Artificial Intelligence to Detect Voice Disorders: An AI-Supported Systematic Review of Accuracy Outcomes

*Charles J. Nudelman, †Virginia Tardini, and ‡Pasquale Bottalico, *Syracuse, NY, †Bologna, Italy, and ‡Urbana, IL

SUMMARY: Background. The objective of the present systematic review is to identify which artificial intelligence (AI) approaches have been used to successfully detect voice disorders. The review examines studies involving patients with non-neurological voice disorders and controls, where AI was applied to detect voice disorders. The primary outcome of interest is the accuracy of these AI models. Additionally, this review demonstrates how the procedures of conducting a systematic review can be supported by AI.

Methods. Studies were eligible for inclusion if they implemented an AI approach to detect non-neurological voice disorders from healthy voice samples. A comprehensive search was conducted using PubMed/MEDLINE, Science Direct, Web of Science, EBSCO, and Scopus databases. Risk of bias was assessed via the Quality Assessment Tool for Diagnostic Accuracy Studies. The occurrences of the most common AI techniques utilized in the literature are presented, and a summary of their abilities to accurately detect a voice disorder is reported.

Results. In total, 79 publications met the inclusion criteria. These studies included patient recordings from a variety of voice databases. The most common AI techniques implemented were Support Vector Machines (SVMs) (n = 28) and Convolutional Neural Networks (CNNs) (n = 22). The mean accuracy of the models in detecting voice disorders was 92% across all studies. Nine studies reported 100% accuracy, and 32 studies reported between 95% and 99%.

Discussion. Strengths of the evidence include high accuracies across diverse models and datasets. Limitations include a limited variety of datasets and a trend of hyperoptimization without sufficient external validation. Clinicians and researchers should recognize that while current AI models show promise, future studies should prioritize robust external validation and more representative datasets.

Key Words: Artificial intelligence–AI–Machine learning–Systematic review–Voice disorders.

INTRODUCTION

Artificial Intelligence (AI) is emerging as a transformative tool for medicine, shifting the field from human observation toward machine-based precision medicine.[1,2] Machine learning can be considered a form of AI and, simultaneously, a driver of AI.[3] Machine learning generally involves a system that can acquire its own knowledge through supervised or unsupervised agents.[4] These machine-based tools can make decisions about—and predictions from—medically relevant patient data.[5] AI has demonstrated efficacy in analyzing acoustic voice signals to detect a dysphonic voice from a healthy voice, which may aid in the early identification of voice disorders and facilitate improved access to specialized care. Various machine learning approaches have been used to detect dysphonic voices from healthy voices. These machine learning approaches commonly involve a training phase and testing phase, which together aim to improve the quality of the AI system. The training phase may be supervised (ie, the machine learning algorithms are trained on a dataset that contains information about the problem at hand[6]) or unsupervised (ie, the machine learning algorithm uses a dataset that is unlabeled and reveals hidden structures within the dataset[5]). Following the training phase, the AI model is tested with test data (novel to the machine), and researchers then evaluate the performance of the tool on the test data.

Types of AI tools for voice disorder detection

There are various types of machine learning tools, which can be generally distinguished as tools that are based on statistical learning and neural network algorithms. Among those based on statistical learning are K-nearest neighbors (KNN), Hidden Markov Modeling (HMM), online sequential extreme learning machine (OSELM), Support Vector Machines (SVMs), and Extreme Gradient Boosting (XGBoost). Neural network algorithms include Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Deep Neural Networks (DNNs). An overview of these machine learning tools can be found in Appendix A.

Voice disorder databases

Within the extant literature, AI tools are viewed increasingly as valuable contributors to improved precision, accuracy, and reliability in detecting voice disorders.[7] Presently, various voice recording databases are utilized by researchers when testing machine learning techniques to detect voice disorders. The Saarbruecken Voice Database (SVD) contains recordings of over 2000 speakers (687 healthy speakers and 1356 patients with 71 distinct voice

Accepted for publication September 12, 2025.

From the *Department of Communication Sciences and Disorders, Syracuse University, Syracuse, NY; †Department of Industrial Engineering, University of Bologna, Bologna, Italy; and the ‡Department of Speech and Hearing Science, University of Illinois Urbana-Champaign, Urbana, IL.

Address correspondence and reprint requests to: Charles J. Nudelman, Department of Communication Sciences and Disorders, Syracuse University, Syracuse, NY. E-mail: cjnudelm@syr.edu

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

https://doi.org/10.1016/j.jvoice.2025.09.021

disorders). Tasks include sustained vowels at standard, high, and low pitches, pitch glides on sustained vowels, and a short sentence spoken in the German language.[8] The Advanced Voice Function Assessment Databases (AVFAD) recorded 709 speakers (363 healthy voice users and 346 speakers with voice disorders, including nodules, polyps, cysts, Reinke's edema, reflux, and unilateral vocal fold paralysis). These speakers recorded sustained vowels, sentences, a read text, and spontaneous speech in the Portuguese language.[9] The Massachusetts Eye and Ear Infirmary (MEEI) database has over 1400 voice recordings (657 samples from speakers with voice disorders and 53 samples from healthy voice users). Tasks in the MEEI database include a sustained vowel and a read text spoken in the English language.[10] The Hospital Universitario Principe de Asturias (HUPA) database consists of 408 total sustained vowels from 239 healthy voice users and 169 patients with voice disorders, including vocal fold nodules, polyps, cysts, and edema.[11-13] The Arabic Voice Pathology Database (AVPD) is another corpus, which involves 366 speakers (187 healthy voice users and 179 patients with vocal fold sulcus, nodules, cysts, paralysis, or polyps) who recorded three vowels, isolated Arabic words, and running speech.[14] The Far Eastern Memorial Hospital voice database (FEMH) contains sustained vowel recordings from 250 speakers, including 50 healthy voice users and 150 patients with vocal fold nodules, polyps, cysts, laryngeal neoplasm, or unilateral vocal fold paralysis.[15] The VOice ICar fEDerico II Database (VOICED) includes 208 sustained vowel recordings (58 healthy voice users and 150 patients). The disorders ascribed to the patients in the VOICED dataset are as follows: prolapse, polyps, hyperkinetic dysphonia, rigid vocal folds, chorditis, Reinke's edema, vocal fold nodules, minor hyperkinetic dysphonia, extraglottic air leak, spasmodic dysphonia, cyst, bilateral vocal fold, laryngitis, conversion dysphonia, vocal fold paralysis, minor hypokinetic dysphonia, glottic insufficiency, presbyphonia, adduction deficit, dysphonia by chordal groove, hypokinetic dysphonia, or laryngopharyngeal reflux.[16] Of note, many of these disorders are not considered to be standard diagnoses (eg, bilateral vocal fold). However, it is possible that this list contains errors due to linguistic translation from the patients' medical records. The MAPACI speech pathology database contains a total of 48 sustained vowel recordings from 24 vocally healthy speakers and 24 patients with voice disorders.[17] Table 1 summarizes these databases.

There are various voice acoustic features that are commonly assessed by AI tools when detecting voice disorders. Prior to feature extraction, a preprocessing step typically occurs, which may involve low-pass filtering of the signal and windowing to segment the speech signal appropriately.[18] Acoustic voice features may be time-domain features (eg, speech segment energy, zero-crossing rate, and short-time energy), perceptual features (eg, pitch, harmonicity, chroma, spectral dispersion, spectral centroid, spectral skewness, and entropy), or physical features (eg,

TABLE 1.
Characteristics of Voice Recording Databases

Database name	Samples (*N*)	Speakers	Voice disorder	Languages	Tasks
Saarbruecken Voice Database (SVD)	2043	687 healthy, 1356 patients	71 distinct voice disorders	German	Sustained vowels, pitch glides, and sentences
Advanced Voice Function Assessment Databases (AVFAD)	709	363 healthy, 346 disordered	Nodules, polyps, cysts, Reinke's edema, reflux, and unilateral vocal fold paralysis	Portuguese	Sustained vowels, sentences, read text, and spontaneous speech
Massachusetts Eye and Ear Infirmary (MEEI)	710	53 healthy, 657 disordered	Various unspecified disorders	English	Sustained vowels, read text
Hospital Universitario Principe de Asturias (HUPA)	408	239 healthy, 169 disordered	Vocal fold nodules, polyps, cysts, and edema	Spanish	Sustained vowels only
Arabic Voice Pathology Database (AVPD)	366	187 healthy, 179 disordered	Vocal fold sulcus, nodules, cysts, paralysis, and polyps	Arabic	Vowels, isolated words, and running speech
Far Eastern Memorial Hospital (FEMH)	200	50 healthy, 150 disordered	Vocal fold nodules, polyps, cysts, laryngeal neoplasm, and unilateral vocal fold paralysis	Mandarin	Sustained vowels only
VOice ICar fEDerico II Database (VOICED)	208	58 healthy, 150 disordered	Extensive list including prolapse, polyps, hyperkinetic dysphonia, etc	Italian	Sustained vowels only
MAPACI speech pathology database	48	24 healthy, 24 disordered	Unspecified	Spanish	Sustained vowels only

spectral slope, group phase delay). Mel-Frequency Cepstral Coefficient feature vectors are also commonly used, and they represent the sound power spectra of a voice in the cepstral domain.[19]

Purpose of systematic review

Due to the variety of AI approaches used to detect voice disorders from healthy voices, this systematic review was developed to synthesize the existing scientific evidence describing the accuracy of these AI tools in detecting primary (ie, non-neurologically-based) voice disorders. The overarching aims of the current systematic review are:

- To synthesize the evidence on AI-based methods for detecting voice disorders.
- To evaluate methodological quality and accuracy performance across studies.
- To provide recommendations for future research in this area.

METHOD

A comprehensive literature search, study selection, data extraction, and assessment of methodological quality were performed following the PRISMA guidelines.[20]

Literature search

This systematic review of literature was performed using five computerized databases to characterize the accuracy of AI-based voice methods in detecting primary voice disorders. The databases were PubMed/MEDLINE (National Library of Medicine, Bethesda, MD), Science Direct (Elsevier, Amsterdam, Netherlands), Web of Science (Clarivate Analytics PLC, Philadelphia, PA), EBSCO (EBSCO Industries, Birmingham, AL), and Scopus (Elsevier, Amsterdam, Netherlands). No year limits were applied to the databases when conducting the literature search.

The search string used was: “(("artificial intelligence" OR "machine learning" OR "deep learning" OR "learning algorithms" OR "machine learning techniques") OR ("neural network" OR "neural networks" OR "convolutional neural network" OR "convolutional neural" OR "support vector machine" OR "svm" OR "vector machine" OR "support vector" OR "classification" OR "feature selection")) AND (("voice" OR "voice disorder" OR "voice disorders" OR "pathological voice" OR "voice pathology" OR "dysphonia") AND ("speech" OR "voice signal" OR "voice samples" OR "cepstral coefficient")) AND (("diagnosis" OR "pathology detection") OR ("sensitivity" AND "specificity") OR ("frequency cepstral coefficients" OR "cepstral coefficients" OR "frequency cepstral")) OR ("mfcc" OR "mel-frequency cepstral coefficients") AND ("voice database" OR "voice samples")”. This search string, developed using litsearchr,[21] included a combination of controlled vocabulary (MeSH terms) and non-MeSH free-text terms. The litsearchr AI approach was trained on a manual naive search. The automated method used text mining and keyword co-occurrence networks to identify the most important terms for a literature review and was implemented in the R (R Development Core Team, 2020) package litsearchr.[21] This automated approach has been shown to reduce bias in search strategy development and has been used by the first author in a prior systematic review.[22,23] This search string was inputted into the databases on December 12, 2024. Of note, research librarians were not involved in this search. The databases and number of articles resulting from the searches are displayed in Table 2.

TABLE 2.
The Databases and Number of Articles Resulting From the Searches

Database	Number of articles resulting from search (*n*)
PubMed/MEDLINE	83
Science Direct	1166
Web of Science	1254
EBSCO	1141
Scopus	297

Study selection

The search string resulted in 3941 potentially relevant publications. Prior to screening, 934 duplicate articles were removed. Thus, following duplicate removal, the total number of potentially relevant publications was 3007.

In the first stage of screening, seven different humans and one AI large language model (LLM) reviewed the titles and abstracts of each potential publication. The LLM was trained via a custom Python script that called an OpenAI LLM to review the tiles and abstract using the same criteria as the human reviewers. A copy of the Python prompt can be found in Appendix B. The first author of this study (reviewer 1, CN) and the LLM reviewed all 3007 titles and abstracts. For confidence, the other seven human reviewers split the 3007 publications’ titles and abstracts to review among themselves. Cohen’s kappa with 95% confidence intervals was calculated comparing reviewer 1 and the other reviewers individually using R version 2022.01.2 (R Development Core Team, 2022). These calculations were based on single-ratings of inclusion or exclusion. The Cohen’s kappa estimates for the title and abstract review are displayed in Table 3. During this first screening, 2806 publications were excluded.

The total remaining 201 publications were included for full paper review. For the full paper review, three inclusion criteria were defined: (1) a machine learning approach was implemented, (2) the machine learning approach detected non-neurological voice disorders from healthy sample, and (3) articles must not be systematic or scoping reviews of literature. Only those publications accessible to the authors and published in peer-reviewed scientific journals written in English were included. Prior to the full paper review, a single additional article was excluded due to being

TABLE 3.
Cohen's Kappa Estimates and 95% Confidence Intervals Calculated Comparing Reviewer 1 and the Other Seven Reviewers, Individually, for the Title and Abstract Screening Process

		95% CI	
Rater	Cohen's Kappa Estimate	Lower Bound	Upper Bound
R1-R2	0.369	0.498	0.841
R1-R3	0.602	0.534	0.805
R1-R4	0.669	0.568	0.770
R1-R5	0.398	0.502	0.837
R1-R6	0.653	0.576	0.763
R1-R7	0.476	0.529	0.812
R1-AI	0.523	0.476	0.571

≤0 = indicates **no** agreement.
0.01-0.20 = **none to slight** agreement.
0.21-0.40 = **fair** agreement.
0.41- 0.60 = **moderate** agreement.
0.61-0.80 = **substantial** agreement.
0.81-1.00 = **almost perfect** agreement.

a duplicate. The remaining 200 articles underwent a full paper review according to the three inclusion criteria. A total of 12 publications were removed due to lacking a machine learning approach, 21 were removed due to lacking detection between disordered and healthy voice, 19 were removed for including neurologically-based voice disorders, seven were removed for being a systematic or scoping review, 60 were removed for being conference proceedings, and two were removed for not being written in English. A total of 79 publications met the inclusion criteria and, therefore, were included in the systematic review.

Figure 1 shows the flowchart of the literature search. All included studies were used for data extraction and methodological quality assessment.

Data extraction and analysis

Data extraction and analysis were achieved through two phases. First, relevant data were extracted from the included publications. The extracted information included year of publication, study population, sample size, and the machine learning/AI technique that achieved the best (highest) accuracy in detecting voice disorders from healthy voices. An overview of the characteristics of the included publications is presented in the Results section in Table 4. Second, quality assessment analysis was performed using the Quality Assessment of Diagnostic Accuracy Studies—Second Edition (QUADAS-2[24]).

Assessment of methodological quality

The first author and the two reviewers with the highest Cohen's kappa values from the title and abstract screening read all the included publications and assessed for methodological quality using the Quality Assessment of Diagnostic Accuracy Studies—Second Edition (QUADAS-2[24]). Any initial disagreement regarding any rating of the included studies was resolved in a consensus meeting. This tool considers seven main components: selection bias, index test bias, reference standard bias, patient flow bias, applicability of included patients, applicability of index test, and applicability of reference standard. For the quality score, each of the seven components was rated on a scale of unclear, low, and high. Ratings are informed by "signaling questions" specific to each domain, which flag aspects of study design related to the potential for bias and guide reviewers toward consistent judgments. These questions are answered as "yes," "no," or "unclear". If all signaling questions for a domain are answered "yes," the risk of bias is judged low; if any are answered "no," potential for bias exists and the rating is adjusted accordingly. The "unclear" category is used when insufficient data are reported to make a determination. For the overall quality score, each of the seven components was rated on a scale of unclear, low, and high.

RESULTS

The 79 articles presented were extracted primarily from the following journals: *Journal of Voice* ($n = 9$); *IEEE Access* ($n = 7$); *Applied Sciences* ($n = 5$); *Computers in Biology and Medicine* ($n = 4$); *Biomedical Signal Processing and Control* ($n = 3$); *IEEE Transactions on Biomedical Engineering* ($n = 3$); *Speech Communication* ($n = 3$); *Scientific Reports* ($n = 2$); *Computers & Electrical Engineering* ($n = 2$); *Healthcare Analytics* ($n = 2$); *International Journal of Healthcare Information Systems and Informatics* ($n = 2$); *International Journal of Systems Science* ($n = 2$). The 79 articles were published between 2004 and 2024. An overview of the characteristics of the included publications is presented in Table 4.

AI approaches

The included studies employed a variety of AI techniques. The most common machine learning techniques implemented were SVMs ($n = 28$) and CNNs ($n = 22$). Additional techniques included KNNs ($n = 4$), LSTMs ($n = 4$), ANNs ($n = 3$), DNNs ($n = 3$), OSELMs ($n = 2$), XGBoost ($n = 2$), Convolutional Bottleneck Network ($n = 1$), Discriminative Paraconsistent Machine ($n = 1$), Feedforward Neural Network ($n = 1$), Generative Adversarial Network ($n = 1$), Hierarchical Extreme Learning Machine ($n = 1$), HMM ($n = 1$), Linear Discriminant Analysis ($n = 1$), Logistic Model Tree algorithm ($n = 1$), Logistic Regression Model ($n = 1$), Learning Vector Quantization ($n = 1$), Naive Bayes classifier ($n = 1$), Quadratic Discriminant Analysis ($n = 1$), Sequential Learning Resource Allocation Neural Network ($n = 1$), and Stochastic Gradient Descent Classifier ($n = 1$).

Seven of the 79 articles presented distinct novel approaches toward AI identification of voice disorders. Four of these seven articles presented novel approaches with SVMs. Amami & Smiti, (2017) developed a novel policy combining a modified density-based clustering algorithm

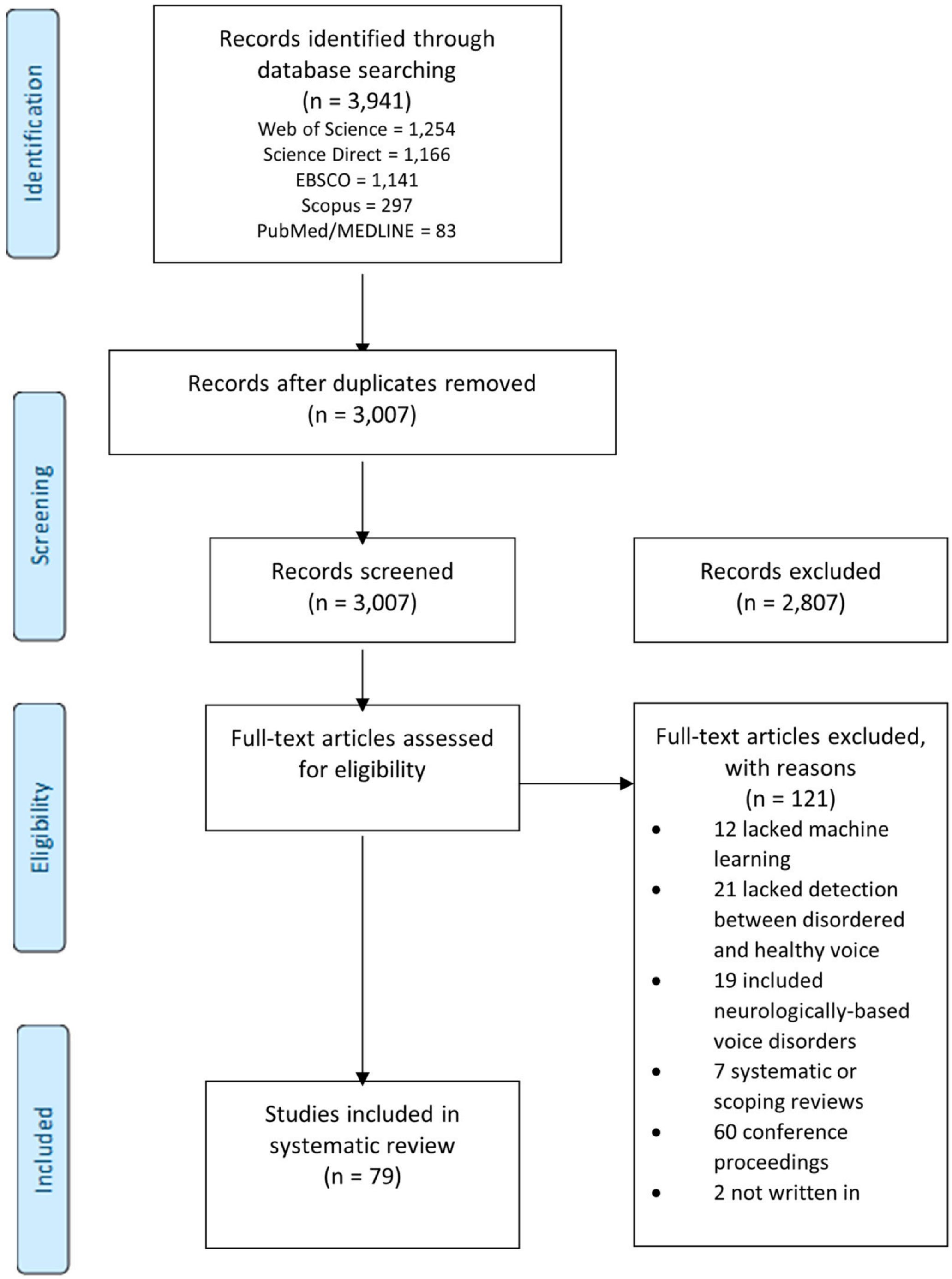

FIGURE 1. Flowchart of the process for identification of included publications.

and SVMs. Ultimately, their novel model demonstrated 98% accuracy in detecting voice disorders using the MEEI dataset.[32] Similarly, Zakariah et al (2024) developed an integrated attention-based decision-making approach with SVMs. To aid in detecting voice disorders, Zakariah et al also introduced Mel-Frequency Energy Line features, which encompass spectral qualities of dysphonia. Ultimately, their SVM integration (called SVM-TabNet) had 100% accuracy in detecting voice disorders within the SVD database.[100] Uloza et al (2010) developed a novel approach toward building SVM committees. Sequential committees aided in the classification of voice features, as each committee selected progressively more voice features as inputs. Their committee approach demonstrated 92% accuracy in detecting voice disorders within a privately collected dataset with 444 total voice recordings.[89] Basalamah et al (2023) developed a novel preprocessing approach to enhance SVM abilities to detect voice disorders. Specifically, they applied linear discriminant analysis as a preprocessing step, which reduced the dimensionality of the voice feature

TABLE 4.
Characteristics of the Included Publications on the Ability for Machine Learning/AI Techniques to Detect Voice Disorders From Healthy Voices

ID	Reference	Title	Year/Country	Dataset	Machine Learning/AI Technique (Highest Accuracy)
1	24	Voice pathology identification system using a deep learning approach based on unique feature selection sets	2023/Iraq	SVD	LSTM
2	25	Voice Pathology Detection and Classification by Adopting Online Sequential Extreme Learning Machine	2021/Malaysia	SVD	OSELM
3	26	Voice pathology detection using deep learning on mobile healthcare framework	2018/Saudi Arabia	SVD	CNN
4	27	Voice pathology detection based on the modified voice contour and SVM	2016/Saudi Arabia	Privately collected dataset	SVM
5	28	Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions	2017/Saudi Arabia	MEEI, SVD, and AVPD	SVM
6	29	Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions	2017/Saudi Arabia	MEEI, SVD, and AVPD	SVM
7	30	A Novel Voice Feature AVA and its Application to the Pathological Voice Detection Through Machine Learning	2023/Malaysia	SVD	Naive Bayes classifier
8	31	An incremental method combining density clustering and support vector machines for voice pathology detection	2017/Tunis	MEEI	SVM
9	32	Voice pathology detection by using the deep network architecture	2021/Turkey	SVD and privately collected dataset	LSTM and CNN
10	33	On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices	2011/Greece	MEEI and privately collected dataset	SVM
11	11	Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients	2011/Greece	MEEI	SVM
12	34	Identification of voice disorders using long-time features and support vector machine with different feature reduction methods	2011/Iran	MEEI	SVM
13	35	An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis, and support vector machine	2012/Iran	MEEI	SVM
14	36	A Highly Accurate Dysphonia Detection System Using Linear Discriminant Analysis	2023/Saudi Arabia	SVD and privately collected dataset	SVM
15	37	Automatic Classification of Disordered Voices Based on a Hybrid HMM-SVM Model	2021/Algeria	Privately collected dataset	Hybrid HMM and SVM
16	16	Voice Disorder Detection via an m-Health System: Design and Results of a Clinical Study to Evaluate Vox4Health	2018/Italy	Privately collected dataset	Logistic Model Tree algorithm
17	38	Deep Neural Network for Automatic Classification of Pathological Voice Signals	2022/China	VOICED	DNN
18	39	Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K-Nearest Neighbor (KNN)	2021/China	VOICED	KNN
19	40	Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework	2023/China	Privately collected dataset	CNN

TABLE 4 (*Continued*)

ID	Reference	Title	Year/Country	Dataset	Machine Learning/AI Technique (Highest Accuracy)
20	41	Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset	2020/China	VOICED	Generative Adversarial Network
21	42	The use of wavelet-packet transform and artificial neural networks in analysis and classification of dysphonic voices	2007/Brazil	Privately collected dataset	ANN
22	43	Assessment of Voice Disorders Using Machine Learning and Vocal Analysis of Voice Samples Recorded through Smartphones	2024/Italy	VOICED	Fine KNN
23	44	Deep connected attention (DCA) ResNet for robust voice pathology detection and classification	2021/China	SVD and privately collected dataset	CNN
24	45	Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method	2021/China	MEEI and SVD	Random Forest
25	46	Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach	2019/Taiwan	MEEI	DNN
26	47	Voice pathology detection on spontaneous speech data using deep learning models	2024/Iran	AVFAD	CNN
27	48	Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM)	2020/Brazil	SVD	Discriminative Paraconsistent Machine
28	49	Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex	2009/Spain	MEEI	ANN
29	50	Automated speech analysis applied to laryngeal disease categorization	2008/Lithuania	Privately collected dataset	SVM
30	51	Consistency of the Signature of Phonotraumatic Vocal Hyperfunction Across Different Ambulatory Voice Measures	2024/United States	Privately collected dataset	Supervised Logistic Regression Model with Nested Cross-Validation and Forward Feature Selection
31	52	Automatic detection of voice impairments by means of short-term cepstral parameters and neural network-based detectors	2004/Spain	MEEI	Learning Vector Quantization
32	53	Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features	2020/Tunisa	Privately collected dataset	SVM
33	54	A new feature constituting approach to detection of vocal fold pathology	2014/Malaysia	MEEI and MAPACI	KNN
34	55	Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study	2021/Taiwan	Privately collected dataset	CNN
35	56	Using SincNet for Learning Pathological Voice Disorders	2022/Taiwan	Privately collected dataset (FEMH)	CNN
36	57	Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals	2022/Canada	SVD	CNN
37	58	A comparison of data augmentation methods in voice pathology detection	2024/Finland	HUPA and SVD	2-Dimensional (2-D) CNN

TABLE 4 (*Continued*)

ID	Reference	Title	Year/Country	Dataset	Machine Learning/AI Technique (Highest Accuracy)
38	59	Optimized early fusion of handcrafted and deep learning descriptors for voice pathology detection and classification	2024/India	AVPD and SVD	KNN
39	60	Voice pathology detection using optimized convolutional neural networks and explainable artificial intelligence-based analysis	2024/India	AVPD, SVD, and VOICED	CNN
40	61	Analysis and Detection of Pathological Voice Using Glottal Source Features	2020/Finland	HUPA and SVD	SVM
41	62	Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy	2020/Korea	Privately collected dataset	1-Dimensional (1-D) CNN
42	63	Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis	2024/Korea	Privately collected dataset	CNN
43	64	Improved Laryngeal Pathology Detection Based on Bottleneck Convolutional Networks and MFCC	2024/Algeria	HUPA	Convolutional Bottleneck Network
44	65	Deep learning approaches for pathological voice detection using heterogeneous parameters	2020/Korea	MEEI and SVD	Feedforward Neural Network
45	66	An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection	2023/Korea	SVD	CNN
46	67	Evaluating the Diagnostic Potential of Connected Speech for Benign Laryngeal Disease Using Deep Learning Analysis	2024/Korea	Privately collected dataset	CNN
47	68	Different Performances of Machine Learning Models to Classify Dysphonic and Non-Dysphonic Voices	2022/Brazil	Privately collected dataset	Stochastic Gradient Descent Classifier
48	69	Integrated Vocal Deviation Index (IVDI): A Machine Learning Model to Classifier of the General Grade of Vocal Deviation	2024/Brazil	Privately collected dataset	XGBoost
49	70	Artificial Neural Network-based Classification to Screen for Dysphonia Using Psychoacoustic Scaling of Acoustic Voice Features	2008/Germany	Privately collected dataset	ANN
50	71	Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection	2007/ United Kingdom	MEEI	Quadratic Discriminant Analysis
51	14	Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms	2017/Saudi Arabia	AVPD and MEEI	SVM
52	72	Deep Learning Approach for Voice Pathology Detection and Classification	2021/India	VOICED	Ensemble-CNN Decision Fusion
53	73	Classification of functional dysphonia using the tunable Q wavelet transform	2023/Finland	VOICED	CNN
54	74	MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection	2023/Turkey	SVD	LSTM
55	75	Voice pathology detection and classification using convolutional neural network model	2020/Iraq	SVD	CNN
56	76	Telephony-based voice pathology assessment using automated speech analysis	2006/Ireland	MEEI	Linear Discriminant Analysis

TABLE 4 (*Continued*)

ID	Reference	Title	Year/Country	Dataset	Machine Learning/AI Technique (Highest Accuracy)
57	77	Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion	2022/Turkey	SVD	SVM
58	78	Decoding phonation with artificial intelligence (DeP AI): Proof of concept	2019/United States	Privately collected dataset	CNN
59	79	A Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation	2021/Finland	HUPA, PC-GITA, and SVD	SVM
60	80	Development of a machine-learning based voice disorder screening tool	2022/Canada	SVD	Hybrid CNN and SVM
61	81	Automatic Voice Disorder Detection Using Self-Supervised Representations	2023/Spain	SVD and AVFAD	DNN
62	82	Support vector wavelet adaptation for pathological voice assessment	2011/Australia	MEEI	SVM
63	83	Wavelet adaptation for automatic voice disorders sorting	2013/Iran	MEEI	SVM
64	18	Advances in Automated Voice Pathology Detection: A Comprehensive Review of Speech Signal Analysis Techniques	2024/India	SVD	Hybrid 2-D CNN and LSTM
65	84	Unraveling the complexities of pathological voice through saliency analysis	2023/India	VOICED	Multi-Layer Perceptron, 1-D CNN, and 2-D CNN
66	85	Hierarchical Multi-Class Classification of Voice Disorders Using Self-Supervised Models and Glottal Features	2023/Finland	SVD	SVM
67	86	The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection	2024/Finland	SVD	SVM
68	87	Categorizing Normal and Pathological Voices: Automated and Perceptual Categorization	2011/Lithuania	Privately collected dataset	SVM
69	88	Voice disorder detection using machine learning algorithms: An application in speech and language pathology	2024/ United Kingdom	SVD, MEEI, and privately collected datasets	SVM
70	89	Exploring similarity-based classification of larynx disorders from human voice	2012/Lithuania	Privately collected dataset	SVM
71	90	Pathological assessment of patients' speech signals using nonlinear dynamical analysis	2010/Iran	MEEI	SVM
72	91	A Deep Learning Approach for Voice Disorder Detection for Smart Connected Living Environments	2022/Italy	MEEI, SVD, and VOICED	Light-CNN
73	92	Voice Disorder Identification by Using Machine Learning Techniques	2018/Italy	SVD	SVM
74	93	A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection	2023/India	VOICED	XGBoost
75	94	Pathological voice classification based on multi-domain features and deep hierarchical extreme learning machine	2023/China	VOICED	Hierarchical Extreme Learning Machine
76	95	Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach	2011/China	MEEI	SVM

TABLE 4 (*Continued*)

ID	Reference	Title	Year/Country	Dataset	Machine Learning/AI Technique (Highest Accuracy)
77	96	Automatic detection of vocal cord disorders using machine learning method for healthcare system	2024/Saudi Arabia	SVD	Sequential Learning Resource Allocation Neural Network
78	97	The accuracy of an Online Sequential Extreme Learning Machine in detecting voice pathology using the Malaysian Voice Pathology Database	2023/Malaysia	Privately collected dataset	OSELM
79	98	Pathological voice classification using MEEL features and SVM-TabNet model	2024/Saudi Arabia	SVD	SVM-TabNet fusion model

Dataset legend: SVD, Saarbruecken Voice Database Dataset; AVFAD, Advanced Voice Function Assessment Databases; AVPD, Arabic voice pathology database; FEMH, Far Eastern Memorial Hospital Voice Database; HUPA, Hospital Universitario Principe de Asturias database; MAPACI, MAPACI Speech Pathology Database; MEEI, Massachusetts Eye & Ear Infirmary Voice Disorders Database; Saarbruecken Voice Database; VOICED, VOice ICar fEDerico II Database.
Machine learning/AI technique legend: ANN, Artificial Neural Network; CNN, Convolutional Neural Network; DNN, Deep Neural Network; HMM, Hidden Markov Model; KNN, K-Nearest Neighbors; LSTM, Long Short-Term Memory; OSELM, Online Sequential Extreme Learning Machine; SVM, Support Vector Machine; XGBoost, Extreme Gradient Boosting.

matrix prior to being fed into the SVMs. With the addition of their novel preprocessing, SVMs demonstrated 95% accuracy in detecting voice disorders.[37] Mohammed et al (2023) developed a novel CNN-based hybrid network. This network, called MMHFNet, involved two CNN streams that extract voice features combined with hybrid connections that concatenate the features across streams. With the SVD voice database, the MMHFNet demonstrated 96% accuracy in detecting voice disorders.[76] Verma et al (2023) developed a hybrid model called VDDMFS that involved LSTM, ANN, and XGBoost. Following feature extraction, the LSTM model processed voice features, the ANN model processed metadata features (ie, age, sex, fundamental frequency, and spectral centroid), and the XGBoost stacked probabilities from both models into a feature matrix before classifying voice samples. Overall, VDDMFS demonstrated 96% accuracy in detecting voice disorders using the VOICED database.[95] Fonseca et al (2020) proposed a discriminant paraconsistent machine to classify voice disorders on a more fine-grained scale compared with the standard binary classification of "healthy versus disordered." Their approach projected the voice input data on a paraconsistent plan that allowed the input to be classified as exclusively one of two classes, neither classification, or both classifications. This can be relevant when a speaker has both Reinke's edema and laryngitis, for example. The discriminant paraconsistent machine demonstrated 96% accuracy in detecting voice disorders using the SVD database.[50]

Databases

The included studies tested their AI tools using a variety of voice databases, with 18 using multiple databases.[14,29,30,33,34,37,46,47,56,60-63,67,81,83,90,93] The most common database utilized across the 79 studies included in the present systematic review was the SVD ($n = 32$), followed by the MEEI ($n = 22$), privately collected databases ($n = 25$), the VOICED ($n = 11$), the AVPD ($n = 5$), the HUPA ($n = 4$), the AVFAD ($n = 2$), the MAPACI ($n = 1$), and the PC-GITA ($n = 1$).

Accuracy and other performance outcomes

The included studies examined the performance of their machine learning tool according to a variety of outcome measures. Of primary importance to the present systematic review is accuracy, which is calculated as

$$Accuracy = \frac{Correct\ Disordered + Correct\ Healthy}{Correct\ Disordered + Misclassfied\ Disorderd + Correct\ Healthy + Misclassfied\ Healthy}$$

Among the present studies, nine reported 100% accuracy in their machine learning model's ability to detect a voice disorder from a healthy.[28,36,44,47,56,84,85,90,100] Thirty-two studies reported between 95% and 99%, and the remaining 38 studies' accuracy ranged from 67% to 94%. Table 5

TABLE 5.
Accuracy Results of the Included Publications on the Ability for Machine Learning/AI Techniques to Detect Voice Disorders From Healthy Voices and the Percent Split Between the Training and Testing Datasets

ID	Reference	Machine Learning/AI Technique (Highest Accuracy)	Highest Accuracy in Detecting Voice Disorders (%)	% Trained Dataset/% Tested Dataset
1	25	LSTM	99.3%	70/30
2	26	OSELM	91.2%	80/20
3	27	CNN	94.1%	23/77
4	28	SVM	100%	70/30
5	29	SVM	99.8%	54/46
6	30	SVM	99.8%	54/46
7	31	Naive Bayes classifier	80%	80/20
8	32	SVM	98%	Not reported
9	33	LSTM and CNN	99.6%	74/26
10	34	SVM	95.9%	75/25
11	11	SVM	99.2%	80/20
12	35	SVM	94.3%	70/30
13	36	SVM	100%	70/30
14	37	SVM	95.2%	Not reported
15	38	Hybrid HMM and SVM	97.4%	70/30
16	39	Logistic Model Tree algorithm	77.4%	Not reported
17	40	DNN	98.6%	65/35
18	41	KNN	93.3%	80/20
19	42	CNN	95%	80/20
20	43	Generative Adversarial Network	95.6%	Not reported
21	44	ANN	100%	Not reported
22	45	Fine KNN	98.3%	71/29
23	46	CNN	82.2%	80/20
24	47	Random Forest	100%	93/7
25	48	DNN	99.3%	Not reported
26	49	CNN	92%	80/20
27	50	Discriminative Paraconsistent Machine	95%	Not reported
28	51	ANN	91%	70/30
29	52	SVM	95.5%	Not reported
30	53	Supervised Logistic Regression Model with Nested Cross-Validation and Forward Feature Selection	74.5%	Not reported
31	54	Learning Vector Quantization	96%	70/30
32	55	SVM	99.3%	90/10
33	56	KNN	100%	70/30
34	57	CNN	66.9%	80/20
35	58	CNN	83.3%	80/20
36	59	CNN	80.3%	80/20
37	60	2-Dimensional (2-D) CNN	80%	Not reported
38	61	KNN	98.5	70/30
39	62	CNN	97.9%	75/25
40	63	SVM	78.4%	95/5
41	64	1-Dimensional (1-D) CNN	85%	80/20
42	65	CNN	97%	80/20
43	66	Convolutional Bottleneck Network	88.8%	80/20
44	67	Feedforward Neural Network	99.3%	70/30
45	68	CNN	98.9%	70/30
46	69	CNN	85.5%	
47	70	Stochastic Gradient Descent Classifier	91%	Not reported
48	71	XGBoost	93.8%	80/20
49	72	ANN	80%	Not reported
50	73	Quadratic Discriminant Analysis	91.8%	Not reported
51	14	SVM	92.7%	80/20
52	74	Ensemble-CNN Decision Fusion	99.1%	90/10
53	75	CNN	67.9%	80/20
54	76	LSTM	96.1%	80/20

TABLE 5 (*Continued*)

ID	Reference	Machine Learning/AI Technique (Highest Accuracy)	Highest Accuracy in Detecting Voice Disorders (%)	% Trained Dataset/% Tested Dataset
55	[77]	CNN	95.4%	80/20
56	[78]	Linear Discriminant Analysis	89.1%	70/30
57	[79]	SVM	90.1%	Not reported
58	[80]	CNN	90%	91/9
59	[81]	SVM	95.4%	67/33
60	[82]	Hybrid CNN and SVM	97.8%	70/30
61	[83]	DNN	94%	Not reported
62	[84]	SVM	100%	75/25
63	[85]	SVM	100%	75/25
64	[18]	Hybrid 2-D CNN and LSTM	83.3%	90/10
65	[86]	Multi-Layer Perceptron, 1-D CNN, and 2-D CNN	97.1%	80/20
66	[87]	SVM	75.7%	Not reported
67	[88]	SVM	75.1%	95/5
68	[89]	SVM	92%	Not reported
69	[90]	SVM	100%	80/20
70	[91]	SVM	89%	90/10
71	[92]	SVM	94.4%	80/20
72	[93]	Light-CNN	84%	80/20
73	[94]	SVM	85.8%	Not reported
74	[95]	XGBoost	95.7%	Not reported
75	[96]	Hierarchical Extreme Learning Machine	99%	79/21
76	[97]	SVM	96.1%	Not reported
77	[98]	Sequential Learning Resource Allocation Neural Network	94.4%	75/25
78	[99]	OSELM	90%	80/20
79	[100]	SVM-TabNet fusion model	100%	80/20

presents the accuracy findings and the proportion of their training and testing data.

In addition to accuracy, several studies reported other performance metrics, including sensitivity, specificity, precision, and F1 score, which provide additional insight into the performance of the AI models. Sensitivity measures the model's ability to correctly identify disordered samples, while specificity captures its ability to correctly classify healthy voices. Precision reflects the proportion of correctly identified disordered samples out of all labeled as disordered, and F1 score represents the harmonic mean of precision and sensitivity, offering a balanced measure. These performance outcomes can be calculated as Chen and Chen[40]

$$Sensitivity = \frac{Correct\ Disordered}{Correct\ disordered + Misclassified\ Healthy}$$

$$Specificity = \frac{Correct\ Healthy}{Correct\ Healthy + Misclassified\ Disordered}$$

$$Precision = \frac{Correct\ Disorderd}{Correct\ Disorderd + Misclassfied\ Disorderd}$$

$$F1score = \frac{2(Correct\ disordered)}{2(Correct\ Disordered) + Misclassified\ Disorderd + Misclassfied\ Healthy}$$

Among the included studies, sensitivity values ranged from 63%[72] to 100%.[36,71,82] Specificity values ranged from 65%[58] to 100%.[36,71] Reported precision scores ranged from 73%[99] to 100%.[29] F1 scores varied from 74%[75] in Mittapalle et al (2023) to 99%.[74]

Quality assessment

The Quality Assessment of Diagnostic Accuracy Studies—Second Edition (QUADAS-2[24]) was used for this review. For selection bias, 54% of the included articles demonstrated a "High" rating. For index test bias, 71% of the included articles demonstrated a "High" rating. For reference standard bias, 94% of included articles demonstrated an "Unclear" rating. For patient flow bias, 46% of included articles demonstrated a "High" rating and 46% demonstrated an "Unclear" rating. For applicability of included patients, 96% of included articles demonstrated a "Low" rating. For applicability of index test, 91% of included articles demonstrated a "Low" rating. For applicability of reference standard, 100% of articles demonstrated a "Low" rating. The full results of the QUADAS-2 assessment are reported in Table 6.

DISCUSSION

This systematic review synthesized 79 research articles that used AI to detect voice disorders. The articles were assessed for methodological quality using the QUADAS-2 scales, and accuracy data were presented. Overall, there was exceptionally high accuracy in detecting voice disorders across the articles (mean accuracy = 92%). The subsequent discussion will integrate the included papers' details and provide recommendations for further research in this area. Across the papers, there is a trend of hyperoptimization—ie, the papers aim for slight increases in AI performance outcomes (on the order of 1%-5% better accuracy, for example). Given that present AI approaches demonstrate exceptionally high accuracy in detecting voice disorders, this review concludes that the hyperoptimization trend may be misguided. Rather than iterating AI tools to achieve slightly better outcomes, it would benefit clinical practice if these tools were made more accessible and if the models were trained using more expansive/representative databases, as opposed to standard datasets selected across studies. The present review extends prior systematic reviews by comparing methodological quality across multiple AI approaches, rather than focus on pooled accuracy estimates.[101] Additionally, the inclusion of the most recent studies through 2024 in the present systematic review builds upon, Idrisoglu et al[102] who examined a similar topic, but did not capture the latest advances in AI models or databases.

Hyperoptimization trend

Among the 79 papers, SVMs were used in over one-third ($n = 28$), and they demonstrated a weighted mean accuracy of approximately 94%. Deep learning families (ie, neural networks) were the second most common choice of AI tool. CNNs demonstrated a mean accuracy of approximately 88%. Other AI architectures (convolutional bottleneck networks, logistic model trees, linear/logistic discriminant analyses, and naive Bayes) demonstrated voice disorder detection accuracy ranging from approximately 77%-91%. These approaches were highly accurate, with the lower end approximating the interrater reliability of specialized speech language pathologists to perceptually rate dysphonia.[103] Across studies, an incremental improvement in accuracy is evident. Approximately 36% of the included papers demonstrated < 2% improvement in accuracy results compared with a prior included study that used the same voice database. These small improvements illuminate the hyperoptimization trend in which technical novelty appears to be prioritized over actual clinical impact.[104] Despite a clustering of highly accurate approaches, it remains unclear if these tools are ecologically valid. These machine learning techniques are not externally validated on active, clinical populations, for example.[105,106] Future work in the area of AI to detect voice disorders would benefit from prioritizing external validation on novel patients, with novel phonation tasks instead of ever-narrower optimization on widely used databases.

Voice database representation and generalizability

The studies included in this review relied heavily on a small set of publicly available voice disorder databases, which limits the generalizability of their findings. The SVD was the most commonly used ($n = 32$ studies). This German-language database includes sustained vowels and short sentences from both healthy speakers and individuals with a variety of laryngeal pathologies, but its recordings were collected under controlled, studio-like conditions.[8] The MEEI database ($n = 22$) was another frequently used corpus, consisting mainly of sustained vowels and short reading passages in English from both disordered and healthy speakers. However, the dataset's small number of healthy samples ($n = 53$) and lack of task variability limit its clinical applicability. The VOICED and AVPD databases, while used in fewer studies, offer some linguistic diversity (Italian and Arabic, respectively), yet also emphasize sustained vowels as the primary phonatory task. Additional databases such as HUPA, AVFAD, and MAPACI have been used infrequently and often with inconsistent disorder labeling, which complicates cross-study comparisons. Most datasets included in the present review feature narrow voice elicitation tasks (typically sustained vowels or simple reading tasks), recorded under ideal acoustic conditions. These constraints likely fail to capture critical dimensions of real-world voice use such as spontaneous speech, speech in noise, or across varying levels of vocal effort, which are routinely evaluated in clinical voice assessments (eg, van Mersbergen et al[107]). Furthermore, the demographic skew toward adult speakers and monolingual samples results in limited representation of pediatric and aging populations, both of which are highly relevant in voice care. Thus, while current machine learning models achieve high accuracy within these datasets, their generalizability to real-world contexts remains unproven.

Toward clinical implementation

Although the voice disorder detection systems driven by AI are increasingly accurate, it remains unclear if they demonstrate adequacy and ease-of-use for clinical deployment. Translation of laboratory results to clinical tools is facilitated by the emerging fields of implementation science[108] and dissemination science.[109] Conceptual frameworks within these fields include the Reach, Effectiveness, Adoption, Implementation, Maintenance framework (RE-AIM[110,111]) and the Consolidated Framework for Implementation Research (CFIR[112]). These frameworks emphasize that to achieve widespread clinical adoption and/or translation of a tool, more than superior technical performance is required. Rather, an innovation must demonstrate feasibility, acceptability, and sustained use in real-world settings. In this regard, the present AI voice disorder

TABLE 6.
Quality Assessment of Included Publications by Means of the Quality Assessment of Diagnostic Accuracy Studies—Second Edition (QUADAS-2)

ID	Year/Country/First Author	Selection Bias	Index Test bias	Reference Standard Bias	Patient Flow Bias	Applicability of Included Patients	Applicability of Index Test	Applicability of Reference Standard
1	2023/Iraq/Abdulmajeed	Low	High	Unclear	Unclear	Low	Low	Low
2	2021/Malaysia/Al-Dhief	Low	High	Unclear	Unclear	Low	Low	Low
3	2018/ Saudi Arabia /Alhussein	High	High	Unclear	High	Low	Low	Low
4	2016/ Saudi Arabia /Ali	High	High	Unclear	High	Low	Low	Low
5	2017/Saudi Arabia/Al-Nasheri	High	High	Unclear	High	Low	Low	Low
6	2017/ Saudi Arabia /Al-Nasheri	High	Low	Unclear	High	Low	Low	Low
7	2023/ Malaysia /Altaf	High	High	Unclear	Unclear	Low	High	Low
8	2017/Tunis/Amami	Unclear	Unclear	Unclear	Unclear	Low	Low	Low
9	2021/Turkey/Ankishan	Unclear	High	Unclear	Unclear	Low	Low	Low
10	2011/Greece/Arias-Londono	Unclear	High	Unclear	Unclear	Low	Low	Low
11	2011/ Greece/Arias-Londono	High	High	Unclear	High	Low	Low	Low
12	2011/Iran/Arjmandi	Low	High	Unclear	Unclear	Low	Low	Low
13	2012/Iran/Arjmandi	Unclear	High	Unclear	Unclear	Low	Low	Low
14	2023/ Saudi Arabia /Basalamah	High	High	Unclear	High	Low	Low	Low
15	2021/Algeria/Benhammoud	High	High	Unclear	High	Low	Low	Low
16	2018/Italy/Cesari	High	Low	Unclear	High	Low	Low	Low
17	2022/China/Chen	High	High	Unclear	High	Low	High	Low
18	2021/China/Chen	Unclear	Unclear	Unclear	Unclear	Low	Low	Low
19	2023/China/Chen	High	High	Unclear	High	Low	Low	Low
20	2020/China/Chui	Unclear	High	Unclear	Unclear	Low	Low	Low
21	2007/Brazil/Crovato	High	Unclear	Unclear	High	Low	Low	Low
22	2024/Italy/Di Cesare	Low	High	Unclear	Unclear	Low	Low	Low
23	2021/China/Ding	High	High	Unclear	High	Low	Low	Low
24	2021/China/Fan	Low	High	Unclear	Unclear	Low	Low	Low
25	2019/Taiwan/Fang	Low	High	Unclear	High	Low	Low	Low
26	2024/Iran/Farazi	High	Low	Unclear	High	Low	Low	Low
27	2020/Brazil/Fonseca	High	High	Unclear	High	Low	Low	Low
28	2009/Spain/Fraile	High	High	Unclear	High	Low	Low	Low
29	2008/Lithuania/Gelzinis	High	High	Unclear	High	Low	Low	Low
30	2024/USA/Ghasemzadeh	High	Low	Unclear	High	Low	Low	Low
31	2004/Spain/Godino-Llorente	High	High	Unclear	Unclear	Low	Low	Low
32	2020/Tunisa/Hammami	High	High	Unclear	High	Unclear	Low	Low
33	2014/Malaysia/Hariharan	Unclear	High	Unclear	High	Low	Low	Low
34	2021/Taiwan/Hu	High	High	Unclear	Low	Low	High	Low
35	2022/ Taiwan /Hung	Low	Low	Low	Low	Low	High	Low
36	2022/Canada/Islam	High	High	Unclear	High	Low	Low	Low
37	2024/Finland/Javanmardi	High	Unclear	Unclear	High	Low	Low	Low
38	2024/ India /Jegan	High	High	Unclear	High	Low	Low	Low
39	2024/India/Jegan	Unclear	High	Unclear	Unclear	Low	Low	Low
40	2020/Finland/Kadiri	Unclear	High	Unclear	Unclear	Low	Low	Low
41	2020/Korea/Kim	High	Low	Low	Low	Unclear	High	Low
42	2024/Korea /Kim	High	High	Unclear	High	Low	Low	Low
43	2024/Algeria/Korba	Unclear	High	Unclear	Unclear	Unclear	Low	Low
44	2020/Korea/Lee	Low	Unclear	Unclear	Unclear	Low	Low	Low
45	2023/Korea/Lee	Low	Unclear	Unclear	Unclear	Low	Low	Low
46	2024/Korea/Lee	High	High	Unclear	High	Low	Low	Low
47	2022/Brazil/Leite	High	High	Unclear	High	Low	Low	Low
48	2024/Brazil/Lima-Filho	High	High	Unclear	High	Low	Low	Low
49	2008/Germany/Linder	High	Unclear	Unclear	High	Low	Low	Low
50	2007/UK/Little	Low	High	Unclear	Unclear	Low	Low	Low
51	2017/Saudi Arabia/Mesallam	Low	Low	Unclear	Unclear	Low	Low	Low
52	2021/India/Mittal	Low	High	Unclear	Unclear	Low	Low	Low
53	2023/Finland/Mittapalle	Low	High	Unclear	Unclear	Low	Low	Low
54	2023/Turkey/Mohammed	Low	Unclear	Unclear	Unclear	Low	Low	Low
55	2020/Iraq/Mohammed	Unclear	High	Unclear	Unclear	Low	Low	Low
56	2006/Ireland/Moran	Low	High	Unclear	Unclear	Low	Low	Low
57	2022/Turkey/Omeroglu	Low	High	Unclear	Unclear	Low	Low	Low
58	2019/USA/Powell	High	High	Low	Low	Low	Low	Low
59	2021/Finland/Reddy	Unclear	High	Unclear	Unclear	Low	Low	Low
60	2022/Canada/Reid	Low	High	Unclear	Unclear	Low	Low	Low
61	2023/Spain/Ribas	Low	High	Unclear	Unclear	Low	Low	Low
62	2011/Australia/Saeedi	High	High	Unclear	High	Low	Low	Low
63	2013/Iran/Saeedi	Low	High	Unclear	Unclear	Low	Low	Low
64	2024/India/Sankaran	High	High	Unclear	High	Low	Low	Low
65	2023/India/Shaikh	Low	Low	Low	Unclear	Low	Low	Low
66	2023/Finland/Tirronen	High	High	Unclear	High	Low	Low	Low
67	2024/Finland/Tirronen	Unclear	High	Unclear	Unclear	Low	Low	Low
68	2011/Lithuania/Uloza	High	Low	Unclear	High	Low	Low	Low
69	2024/UK/Ur Rehman	High	Low	Unclear	High	Low	Low	Low
70	2012/Lithuania/Vaiciukynas	Unclear	High	Unclear	Unclear	Low	Low	Low
71	2010/Iran/Vaziri	High	High	Unclear	High	Low	Low	Low
72	2022/Italy/Verde	High	High	Unclear	High	Low	Low	Low
73	2018/Italy/Verde	High	Unclear	Unclear	High	Low	Low	Low
74	2023/India/Verma	Low	Unclear	Unclear	Unclear	Low	Low	Low
75	2023/China/Wang	High	Low	Low	Unclear	Low	High	Low
76	2011/China/Wang	Unclear	High	Unclear	Unclear	Low	Low	Low
77	2024/Saudi Arabia/Yadav	High	High	Unclear	Unclear	Low	Low	Low
78	2023/Malaysia/Za'im	High	Unclear	Unclear	High	Low	Low	Low
79	2024/Saudi Arabia/Zakariah	High	Low	Unclear	Low	Low	High	Low

detection methods remain untested. That is, few of the included studies report how their approach may influence clinical decision-making. Additionally, none of the included studies were deployed in clinical workflows. Future efforts should prioritize ecological validity by integrating models into routine screening in high-risk populations (eg, teachers[113-117]). Importantly, AI voice disorder detection methods are not intended to replace the existing multidimensional approach taken to assess voice disorders, which encompasses auditory-perceptual evaluation,[118] acoustic analysis,[119,120] and patient self-reports,[121] among other evaluation techniques.

CONCLUSION

The present systematic review found that, across 79 studies, AI models consistently achieved high accuracy in detecting voice disorders when tested on established datasets. Machine learning approaches such as SVMs and CNNs were the most commonly implemented, with average classification accuracy reaching 92% across studies. While the AI models demonstrated high accuracy in detecting voice disorders, the quality assessment indicated that a majority of included studies exhibited high or unclear risk of bias in several domains. Overall, the limited diversity of datasets and the frequent reliance on sustained vowels recorded in controlled settings restrict the generalizability of the findings to real-world clinical populations. Future research would benefit from greater emphasis on ecological validity, including testing on spontaneous speech, noisy environments, and diverse populations, to support translation into clinical screening tools.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Gonzalo Farid Saud Medina, Ryan Anderson, Allie Benivegna, Naomi Ha, Kate Harty, and Kaliyah Houe-Henry for their involvement.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jvoice.2025.09.021.

References

1. Bhattad PB, Jain V. Artificial intelligence in modern medicine–the evolving necessity of the present and role in transforming the future of medical care. *Cureus*. 2020;12:e8041.
2. Piccialli F, Di Somma V, Giampaolo F, et al. A survey on deep learning in medicine: Why, how and when? *Inf Fusion*. 2021;66:111–137.
3. Kühl N, Schemmer M, Goutier M, Satzger G. Artificial intelligence and machine learning. *Electr Markets*. 2022;32:2235–2244.
4. Panesar A. *Machine Learning and AI for Healthcare*. New York, NY: Springer,; 2019.
5. Aldrich C, Auret L. *Unsupervised Process Monitoring and Fault Diagnosis With Machine Learning Methods*. New York, NY: Springer,; 2013.
6. Uçar MK, Nour M, Sindi H, Polat K. The effect of training and testing process on machine learning in biomedical datasets. *Math Problems Eng*. 2020;2020:2836236. https://doi.org/10.1155/2020/2836236.
7. Syed SA, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. *Math Biosci Eng MBE*. 2020;17:7958–7979. https://doi.org/10.3934/mbe.2020404.
8. Barry W, Pützer, M. Saarbrucken voice database. Institute of Phonetics; 2007. [Online]. Available: http://stimmdatenbank.coli.unisaarland.de.
9. Jesus L, Hall A, Belo I, Machado J. The advanced voice function assessment databases (AVFAD): tools for voice clinicians and speech research. In: Fernandes FDM, ed. *Advances in Speech-language Pathology*. London, UK: IntechOpen; 2017. https://doi.org/10.5772/intechopen.69643.
10. MEEI. Voice disorders database, version. 1.03 (cd-rom) [Dataset]. Kay Elemetrics Corporation; 1994.
11. Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Trans Bio-Med Eng*. 2011;58:370–379. https://doi.org/10.1109/TBME.2010.2089052. (MEDLINE).
12. Godino-Llorente, J.I. Personal Communication via Email with Juan Ignacio Godino-Llorente [Personal communication]; 2025.
13. Zhang Y, Qian J, Zhang X, et al. Pathological voice detection using joint subsapce transfer learning. *Appl Sci*. 2022;12:8129. https://doi.org/10.3390/app12168129.
14. Mesallam TA, Farahat M, Malki KH, et al. Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. *J Healthc Eng*. 2017;2017:8783751.
15. Ramalingam A, Kedari S, Vuppalapati C. IEEE FEMH voice data challenge; 2018. 5272-5276.
16. Cesari U, De Pietro G, Marciano E, et al. A new database of healthy and pathological voices. *Comput Electr Eng*. 2018;68:310–321. https://doi.org/10.1016/j.compeleceng.2018.04.008.
17. MAPACI P. Voice Disorder Database; 2004.
18. Sankaran A, Kumar LS. Advances in automated voice pathology detection: a comprehensive review of speech signal analysis techniques. *IEEE Access*. 2024:18114–181127. https://doi.org/10.1109/ACCESS.2024.3508884. (Scopus).
19. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. *IEEE Access*. 2022;10:122136–122158.
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*. 2021;372:n71.
21. Grames EM, Stillman A, Tingley MW, Elphick C. Litsearchr: Automated search term selection and search strategy for systematic reviews; 2019b.
22. Grames EM, Stillman AN, Tingley MW, Elphick CS. An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods Ecol Evol*. 2019;10:1645–1654.
23. Nudelman CJ, Bottalico P, Cantor-Cutiva LC. The effects of room acoustics on self-reported vocal fatigue: a systematic review. *J Voice*. 2025;39:1131.e11–1131.e30.
24. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Int Med*. 2011;155:529–536.

25. Abdulmajeed NQ, Al-Khateeb B, Mohammed MA. Voice pathology identification system using a deep learning approach based on unique feature selection sets. *Exp Syst.* 2023;42:e13327.
26. Al-Dhief FT, Baki MM, Latiff NMA, et al. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access.* 2021;9:77293–77306. https://doi.org/10.1109/ACCESS.2021.3082565. (Scopus).
27. Alhussein M, Muhammad G. Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access.* 2018;6:41034–41041. https://doi.org/10.1109/ACCESS.2018.2856238. (Scopus).
28. Ali Z, Alsulaiman M, Elamvazuthi I, et al. Voice pathology detection based on the modified voice contour and SVM. *Biol Inspir Cogn Arch.* 2016;15:10–18. https://doi.org/10.1016/j.bica.2015.10.004.
29. Al-Nasheri A, Muhammad G, Alsulaiman M, et al. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access.* 2017;6:6961–6974. https://doi.org/10.1109/ACCESS.2017.2696056. (Scopus).
30. Al-Nasheri A, Muhammad G, Alsulaiman M, Ali Z. Investigation of voice pathology detection and classification on different frequency regions using correlation functions. *J Voice.* 2017;31:3–15. https://doi.org/10.1016/j.jvoice.2016.01.014. (Music Index).
31. Altaf A, Mahdin H, Maskat R, et al. A novel voice feature AVA and its application to the pathological voice detection through machine learning. *Int J Adv Comput Sci Appl.* 2023;14:1085–1092. https://doi.org/10.14569/IJACSA.2023.01409113. (Scopus).
32. Amami R, Smiti A. An incremental method combining density clustering and support vector machines for voice pathology detection. *Comput Electr Eng.* 2017;57:257–265.
33. Ankışhan H, İnam SÇ. Voice pathology detection by using the deep network architecture. *Appl Soft Comput.* 2021;106:107310. https://doi.org/10.1016/j.asoc.2021.107310.
34. Arias-Londoño JD, Godino-Llorente JI, Markaki M, Stylianou Y. On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. *Logop Phoniatr Vocol.* 2011;36:60–69.
35. Arjmandi MK, Pooyan M, Mikaili M, et al. Identification of voice disorders using long-time features and support vector machine with different feature reduction methods. *J Voice Off J Voice Found.* 2011;25:e275–e289. https://doi.org/10.1016/j.jvoice.2010.08.003. (MEDLINE).
36. Arjmandi M, Pooyan M. An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomed Signal Process Control.* 2012;7:3–19. https://doi.org/10.1016/j.bspc.2011.03.010.
37. Basalamah A, Hasan M, Bhowmik S, Shahriyar SA. A highly accurate dysphonia detection system using linear discriminant analysis. *Comput Syst Sci Eng.* 2023;44:1921–1938. https://doi.org/10.32604/csse.2023.027399. (Scopus).
38. Benhammoud R, Kacha A. Automatic classification of disordered voices based on a hybrid HMM-SVM Model. *J Commun Technol Electron.* 2021;66:S139–S148. https://doi.org/10.1134/S1064226921140023.
39. Cesari U, De Pietro G, Marciano E, et al. Voice disorder detection via an m-health system: design and results of a clinical study to evaluate Vox4Health. *BioMed Res Int.* 2018;2018:8193694. https://doi.org/10.1155/2018/8193694.(MEDLINE Ultimate).
40. Chen L, Chen J. Deep neural network for automatic classification of pathological voice signals. *J Voice.* 2022;36:288.e15–288.e24. https://doi.org/10.1016/j.jvoice.2020.05.029. (Scopus).
41. Chen L, Wang C, Chen J, et al. Voice disorder identification by using Hilbert-Huang Transform (HHT) and K nearest neighbor (KNN). *J Voice.* 2021;35:932-e1. https://doi.org/10.1016/j.jvoice.2020.03.009.
42. Chen Z, Zhu P, Qiu W, et al. Deep learning in automatic detection of dysphonia: comparing acoustic features and developing a generalizable framework. *Int J Lang Commun Disord.* 2023;58:279–294. https://doi.org/10.1111/1460-6984.12783. (MEDLINE).
43. Chui KT, Lytras MD, Vasant P. Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset. *Appl Sci (Switzerland).* 2020;10:4571. https://doi.org/10.3390/app10134571. (Scopus).
44. Crovato CDP, Schuck A. The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices. *IEEE Trans Biomed Eng.* 2007;54:1898–1900. https://doi.org/10.1109/TBME.2006.889780. (Scopus).
45. Di Cesare MG, Perpetuini D, Cardone D, Merla A. Assessment of voice disorders using machine learning and vocal analysis of voice samples recorded through smartphones. *Bio Med Inf.* 2024;4:549–565.
46. Ding H, Gu Z, Dai P, et al. Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. *Biomed Signal Process Control.* 2021;70:102973. https://doi.org/10.1016/j.bspc.2021.102973. (Scopus).
47. Fan Z, Wu Y, Zhou C, et al. Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. *Appl Sci (Switzerland).* 2021;11:3450. https://doi.org/10.3390/app11083450. (Scopus).
48. Fang S-H, Tsao Y, Hsiao M-J, et al. Detection of pathological voice using cepstrum vectors: a deep learning approach. *J Voice.* 2019;33:634–641. https://doi.org/10.1016/j.jvoice.2018.02.003. (Music Index).
49. Farazi S, Shekofteh Y. Voice pathology detection on spontaneous speech data using deep learning models. *Int J Speech Technol.* 2024;27:739–751. https://doi.org/10.1007/s10772-024-10134-4.(Communication Source).
50. Fonseca ES, Guido RC, Junior SB, et al. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). *Biomed Signal Process Control.* 2020;55:101615. https://doi.org/10.1016/j.bspc.2019.101615. (Scopus).
51. Fraile R, Sáenz-Lechón N, Godino-Llorente JI, et al. Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex. *Folia Phoniatr Logop Off Organ Int Assoc Logop Phoniatr (IALP).* 2009;61:146–152. https://doi.org/10.1159/000219950. (MEDLINE).
52. Gelzinis A, Verikas A, Bacauskiene M. Automated speech analysis applied to laryngeal disease categorization. *Comput Methods Programs Biomed.* 2008;91:36–47.
53. Ghasemzadeh H, Hillman RE, Mehta DD. Consistency of the signature of phonotraumatic vocal hyperfunction across different ambulatory voice measures. *J Speech Lang Hear Res.* 2024;67:1997–2020. https://doi.org/10.1044/2024_JSLHR-23-00515.
54. Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Trans Biomed Eng.* 2004;51:380–384.
55. Hammami I, Salhi L, Labidi S. Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features. *IRBM.* 2020;41:161–171. https://doi.org/10.1016/j.irbm.2019.11.004. (Scopus).
56. Hariharan M, Polat K, Yaacob S. A new feature constituting approach to detection of vocal fold pathology. *Int J Syst Sci.* 2014;45:1622–1634. https://doi.org/10.1080/00207721.2013.794905.
57. Hu H-C, Chang S-Y, Wang C-H, et al. Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study. *J Med Int Res.* 2021;23:1645–1654.N.PAG-N.PAG. Library & Information Science Source.
58. Hung CH, Wang SS, Wang CT, Fang SH. Using SincNet for learning pathological voice disorders. *Sensors (Basel, Switzerland).* 2022;22:6634. https://doi.org/10.3390/s22176634. (MEDLINE Ultimate).
59. Islam R, Abdel-Raheem E, Tarique M. Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. *Comput Methods Programs Biomed Update.* 2022;2:100074. https://doi.org/10.1016/j.cmpbup.2022.100074. (Scopus).
60. Javanmardi F, Kadiri SR, Alku P. A comparison of data augmentation methods in voice pathology detection. *Comput Speech Lang.* 2024;83:1–16.

61. Jegan R, Jayagowri R. Optimized early fusion of handcrafted and deep learning descriptors for voice pathology detection and classification. *Healthcare Anal.* 2024;6:100369. https://doi.org/10.1016/j.health.2024.100369. (Scopus).
62. Jegan R, Jayagowri R. Voice pathology detection using optimized convolutional neural networks and explainable artificial intelligence-based analysis. *Comput Methods Biomech Biomed Eng.* 2024;27:2041–2057.
63. Kadiri S, Alku P. Analysis and detection of pathological voice using glottal source features. *IEEE J Select Topics Signal Process.* 2020;14:367–379. https://doi.org/10.1109/JSTSP.2019.2957988.
64. Kim H, Jeon J, Han YJ, et al. Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. *J Clin Med.* 2020;9:3415.
65. Kim H-B, Song J, Park S, Lee YO. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. *Sci Rep.* 2024;14:1–13.
66. Korba M, Doghmane H, Khelil K, Messaoudi K. Improved laryngeal pathology detection based on Bottleneck Convolutional Networks and MFCC. *IEEE Access.* 2024;12:124801–124815. https://doi.org/10.1109/ACCESS.2024.3454825.
67. Lee J, Choi H-J. Deep learning approaches for pathological voice detection using heterogeneous parameters. *IEICE Trans Inf Syst.* 2020;E103D:1920–1923. https://doi.org/10.1587/transinf.2020EDL8031. (Scopus).
68. Lee J-N, Lee J-Y. An efficient SMOTE-based deep learning model for voice pathology detection. *Appl Sci (Switzerland).* 2023;13:3571. https://doi.org/10.3390/app13063571.(Scopus).
69. Lee JH, Seok J, Kim JY, et al. Evaluating the Diagnostic Potential of Connected Speech for Benign Laryngeal Disease Using Deep Learning Analysis. (MEDLINE). *Journal of Voice: Official Journal of the Voice Foundation.* 2024. https://doi.org/10.1016/j.jvoice.2024.01.015.
70. Leite DRA, de Moraes RM, Lopes LW. Different performances of machine learning models to classify dysphonic and non-dysphonic voices. *J Voice Off J Voice Found.* 2022;39:577–590. https://doi.org/10.1016/j.jvoice.2022.11.001. (MEDLINE).
71. Lima-Filho LMA, Lopes LW, Filho T de MES. Integrated vocal deviation index (IVDI): a machine learning model to classifier of the general grade of vocal deviation. *J Voice Off J Voice Found.* 2024. https://doi.org/10.1016/j.jvoice.2024.11.002. 0892-1997, (MEDLINE).
72. Linder R, Albers AE, Hess M, et al. Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *J Voice.* 2008;22:155–163. https://doi.org/10.1016/j.jvoice.2006.09.003. (Music Index).
73. Little MA, McSharry PE, Roberts SJ, et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMed Eng OnLine.* 2007;6:23.
74. Mittal V, Sharma R. Deep learning approach for voice pathology detection and classification. *Int J Healthc Inf Syst Inf.* 2021;16:1–30. https://doi.org/10.4018/IJHISI.20211001.oa28.
75. Mittapalle KR, Yagnavajjula MK, Alku P. Classification of functional dysphonia using the tunable Q wavelet transform. *Speech Commun.* 2023;155:102989.
76. Mohammed HMA, Omeroglu AN, Oral EA. MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection. *Expert Syst Appl.* 2023;223:119790.
77. Mohammed MA, Abdulkareem KH, Mostafa SA, et al. Voice pathology detection and classification using convolutional neural network model. *Appl Sci (Switzerland).* 2020;10:3723. https://doi.org/10.3390/app10113723. (Scopus).
78. Moran RJ, Reilly RB, de Chazal P, Lacy PD. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Trans Bio-Med Eng.* 2006;53:468–477. https://doi.org/10.1109/TBME.2005.869776. (MEDLINE).
79. Omeroglu AN, Mohammed HMA, Oral EA. Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Eng Sci Technol Int J.* 2022;36:101148. https://doi.org/10.1016/j.jestch.2022.101148. (Scopus).
80. Powell ME, Rodriguez Cancio M, Young D, et al. Decoding phonation with artificial intelligence (DeP AI): proof of concept. *Laryngosc Investig Otolaryngol.* 2019;4:328–334.
81. Reddy M, Alku P. A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation. *IEEE ACCESS.* 2021;9:135953–135963. https://doi.org/10.1109/ACCESS.2021.3117665.
82. Reid J, Parmar P, Lund T, et al. Development of a machine-learning based voice disorder screening tool. *Am J Otolaryngol.* 2022;43:103327. https://doi.org/10.1016/j.amjoto.2021.103327. (MEDLINE).
83. Ribas D, Pastor MA, Miguel A, et al. Automatic voice disorder detection using self-supervised representations. *IEEE Access.* 2023;11:14915–14927. https://doi.org/10.1109/ACCESS.2023.3243986. (Scopus).
84. Saeedi NE, Almasganj F, Torabinejad F. Support vector wavelet adaptation for pathological voice assessment. *Comput Biol Med.* 2011;41:822–828. https://doi.org/10.1016/j.compbiomed.2011.06.019. (MEDLINE).
85. Saeedia NE, Almasganj F. Wavelet adaptation for automatic voice disorders sorting. *Comput Biol Med.* 2013;43:699–704.
86. Shaikh AAS, Bhargavi MS, Naik GR. Unraveling the complexities of pathological voice through saliency analysis. *Comput Biol Med.* 2023;166:107566. https://doi.org/10.1016/j.compbiomed.2023.107566. (Scopus).
87. Tirronen S, Kadiri S, Alku P. Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. *IEEE Open J Signal Process.* 2023;4:80–88. https://doi.org/10.1109/OJSP.2023.3242862.
88. Tirronen S. The effect of the MFCC frame length in automatic voice pathology detection. *J Voice.* 2022;8:975–982.
89. Uloza V, Verikas A, Bacauskiene M, et al. Categorizing normal and pathological voices: automated and perceptual categorization. *J Voice.* 2010;25:700–708.(in press-in press.).
90. Ur Rehman M, Shafique A, Azhar Q-U-A, et al. Voice disorder detection using machine learning algorithms: an application in speech and language pathology. *Eng Appl Artif Intell.* 2024;133:108047.
91. Vaiciukynas E, Verikas A, Gelzinis A, et al. Exploring similarity-based classification of larynx disorders from human voice. *Speech Commun.* 2012;54:601–610.
92. Vaziri G, Almasganj F, Behroozmand R. Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Comput Biol Med.* 2010;40:54–63.
93. Verde L, Brancati N, De Pietro G, et al. A deep learning approach for voice disorder detection for smart connected living environments. *ACM Trans Int Technol.* 2022;22:1–16. https://doi.org/10.1145/3433993. (Scopus).
94. Verde L, De Pietro G, Sannino G. Voice disorder identification by using machine learning techniques. *IEEE Access.* 2018;6:16246–16255. https://doi.org/10.1109/ACCESS.2018.2816338. (Scopus).
95. Verma V, Benjwal A, Chhabra A, et al. A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Sci Rep.* 2023;13:22719. https://doi.org/10.1038/s41598-023-49869-6. (MEDLINE).
96. Wang J, Xu H, Peng X, et al. Pathological voice classification based on multi-domain features and deep hierarchical extreme learning machine. *J Acoust Soc Am.* 2023;153:423–435.
97. Wang X, Zhang J, Yan Y. Discrimination between pathological and normal voices using GMM-SVM approach. *J Voice.* 2011;25:38–43. https://doi.org/10.1016/j.jvoice.2009.08.002. (Music Index).
98. Yadav K. Automatic detection of vocal cord disorders using machine learning method for healthcare system. *Int J Syst Assur Eng Manag.* 2024;15:429–438. https://doi.org/10.1007/s13198-022-01761-8. (Scopus).

99. Za'im NAN, AL-Dhief FT, Azman M, et al. The accuracy of an online sequential extreme learning machine in detecting voice pathology using the Malaysian Voice Pathology Database. *J Otolaryngol Head Neck Surg.* 2023;52:1–11.
100. Zakariah M, Al-Razgan M, Alfakih T. Pathological voice classification using MEEL features and SVM-TabNet model. *Speech Commun.* 2024;162:103100.
101. Al-Hussain G, Shuweihdi F, Alali H, et al. The effectiveness of supervised machine learning in screening and diagnosing voice disorders: systematic review and meta-analysis. *J Med Int Res.* 2022;24:e38472. https://doi.org/10.2196/38472.
102. Idrisoglu A, Dallora AL, Anderberg P, Berglund JS. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: systematic literature review. *J Med Int Res.* 2023;25:e46105. https://doi.org/10.2196/46105.
103. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech-Lang Pathol.* 2011;20:14–22. https://doi.org/10.1044/1058-0360(2010/09-0105).
104. Balagopalan A, Baldini I, Celi LA, et al. Machine learning for healthcare that matters: reorienting from technical novelty to equitable impact. *PLOS Digital Health.* 2024;3:e0000474.
105. Cabitza F, Campagner A, Soares F, et al. The importance of being external. Methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed.* 2021;208:106288. https://doi.org/10.1016/j.cmpb.2021.106288.
106. Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PloS One.* 2020;15:e0234722.
107. van Mersbergen M, Beckham BH, Hunter EJ. Do we need a measure of vocal effort? Clinician's report of vocal effort in voice patients. *Perspect ASHA Spec Int Groups.* 2021;6:69–79. https://doi.org/10.1044/2020_persp-20-00258.
108. Douglas NF, Feuerstein JL, Oshita JY, et al. Implementation science research in communication sciences and disorders: a scoping review. *Am J Speech-Lang Pathol.* 2022;31:1054–1083.
109. Hart MK, Laures-Gore J, Peele S. What about dissemination science? Practical recommendations for the clinical researcher in communication sciences and disorders. *Am J Speech-Lang Pathol.* 2025;34:2351–2358.
110. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. *Am J Publ Health.* 2013;103:e38–e46.
111. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Publ Health.* 1999;89:1322–1327.
112. Drake CD, Lewinski AA, Zullig LL. Consolidated framework for implementation research (CFIR). *Int Encycl Health Commun.* 2023:1–9. https://doi.org/10.1002//9781119678816.iehc0944.
113. Behlau M, Zambon F, Guerrieri AC, Roy N. Epidemiology of voice disorders in teachers and nonteachers in Brazil: prevalence and adverse effects. *J Voice.* 2012;26:665.e9–665.e18. https://doi.org/10.1016/j.jvoice.2011.09.010.
114. Nudelman CJ, Flaherty MM, Bottalico P. Altered auditory feedback in teachers: a preliminary investigation. *J Voice.* 2024. https://doi.org/10.1016/j.jvoice.2024.10.015. S089219972400359X.
115. Pinto Giannini SP, Ferreira LP. Voice disorders in teachers and the international classification of functioning, disability and health (ICF). *Revista de Investigación e Innovación En Ciencias de La Salud.* 2021;3:33–47.
116. Roy N, Merrill RM, Thibeault S, et al. Prevalence of voice disorders in teachers and the general population. *J Speech Lang Hear Res.* 2004;47:281–293.
117. Trinite B. Epidemiology of voice disorders in latvian school teachers. *J Voice Off J Voice Found.* 2017;31:508.e1–508.e9. https://doi.org/10.1016/j.jvoice.2016.10.014.
118. Kempster GB, Nagle KF, Solomon NP. Development and rationale for the Consensus Auditory-Perceptual Evaluation Of Voice—revised (CAPE-Vr). *J Voice.* 2025. https://doi.org/10.1016/j.jvoice.2025.01.022.
119. Bottalico P, Codino J, Cantor-Cutiva LC, et al. Reproducibility of voice parameters: the effect of room acoustics and microphones. *J Voice.* 2018;34:320–334. https://doi.org/10.1016/j.jvoice.2018.10.016.
120. Castillo-Allendes A, Codino J, Cantor-Cutiva LC, et al. Clinical utility and validation of the acoustic voice quality and acoustic breathiness indexes for voice disorder assessment in english speakers. *J Clin Med.* 2023;12:7679.
121. Nudelman CJ, Bottalico P, van Mersbergen M, Nanjundeswaran C. Toward enhanced voice-related self-reports: translation, cross-cultural adaptation, and validity. *J Voice.* 2024.