 2013;35:1872–1886.
38. Joakim A, Stéphane M. Deep scattering spectrum. *IEEE Trans Signal Process.* 2014;62:4114–4128.
39. Yagnavajjula MK, Mittapalle KR, Alku P, Rao KS, Mitra P. Automatic classification of neurological voice disorders using wavelet scattering features. *Speech Commun.* 2024;157:1–10.
40. Reddy MK, Keerthana YM, Alku P. End-to-end pathological speech detection using wavelet scattering network. *IEEE Signal Processing Letts.* 2022;29:1863–1867.
41. Mittapalle KR, Alku P. Classification of phonation types in singing voice using wavelet scattering network-based features. *JASA Express Letts.* 2024;4:1–7.
42. Oneata D, Verbeek J, Schmid C. Action and event recognition with fisher vectors on a compact feature set. In: Proceedings of the IEEE International Conference on Computer Vision, 2013, 1817–1824.
43. Gosztolya G. Using the fisher vector representation for audio-based emotion recognition. *Acta Polytech Hung.* 2020;17:7–23.
44. Tian Y, He L, Li Z-Y, Wu W-l , Zhang W-Q, Liu J. Speaker verification using fisher vector, In: The 9th International Symposium on Chinese Spoken Language Processing, 2014, 419–422.
45. Hegde S, Achary KK, Shetty S. Feature selection using fisher's ratio technique for automatic speech recognition, 2015 [Online]. Available at: https://arxiv.org/abs/1505.03239.
46. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* 2002;46:389–422.
47. He J, Ding L, Jiang L, Ma L. Kernel ridge regression classification, In: 2014 International Joint Conference on Neural Networks, 2014, 2263–2267.
48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.
49. Egas-López J, Orozco JR, Gosztolya G. Assessing Parkinson's disease from speech using fisher vectors, In: The 20th Annual Conference of the International Speech Communication Association, 2019, 3063–3067.
50. Abakarim F, Abenaou A. Voice pathology detection using the adaptive orthogonal transform method, svm and mlp. *Int J Online Biomed Eng.* 2021;17:90–102.
51. Barizão AH, Fermino MA , Dajer ME, Liboni LH, Spatti DH. Voice disorder classification using mlp and wavelet packet transform, In: 2018 International Joint Conference on Neural Networks, 2018, 1–8.
52. Statistics L. Mann-Whitney *U* Test using SPSS Statistics, 2018. [Online]. Available at https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php.
53. Andén J, Mallat S. Deep scattering spectrum. *IEEE Trans Sign Process.* 2014;62:4114–4128.
54. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. *IEEE Access.* 2022;10:122136–122158.
55. Stone TC, Erickson ML. Experienced and inexperienced listeners' perception of vocal strain. *J Voice.* 2024. Online ahead of print.
56. Anand S, Kopf LM, Shrivastav R, Eddins DA. Objective indices of perceived vocal strain. *J Voice.* 2019;33:838–845.
57. Sundberg J. Level and center frequency of the singer's formant. *J Voice.* 2001;15:176–186.
58. Echternach M, Burk F, Kirsch J, et al. Articulatory and acoustic differences between lyric and dramatic singing in western classical music. *J Acoust Soc Am.* 2024;155:2659–2669.
59. Sundberg J, Nordström P-E. Raised and lowered larynx-the effect on vowel formant frequencies. *STL-QPSR.* 1976;17:035–039.
60. Millgård M, Fors T, Sundberg J. Flow glottogram characteristics and perceived degree of phonatory pressedness. *J Voice.* 2016;30:287–292.
61. Morrison MD, Rammage LA, Belisle GM, Pullan CB, Nichol H. Muscular tension dysphonia. *J Otolaryngol.* 1983;12:302–306.
62. Lowell SY, Kelley RT, Colton RH, Smith PB, Portnoy JE. Position of the hyoid and larynx in people with muscle tension dysphonia. *Laryngoscope.* 2012;122:370–377.
63. Cleveland TF, Sundberg PJ, Prokop J, et al. Aerodynamic and acoustical measures of speech, operatic, and broadway vocal styles in a professional female singer. *J Voice.* 2003;17:283–297.
64. Saldias M, Guzman M, Miranda G, Laukkanen A-M. A computerized tomography study of vocal tract setting in hyperfunctional dysphonia and in belting. *J Voice.* 2019;33:412–419.
65. Bergan CC, Titze IR, Story B. The perception of two vocal qualities in a synthesized vocal utterance: ring and pressed voice. *J Voice.* 2004;18:305–317.
66. Collier WG, Hubbard TL. Judgments of happiness, brightness, speed and tempo change of auditory stimuli varying in pitch and tempo. *Psychomusicol J Res Music Cogn.* 1998;17:36.
67. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech LangHear Res.* 1993;36:21–40.
68. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am.* 1998;104:1598–1608.
69. Webb A, Carding P, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Oto-Rhino-Laryngol Head Neck.* 2004;261:429–434.
70. Donati E, Chousidis C, Ribeiro HDM, Russo N. Classification of speaking and singing voices using bioimpedance measurements and deep learning. *J Voice.* 2023:1–8.
71. Jones SP, Reni SK, Kale I. Machine learning for monitoring vocal health and performance of professional singers, In: 2024 IEEE International Symposium on Circuits and Systems, 2024, 1–5.