# Speech and Voice Quality as Digital Biomarkers in Depression: A Systematic Review☆

*,†,‡Giovanni Briganti, and §,¶,‖,**Jérôme R. Lechien, *§*Mons,* †*Liège,* ‡*Bruxelles,* ¶*Baudour, Belgium, and* ‖**Paris, France*

**Summary: Objective**. To review the current evidence on the use of artificial intelligence-driven speech and voice analysis as a biomarker for depression.

**Methods**. PubMed, Scopus, and Cochrane databases were reviewed by two independent investigators for studies investigating the use of artificial intelligence-driven speech and voice quality outcomes as biomarkers for depression according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statements. The methodological quality and risk of bias of each included study were assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool.

**Results**. Of the 108 identified records, 12 studies met the inclusion criteria. The studies examined 16 872 participants, including patients with major depressive disorder ($n$ = 1535), bipolar disorder ($n$ = 111), schizophrenia spectrum disorders ($n$ = 35), and anxiety disorders ($n$ = 224). Control groups included a total of 1204 healthy individuals. Speech and voice quality outcomes consistently distinguished depression from controls (AUC = 0.71-0.93), with prosodic, spectral, and perturbation measures showing significant correlations with standardized depression scales. Classification accuracies ranged from 78% to 96.5%. Six studies demonstrated high risk of methodological bias, primarily in patient selection and validation techniques. Voice recording contexts varied between clinical settings and mobile technologies.

**Conclusion**. The findings of this review highlight the potential of voice biomarkers as a novel tool for depression detection and monitoring. While current evidence demonstrates promising classification accuracy, methodological heterogeneity and generalizability concerns must be addressed before widespread clinical adoption.

**Key Words:** Speech—Voice—Otolaryngology—Otorhinolaryngology—Laryngeal—Larynx—Acoustic—Biomarker—Mood monitoring—Machine learning—Artificial intelligence.

## INTRODUCTION

Major depressive disorder (MDD) is a prevalent and debilitating psychiatric condition characterized by persistent low mood, cognitive disturbances, and psychomotor changes.[1] Depression significantly impairs daily functioning and is associated with an increased risk of suicide, cardiovascular disease, and metabolic disorders.[2] Despite advancements in pharmacological and psychotherapeutic interventions, accurate and timely assessment of depressive symptoms remains a major challenge in clinical practice. Current diagnostic and monitoring approaches rely primarily on clinical interviews and self-reported symptom scales, which, although widely used, are subject to recall bias, variability in patient adherence, and limited accessibility.[3] There is a critical need for objective biomarkers that can facilitate real-time, accurate, and scalable monitoring of depression to improve clinical outcomes.[4]

Speech analysis has emerged as a promising noninvasive tool for detecting and monitoring depression. While artificial intelligence applications represent recent innovations, the foundation of speech analysis for detecting emotions and psychological states dates back to the early 1970s. Pioneering research established fundamental connections between vocal acoustics and emotional states, developing theoretical frameworks that linked physiological changes to specific vocal modifications.[5–7] Alterations in vocal parameters, including pitch, prosody, speech rate, and spectral features, have been hypothesized to reflect the neurophysiological and psychomotor changes associated with depressive states.[8] Depressive speech is often characterized by reduced vocal intensity, slower speech tempo, and increased acoustic perturbations, which may serve as measurable biomarkers for mood disorders.[9,10] With the advancement of artificial intelligence-driven voice analysis, particularly through machine learning and deep learning techniques, it has become possible to extract and classify complex speech features associated with depression. The integration of voice biomarkers into digital health platforms, including smartphone-based and telemedicine applications, has expanded the potential for continuous, ecologically valid monitoring of patients in real-world settings.

Accepted for publication May 1, 2025.
☆ This research received no external funding.

From the *Unit of Computational Medicine and Neuropsychiatry, Faculty of Medicine, Pharmacy and Biomedical Sciences, University of Mons (UMONS), Mons, Belgium; †Department of Clinical Sciences, Faculty of Medicine, University of Liège, Liège, Belgium; ‡Faculty of Medicine, Université libre de Bruxelles, Bruxelles, Belgium; §Surgery Department, Research Institute for Language Science and Technology, University of Mons (UMons), Mons, Belgium; ¶Division of Laryngology and Bronchoesophagology, Department of Otolaryngology Head Neck Surgery, EpiCURA Hospital, Baudour, Belgium; ‖Department of Otolaryngology-Head and Neck Surgery, Foch Hospital, School of Medicine, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France; and the **Department of Otolaryngology, Elsan Hospital, Paris, France.

Address correspondence and reprint requests to: Giovanni Briganti, M.D., M.Sc., Ph.D., Unit of Computational Medicine and Neuropsychiatry, Faculty of Medicine, University of Mons, Avenue du Champ de Mars 6, 7000 Mons, Belgium. E-mail: giovanni.briganti@hotmail.com

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



https://doi.org/10.1016/j.jvoice.2025.05.002

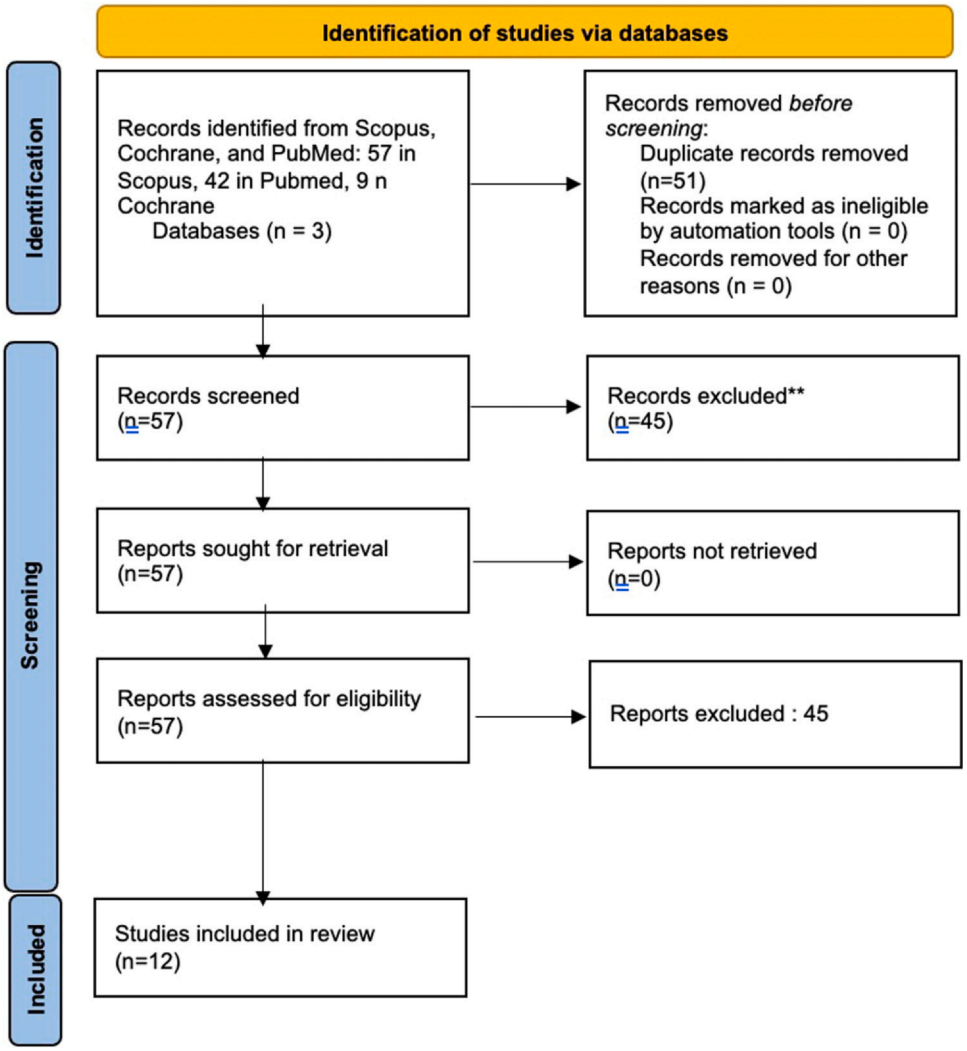


**FIGURE 1.** PRISMA flowchart.

Several studies have investigated the relationship between voice characteristics and depression, applying a wide range of methodological approaches and artificial intelligence models: these studies differ in the type of speech data analyzed, whether from controlled reading tasks or spontaneous speech, the machine learning techniques applied, and the validation methods used.[11–22] While findings suggest that voice-based biomarkers hold potential for depression detection, the variability in study designs and the lack of standardized methodologies limit their generalizability and clinical implementation.

The objective of this systematic review was to summarize the current evidence on the use of artificial intelligence-driven voice analysis as a biomarker for depression. The authors will analyze existing studies on acoustic and linguistic features of speech in depressed individuals, as well as the machine learning models used for classification. This review aims to assess the reliability, validity, and limitations of voice-based biomarkers in the context of depression detection Figure 1.

## MATERIALS AND METHODS

### Framework for data extraction

Two independent investigators conducted the systematic review and data extraction following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for systematic reviews.[8] The inclusion and exclusion criteria were based on the Population, Intervention, Comparison, Outcome, Timing, and Setting framework. Data were extracted systematically using a predefined approach to ensure comprehensive and reproducible findings. Two reviewers performed data extraction independently to maintain accuracy and consistency, with discrepancies resolved through discussion.

#### *Patient population*

The systematic review included prospective, retrospective, controlled, and uncontrolled studies investigating the use of artificial intelligence-based voice analysis in patients with

depression. Eligible studies focused on adults or adolescents diagnosed with MDD or related depressive conditions according to standardized diagnostic criteria, including DSM-5 and ICD-10. Studies specifying sample populations, including demographics, depression severity categorized as mild, moderate, or severe, comorbid psychiatric conditions, and control groups, were included. Studies analyzing mixed diagnostic groups without separate analyses for MDD were excluded.

### *Voice biomarkers and AI models*

Studies assessing acoustic, prosodic, spectral, and linguistic voice features in individuals with depression were included. Controlled speech tasks, such as reading standardized texts, and naturalistic speech samples, including spontaneous speech, telemedicine calls, and smartphone-based recordings, were considered. The extracted voice features included fundamental frequency, jitter, shimmer, harmonic-to-noise ratio, MFCCs, spectral tilt, speech rate, intensity variation, and word choice analyzed through natural language processing. Studies implementing machine learning or deep learning models to classify or predict depression status based on these features were reviewed. Machine learning approaches included supervised models such as SVMs, random forests, and deep neural networks, unsupervised learning methods such as clustering and principal component analysis, and hybrid models.

### *Outcomes*

Primary outcomes included classification accuracy, sensitivity, specificity, and area under the curve of artificial intelligence-based models for detecting depression from voice features. Studies that assessed the predictive validity of voice biomarkers for depression severity or treatment response were included. Secondary outcomes focused on correlations between voice features and clinical rating scales, including the Hamilton Depression Rating Scale,[23] the Patient Health Questionnaire-9,[24] and the Montgomery-Åsberg Depression Rating Scale.[25] Studies that provided comparisons between depressed and non-depressed individuals or examined longitudinal variations in speech features across different mood states were analyzed. Validation strategies, including k-fold cross-validation, external test datasets, and reproducibility checks, were documented where reported.

### *Timing and settings*

The included studies covered a range of settings, including clinical environments such as hospitals, psychiatric clinics, and structured laboratory assessments, as well as smartphone-based assessments conducted through mobile applications and telemedicine platforms. Some studies collected voice samples in real-world conditions, including conversational artificial intelligence interactions, wearable devices, and telephone calls. Studies employing single-session assessments without concurrent clinical evaluation or without specification of timing and settings were excluded. Studies with longitudinal monitoring of mood variations over weeks or months were included.

### *Search strategy*

A systematic search was conducted using PubMed, Scopus, and Cochrane databases to identify studies evaluating artificial intelligence-driven voice analysis for depression detection. Studies published between January 2015 and March 2025 were considered. The search strategy combined terms related to depression and voice. The extracted data included study design, country, setting, sample characteristics such as age, gender, and depression severity, voice biomarkers analyzed, artificial intelligence modeling techniques, validation methods, and outcome measures. Only peer-reviewed studies published in English were included.

### *Bias analysis*

The methodological quality and risk of bias of each included study were assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. This framework evaluates the risk of bias across four domains, including patient selection, index test, reference standard, and flow and timing, along with applicability concerns related to patient selection and the generalizability of the artificial intelligence model. Each study was rated based on transparency in data reporting, patient inclusion methodology, model validation, and the presence of potential confounding factors.

## RESULTS

A total of 108 records were identified through database searches, including 57 from Scopus, 42 from PubMed, and 9 from Cochrane. Following the removal of duplicates and screening for eligibility, 12 studies met the inclusion criteria for systematic review.[11–22] These studies included nine cross-sectional studies [11–14,16,17,19,20,22] and three prospective studies.[15,18,21] Eight studies incorporated a control group for comparison.[12–17,19,20,22] Study demographics, designs, and outcomes are presented in Table 1, with detailed population characteristics summarized in Table 2.

The included studies examined a total of 16 872 participants, comprising patients with MDD ($n = 1535$), bipolar disorder ($n = 111$), schizophrenia spectrum disorders ($n = 35$), and anxiety disorders ($n = 224$). Control groups included a total of 1204 healthy individuals. The mean age of participants across studies varied between 26.2 and 64 years. Two studies did not report age data, while one study provided only an age range of 23-69 years. Gender data were inconsistently reported, with 754 females and 464 males identified across studies, leaving 15 654 participants without specified gender information in the original

**TABLE 1.**
**Study Design and Outcomes**

| References | Design | Settings | *N* | F/M | Age (m/md) | Outcomes | Results |
|---|---|---|---|---|---|---|---|
| Mazur et al 2025[11] | Cross-sectional | Remote recording (personal devices) | 14 898 | 69.4%/ 27.9%* | 37.3 (md) | Detection of depression with acoustic (F0, jitter shimmer) and prosodic features | Sensitivity: 71.3% Specificity: 73.5% |
| Menne et al 2024[12] | Cross-sectional Controlled | Laboratory (tablet microphone) | 96 (44 MDD 52 HC) | 38/58 | 26.2 (m) | Detection of depression with acoustic temporal and lexical features | AUC = 0.93 (speech model) |
| Ghosh et al 2024[13] | Cross-sectional Controlled | Laboratory (DAIC-WOZ dataset) | 189 (59 depressed 130 HC) | NS | NS | Depression detection with acoustic textual and visual features | Accuracy: 81.6% (BERT) 68% (BiLSTM) |
| Huang et al 2024[14] | Cross-sectional Controlled | Laboratory (DAIC-WOZ dataset) | 189 (56 depressed 133 HC) | 87/102 | NS | Depression detection and severity classification with wav2vec 2.0 features | Binary accuracy: 96.49% Multiclass: 94.81% RMSE: 0.1875 |
| Ronneberg et al 2024[15] | RCT-controlled | Voice-based AI coaching | 200 (mild-moderate depression) | NS | > 18 | Depression symptom reduction with voice-based AI intervention | Neural engagement correlated with symptom reduction |
| Cansel et al 2023[16] | Cross-sectional Controlled | Smartphone | 104 (15 SZ 24 ANX 25 MDD 15 BD 25 HC) | 78/26 | 28.9-37.5 | Disorder classification with acoustic features (MFCC GTCC) | SVM: 96.48% kNN: 96.94% accuracy |
| Berardi et al 2023[17] | Cross-sectional Controlled | Digital voice recorder | 60 (20 SZ 20 MDD 20 HC) | 23/37 | 39.3-41.6 | Disorder classification with acoustic features | HC vs MDD: 92.0% SZ vs MDD: 93.2% |
| Li et al 2023[18] | Cross-sectional | Clinical interview recordings | 329 (233 MDD 96 BD) | 217/112 | 31.4 | Depression severity classification with NLP | 4-level severity: F1 = 0.719 Binary (mild vs severe): F1 = 0.884 |
| Kim et al 2023[19] | Cross-sectional Controlled | Smartphone | 318 (153 MDD 165 HC) | 225/93 | 37.2-38.6 | Depression detection with acoustic features | CNN: 78.14% RF: 72.80% AUC: 0.86 |
| Wasserzug et al 2023[20] | Cross-sectional Controlled | Mobile phone conversations | 144 (40 MDD including 14 in remission 104 HC) | 73/71 | NS | Depression detection and remission with prosodic features | Significant differentiation ($P < 0.0001$) correlation with HAM-D ($r = 0.364$) |

TABLE 1 (*Continued*)

| References | Design | Settings | *N* | F/M | Age (m/md) | Outcomes | Results |
|---|---|---|---|---|---|---|---|
| Lin et al 2022[21] | Cross-sectional Controlled | Smartphone app | 263 (240 with depression history 23 HC) | 155/105* | 63.0 | Depression detection with NLP | AUC (PHQ-8) = 0.82 PPV=0.54 NPV = 0.88 |
| Hansen et al 2022[22] | Cross-sectional Controlled | Clinical interview recordings | 82 (40 MDD including 25 in remission 42 HC) | 64/18 | 30.9-37.3 | Depression detection and remission with acoustic and emotion features | AUC = 0.71 (MDD vs HC) |

* The gender of some individuals was not specified. *Abbreviations:* ANX, anxiety disorder; AUC, area under the curve; BD, bipolar disorder; BERT, bidirectional encoder representations from transformers; BiLSTM, bidirectional long short-term memory; CNN, convolutional neural network; DAIC-WOZ, Distress Analysis Interview Corpus—Wizard of Oz; F0, fundamental frequency (pitch); F1, F1 score (harmonic mean of precision and recall); F/M, female/male ratio; GTCC, gamma-tone cepstral coefficients; HAM-D/HAM-D, Hamilton Depression Rating Scale; HC, healthy controls; kNN, k-nearest neighbors; m, mean; MDD, major depressive disorder; MFCC, mel-frequency cepstral coefficients; N, sample size; NLP, natural language processing; NPV, negative predictive value; NS, not specified; PHQ-8/9, Patient Health Questionnaire (8 or 9 items); PPV, positive predictive value; RCT, randomized controlled trial; RF, random forest; RMSE, root mean square error; SD, standard deviation; SVM, support vector machine; SZ, schizophrenia.

**TABLE 2.**
**Summary of Demographics, Populations, and Outcomes**

| Category | *N* |
|---|---|
| *Gender* | |
| Females | 754 |
| Males | 464 |
| Unaccounted gender | 15 654 |
| Mean age (years) | Varied by study, range: 26-64; mean 37.5 |
| *Populations* | |
| Major depressive disorder | 1535 |
| Bipolar disorders | 111 |
| Schizophrenia spectrum disorders | 35 |
| Anxiety disorder | 224 |
| Healthy individuals | 1204 |
| Unaccounted population | 13 763 |
| *Setting/Devices* | |
| Clinical evaluations (hospital) | 8 |
| Smartphone | 4 |
| *Outcomes* | |
| F0 (standard deviation, variability, and mean) | 10 |
| Spectral signal and formants | 8 |
| Percent jitter | 6 |
| Voice intensity | 5 |
| Speech rate | 4 |
| Percent shimmer | 6 |
| Pause duration (connected speech) | 3 |
| Harmonic-to-noise ratio | 3 |
| Mel-frequency cepstral coefficients | 5 |

N, number; NS, not specified.

publications. Our analysis identified significant reporting gaps: 13 763 participants (81.6% of the total sample) lacked clear diagnostic categorization, while 15 654 participants (92.8%) had unspecified gender information. These gaps primarily stem from one study of 14 898 participants,[11] which reported depression screening but provided limited demographic and diagnostic details.

## Voice biomarkers and AI models

All included studies assessed voice as a potential biomarker for depression, incorporating various acoustic, spectral, and linguistic analyses. Prosodic features were examined in ten studies,[11–14,16,17,19,20,22] primarily focusing on fundamental frequency variations, speech rate, and voice intensity. These studies consistently reported decreased fundamental frequency and reduced speech tempo in depressed individuals. While voice intensity was analyzed in several studies, quantitative findings specific to this parameter were not consistently reported, representing a gap in the current literature that warrants further investigation. Spectral features, including MFCCs and spectral tilt, were analyzed in eight studies.[12–14,16,17,19,20,22] Jitter and

**TABLE 3.**
**QUADAS-2 Analysis**

| References | Patient Selection Bias | Index Test Bias | Reference Standard Bias | Flow and Timing Bias | Applicability Concerns | Final QUADAS-2 Judgment |
|---|---|---|---|---|---|---|
| Mazur et al, 2025[11] | High (social media recruitment, selection bias) | Low (validated ML system, rigorous feature extraction) | High (self-reported PHQ-9, no independent clinical verification) | High (cross-sectional, no predictive validity assessment) | Moderate (US and Canada sample, but limited generalizability) | High risk of bias due to selection method and lack of clinical validation |
| Menne et al, 2024[12] | High (small psychiatric sample, voluntary recruitment) | Low (validated ML model, rigorous speech feature extraction) | Moderate (psychiatric evaluation, but no external validation) | High (cross-sectional, no follow-up to assess predictive validity) | Moderate (Germany and France sample, limited generalizability) | Moderate risk of bias due to limited sample and lack of external validation |
| Ghosh et al, 2024[13] | High (public dataset, potential sample bias) | Low (transformer-based model, multimodal validation) | High (clinical interview dataset, but no direct clinician validation) | High (cross-sectional, dataset augmentation instead of longitudinal validation) | Moderate (DAIC-WOZ dataset, but no broad demographic diversity) | High risk of bias due to reliance on a single dataset and lack of real-world validation |
| Huang et al, 2024[14] | High (small dataset, limited linguistic and cultural diversity) | Low (wav2vec 2.0 pretrained model, automatic feature extraction) | High (PHQ-8 labels, no clinician confirmation) | High (cross-sectional study, no long-term follow-up) | Moderate (DAIC-WOZ dataset, unclear applicability to wider populations) | High risk of bias due to dataset constraints and no external clinical validation |
| Ronneberg et al, 2024[15] | Moderate (randomized trial, limited to mild-moderate depression) | Low (validated AI-PST system, previous pilot RCT) | Low (validated psychological scales, clinical supervision) | Moderate (short-term follow-up but no long-term validation) | Moderate (US-based clinical sample, potential recruitment bias) | Moderate risk of bias due to trial limitations but higher validity than others |
| Cansel et al, 2023[16] | High (small, nonrandom sample, voluntary participation) | Low (clearly defined ML model, rigorous feature extraction) | High (clinical diagnosis, but no independent validation) | High (cross-sectional, no follow-up for predictive validity) | Moderate (single-site study, no linguistic diversity) | High risk of bias due to sample selection and lack of independent validation |
| Berardi et al, 2023[17] | High (small, nonrandom sample, psychiatric in-/outpatients) | Low (validated ML model, rigorous acoustic feature extraction) | High (clinical diagnosis confirmed by SCID-I, but no external validation) | High (cross-sectional study, no predictive validation) | High (single-language study, limited generalizability) | High risk of bias due to small sample size and lack of longitudinal validation |
| Li et al, 2023[18] | High (no control group, only depressive patients) | Moderate (validated NLP approach, but no speech features) | High (HAM-D-17 clinical validation, but no external standard) | High (cross-sectional study, no predictive tracking) | High (linguistic based only, lacks multimodal assessment) | High risk of bias due to missing speech features and lack of observational behavioral data |

TABLE 3 (*Continued*)

| References | Patient Selection Bias | Index Test Bias | Reference Standard Bias | Flow and Timing Bias | Applicability Concerns | Final QUADAS-2 Judgment |
|---|---|---|---|---|---|---|
| Kim et al, 2023[19] | High (selection bias, no external validation, and controlled recording environment) | Low (validated CNN model, rigorous acoustic feature extraction) | High (clinical diagnosis confirmed by DSM-5, but no external validation) | High (cross-sectional study, no predictive tracking) | High (limited to Korean language, controlled speech task, and lacks real-world generalizability) | High risk of bias due to selection bias, controlled recording, and lack of external validation |
| Wasserzug et al, 2023[20] | High (small sample, nonrandom recruitment, and potential commercial bias) | Low (validated ML model, rigorous acoustic feature extraction) | High (clinical diagnosis confirmed by HAM-D, but no external validation) | High (cross-sectional study, no predictive tracking) | High (language-independent but commercially driven, limited dataset) | High risk of bias due to small sample, commercial funding, and lack of external validation |
| Lin et al, 2022[21] | High (small control group, self-reported history of depression) | Moderate (validated NLP model, but no acoustic analysis) | High (PHQ-8 and GAD-7 used, but no independent clinical validation) | High (cross-sectional study, no predictive tracking) | High (smartphone-based, excludes nonusers, and lacks multimodal assessment) | High risk of bias due to small control group, lack of independent clinical validation, and commercial funding |
| Hansen et al, 2022[22] | High (small sample, limited to first-episode MDD patients) | Moderate (validated ML model, trained on emotional speech, and no clinical depression) | High (HAM-D used, but no external clinical validation) | High (cross-sectional study, no predictive tracking, and no follow-up on nonremitters) | High (limited to Danish language, may not generalize to non-Germanic languages) | High risk of bias due to small sample size, lack of external validation, and limited generalizability |

*Abbreviations*: AI-PST, artificial intelligence-problem-solving therapy; CNN, convolutional neural network; DAIC-WOZ, Distress Analysis Interview Corpus—Wizard of Oz; DSM-5, Diagnostic and Statistical Manual of Mental Disorders, 5th Edition; GAD-7, Generalized Anxiety Disorder 7-item scale; HAM-D/HAM-D, Hamilton Depression Rating Scale; HAM-D-17, Hamilton Depression Rating Scale (17-item version); ML, machine learning; MDD, major depressive disorder; NLP, natural language processing; PHQ-8, Patient Health Questionnaire (8-item version); PHQ-9, Patient Health Questionnaire (9-item version); QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies, 2nd version; RCT, randomized controlled trial; SCID-I, Structured Clinical Interview for DSM Disorders, Axis I; US, United States; wav2vec 2.0, speech representation learning framework.

shimmer, representing measures of voice perturbation, were investigated in six studies,[11–14,16,19] with findings suggesting increased perturbation among individuals with depression compared with healthy controls. Pause duration and harmonic-to-noise ratio were assessed in three studies,[12,17,20] indicating longer pause times and decreased harmonicity in depressed speech patterns.

Supervised machine learning was implemented in all 12 studies, with SVMs and random forest classifiers being the most used algorithms. Classification accuracies varied across studies, with SVMs achieving performance ranging from 78% to 96.48%. Deep learning models, particularly convolutional neural networks, were utilized in three studies,[13,14,21] demonstrating good predictive performance in voice-based depression detection. Cross-validation techniques were applied in seven studies,[12–14,16,19,20,22] with fivefold and tenfold cross-validation being the most frequently used approaches.

### *Predictive validity of voice biomarkers*

The diagnostic accuracy of voice-based artificial intelligence models varied depending on the specific classification task and methodological approach. In distinguishing depressed individuals from healthy controls, classification performance ranged from AUC values of 0.71 to 0.93.[12,22] The differentiation between depression and other psychiatric conditions yielded mixed results, performing in distinguishing depression from schizophrenia with 93.2% accuracy[17] and in differentiating MDD from bipolar disorder.[18] Longitudinal studies demonstrated the feasibility of using voice biomarkers for symptom tracking, with individual-specific models showing stronger predictive correlations compared with population-level approaches.[15,21] Home monitoring systems reported promising results, with specific correlations between vocal features and standardized depression scales.[11,19,20] Sensitivity values in real-world applications ranged from 71.3% to 96.05%,[11,14] while specificity values ranged from 73.5% to 99.23%.[11]

### *Correlation with symptom scales and clinical validation*

The correlation between voice biomarkers and established depression rating scales was assessed in most included studies. Significant correlations were observed between fundamental frequency variability and Hamilton Depression Rating Scale scores, with vocal parameters associated with symptom severity.[20] Several studies demonstrated moderate correlations between voice-derived features and depression severity as measured by the Patient Health Questionnaire-9,[21] while others reported significant associations between jitter, shimmer, and Clinical Global Impression scores.

Temporal stability of these correlations was evaluated in longitudinal studies,[15,21] with follow-up durations varying across studies. Individual-specific models demonstrated stronger clinical correlations than population-based models, supporting the potential for personalized voice-based mental health monitoring.[15]

### *Timing and settings*

Data collection strategies varied among the studies, as four integrated mobile sensing technologies,[11,19–21] while eight implemented structured voice recording tasks in clinical settings.[12–18,22] The integration of artificial intelligence voice analysis into digital health platforms was evident, with studies using real-world voice samples collected through smartphone applications and telemedicine platforms. Recording frequency ranged from single-session clinical assessments to repeated voice recordings collected over several months. Circadian influences on voice characteristics were addressed in two studies,[18,21] with findings suggesting potential variations in voice parameters depending on time-of-day recording.

### *Bias analysis*

The methodological quality of included studies was assessed using the QUADAS-2 tool. High risk of bias was identified in six studies,[11,14,18–21] primarily due to patient selection biases and lack of external validation. Issues related to reference standard bias were present in four studies,[11,14,18,19] where voice-derived features were compared against self-reported symptom scales rather than clinician-rated assessments. Cross-validation and independent test set validation were adequately performed in seven studies,[12–14,16,19,20,22] while three studies lacked robust validation techniques.[11,18,21] The generalizability of findings was limited in studies that focused on a single language or cultural group, with applicability concerns highlighted in four studies.[17–19,22]

## DISCUSSION

This systematic review synthesized evidence from twelve studies investigating the role of voice biomarkers in the detection and monitoring of depression. The findings suggest that acoustic and prosodic features of speech, when analyzed using machine learning techniques, can provide valuable insights into depression severity, symptom progression, and mood state classification. While the results demonstrate promising classification performance, methodological heterogeneity, limitations in generalizability, and potential biases warrant further investigation.

Across studies, fundamental frequency (F0) variability, speech rate, and voice intensity were consistently identified as key prosodic markers of depression.[11–14,17,19] Depressed individuals exhibited lower mean F0, decreased speech tempo, and reduced vocal intensity, supporting prior literature on psychomotor retardation in depression. Spectral features, including mel-frequency cepstral coefficients (MFCC) and spectral tilt, were also implicated.[12–14,16,17,19]

Additionally, voice perturbation measures such as jitter and shimmer were frequently elevated in individuals with depression,[11–13,16,19] reflecting increased vocal instability.

Machine learning models demonstrated moderate-to-high accuracy in differentiating depressed individuals from healthy controls. Support vector machines (SVM) and random forest classifiers were the most frequently employed supervised learning techniques,[12,13,16,17,19] achieving classification accuracies ranging from 70.9% to 96.5%. Deep learning models, particularly convolutional neural networks (CNNs), were implemented in select studies,[13,14,19] showing good performance in voice-based depression detection. However, despite encouraging results, variability in feature selection, classification tasks, and performance metrics complicates direct comparison across studies.

Longitudinal studies assessing the predictive validity of voice biomarkers revealed that individualized models outperformed population-based approaches.[15,21] Personalized models demonstrated stronger correlations with symptom severity as measured by standard depression scales,[12,18–20,22] reinforcing the potential for real-world deployment of voice-based monitoring tools. Notably, smartphone-based assessments emerged as a growing trend, with studies incorporating passive voice monitoring and structured voice recording tasks.[11,19–21]

Despite promising results, several limitations must be addressed. First, methodological heterogeneity across studies limits comparability. Differences in sample sizes, diagnostic criteria, recording environments, and feature extraction methods introduce variability in findings. Some studies[11,14,21] relied on self-reported symptom scales for depression classification rather than clinician-rated assessments, raising concerns about reference standard bias. Additionally, the absence of standardized voice data collection protocols poses challenges for reproducibility and external validation.

Second, demographic disparities and reporting inconsistencies hinder generalizability. A substantial proportion of participants had unspecified demographic information, particularly regarding gender distribution.[11,13–15,18,21] Where reported, female participants were overrepresented in MDD cohorts, potentially affecting model performance and applicability across populations. Furthermore, the age range of participants varied widely, from early adulthood to late-life depression, suggesting that voice biomarkers may exhibit different predictive properties across developmental stages.

Third, cultural and linguistic variability remains a critical issue. Most studies did not account for language differences in voice analysis, which may influence the acoustic properties of speech and the effectiveness of AI-based classification models. As voice biomarkers are inherently influenced by phonetic structure, future research should consider cross-linguistic validation to enhance model robustness and global applicability.

To advance the field, standardized protocols could include recording both structured speech tasks and naturalistic conversation in the same individuals; collecting samples at multiple time points to account for circadian variations; documenting acoustic parameters using open-source analysis tools with transparent parameters; and implementing machine learning with mandatory cross-validation techniques. Additionally, future datasets should prioritize demographic diversity through balanced gender representation, inclusion of multiple age groups (adolescent, adult, and geriatric), cross-linguistic sampling with standardized translation protocols, and representation of varied depression severities and subtypes verified through standardized clinical assessments.

To advance the field of voice-based depression detection, future studies should prioritize methodological standardization, demographic diversity, and cross-linguistic validation. Establishing large-scale, multisite datasets with standardized voice recording protocols will enhance reproducibility and facilitate model comparison. Additionally, leveraging explainable AI techniques could improve clinical interpretability by identifying which specific vocal features contribute most significantly to depression classification.

Further research is also needed to explore the ecological validity of voice biomarkers in real-world settings. The transition toward passive voice monitoring via smartphones and wearable devices presents an opportunity for continuous, nonintrusive mental health assessment.[24,25] However, ensuring data privacy and ethical considerations remains paramount. Finally, integrating voice biomarkers with multimodal data—such as facial expression analysis and physiological signals—could enhance predictive accuracy and enable a more comprehensive approach to digital mental health monitoring.

## CONCLUSION

The findings of this review highlight the potential use of AI to analyze voice biomarkers as a novel tool for depression detection and monitoring. While current evidence demonstrates promising classification accuracy, methodological heterogeneity and generalizability concerns must be addressed before widespread clinical adoption. Standardized protocols, diverse datasets, and real-world validation efforts will be crucial for translating voice-based AI models into practical psychiatric applications. With continued advancements in AI and digital health, voice biomarkers could play a pivotal role in the future of remote mental health assessment and personalized treatment strategies.

### Data Availability Statement

Not applicable.

### Declaration of Competing Interest

The authors have no financial interest in the subject under discussion. All authors have read and approved the paper.

## Acknowledgments

None.

## Author Contributions

**Giovanni Briganti**: Design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. **Jerome R. Lechien**: Design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## Institutional Review Board Statement

Not required.

## Informed Consent Statement

Not applicable.

## Supplementary Materials

None.

## References

1. Goldberg D. The heterogeneity of "major depression". *World Psychiatry.* 2011;10:226–228.
2. Belmaker RH, Agam G. Major depressive disorder. *N Engl J Med.* 2008;358:55–68.
3. Marx W, Penninx BW, Solmi M, et al. Major depressive disorder. *Nat Rev Dis Primer.* 2023;9:44.
4. Strawbridge R, Young AH, Cleare AJ. Biomarkers for depression: recent insights, current challenges and future prospects. *Neuropsychiatr Dis Treat.* 2017;13:1245–1262.
5. Alpert M, Kurtzberg RL, Friedhoff AJ. Transient voice changes associated with emotional stimuli. *Arch Gen Psychiatry.* 1963;8:362–365.
6. Scherer KR. Vocal affect expression: a review and a model for future research. *Psychol Bull.* 1986;99:143.
7. Scherer KR, Banse R, Wallbott HG, Goldbeck T. Vocal cues in emotion encoding and decoding. *Motiv Emot.* 1991;15:123–148.
8. Arevian AC, Bone D, Malandrakis N, et al. Clinical state tracking in serious mental illness through computational analysis of speech. *PloS One.* 2020;15:e0225695.
9. Koops S, Brederoo SG, De Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. *CNS Neurol Disord - Drug Targets.* 2023;22:152–160.
10. Talavera JA, Saiz-Ruiz J, Garcia-Toro M. Quantitative measurement of depression through speech analysis. *Eur Psychiatry.* 1994;9:185–193.
11. Mazur A, Costantino H, Tom P, Wilson MP, Thompson RG. Evaluation of an AI-based voice biomarker tool to detect signals consistent with moderate to severe depression. *Ann Fam Med.* 2025;23:60–65.
12. Menne F, Dörr F, Schräder J, et al. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry.* 2024;24:794.
13. Ghosh D, Karande H, Gite S, Pradhan B. Psychological disorder detection: a multimodal approach using a transformer-based hybrid model. *MethodsX.* 2024;13:102976.
14. Huang X, Wang F, Gao Y, et al. Depression recognition using voice-based pre-training model. *Sci Rep.* 2024;14:12734.
15. Ronneberg CR, Lv N, Ajilore OA, et al. Study of a PST-trained voice-enabled artificial intelligence counselor for adults with emotional distress (SPEAC-2): design and methods. *Contemp Clin Trials.* 2024;142:107574.
16. Cansel N, Faruk Alcin Ö, Furkan Yılmaz Ö, Ari A, Akan M, Ucuz İ.A. New artificial intelligence-based clinical decision support system for diagnosis of major psychiatric diseases based on voice analysis. *Psychiatr Danub.* 2023;35:489–499.
17. Berardi M, Brosch K, Pfarr JK, et al. Relative importance of speech and voice features in the classification of schizophrenia and depression. *Transl Psychiatry.* 2023;13:298.
18. Li N, Feng L, Hu J, et al. Using deeply time-series semantics to assess depressive symptoms based on clinical interview speech. *Front Psychiatry.* 2023;14:1104190.
19. Kim AY, Jang EH, Lee SH, Choi KY, Park JG, Shin HC. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. *J Med Internet Res.* 2023;25:e34474.
20. Wasserzug Y, Degani Y, Bar-Shaked M, et al. Development and validation of a machine learning-based vocal predictive model for major depressive disorder. *J Affect Disord.* 2023;325:627–632.
21. Lin D, Nazreen T, Rutowski T, et al. Feasibility of a machine learning-based smartphone application in detecting depression and anxiety in a generally senior population. *Front Psychol.* 2022;13:811517.
22. Hansen L, Zhang YP, Wolf D, Sechidis K, Ladegaard N, Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatr Scand.* 2022;145:186–199.
23. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23:56–62.
24. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16:606–613.
25. Davidson J, Turnbull CD, Strickland R, Miller R, Graves K. The Montgomery-Asberg depression scale: reliability and validity. *Acta Psychiatr Scand.* 1986;73:544–548.