Review

Performance of Automatic Speech Analysis in Detecting Depression: Systematic Review and Meta-Analysis

Patricia Laura Maran[1,2], MSc; María Dolores Braquehais[1,3,4,5], MD, PhD; Alexandra Vlaic[6], MD; María Teresa Alonzo-Castillo[6], BSc; Júlia Vendrell-Serres[1,6], MD; Josep Antoni Ramos-Quiroga[1,2,3,6], MD, PhD; Amanda Rodríguez-Urrutia[1,2,3,6], MD, PhD

[1]Psychiatry, Mental Health and Addictions Group, Vall d'Hebron Research Institute (VHIR), Instituto de Investigación Sanitaria Acreditado Instituto de Investigación - Hospital Universitario Vall d'Hebron (IR-HUVH), Barcelona, Catalonia, Spain

[2]Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain

[3]Biomedical Network Research Centre on Mental Health (CIBERSAM), Barcelona, Spain

[4]Integral Care Programme for Sick Health Professionals, Galatea Clinic, Barcelona, Spain

[5]School of Medicine, Universitat Internacional de Catalunya, Barcelona, Spain

[6]Department of Psychiatry, Vall d'Hebron Hospital Universitari, Barcelona, Spain

Corresponding Author:
María Dolores Braquehais, MD, PhD
Psychiatry, Mental Health and Addictions Group, Vall d'Hebron Research Institute (VHIR), Instituto de Investigación Sanitaria Acreditado Instituto de Investigación - Hospital Universitario Vall d'Hebron (IR-HUVH)
VHIR Edifici Central, Pg. de la Vall d'Hebron, 129, Horta-Guinardó
Barcelona, Catalonia, 08035
Spain
Phone: 34 932057267
Email: dolores.braquehais@vhir.org

Abstract

Background: Despite the high prevalence and significant burden of depression, underdiagnosis remains a persistent challenge. Automatic speech analysis (ASA) has emerged as a promising method for depression assessment. However, a comprehensive quantitative synthesis evaluating its diagnostic accuracy is still lacking.

Objective: This systematic review and meta-analysis aimed to assess the diagnostic performance of ASA in detecting depression, considering both machine learning and deep learning approaches.

Methods: We conducted a systematic search across 8 databases, including MEDLINE, PsycInfo, Embase, CINAHL, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar from January 2013 to April 1, 2025. We included studies published in English that evaluated the accuracy of ASA for detecting depression, and reported performance metrics such as accuracy, sensitivity, specificity, precision, or confusion matrices. Study quality was assessed using a modified version of the Quality Assessment of Studies of Diagnostic Accuracy-Revised. A 3-level meta-analysis was performed to estimate the pooled highest and lowest accuracy, sensitivity, specificity, and precision. Meta-regressions and subgroup analyses were performed to explore heterogeneity across various factors, including type of publication, artificial intelligence algorithms, speech features, speech-eliciting tasks, ground truth assessment, validation approach, dataset, dataset language, participants' mean age, and sample size.

Results: Of the 1345 records identified, 105 studies met the inclusion criteria. The pooled mean of the highest accuracy, sensitivity, specificity, and precision were 0.81 (95% CI 0.79 to 0.83), 0.84 (95% CI 0.81 to 0.86), 0.83 (95% CI 0.79 to 0.86), and 0.81 (95% CI 0.77 to 0.84), respectively, whereas the pooled mean of the lowest accuracy, sensitivity, specificity, and precision were 0.66 (95% CI 0.63 to 0.69), 0.63 (95% CI 0.58 to 0.68), 0.60 (95% CI 0.55 to 0.66), and 0.64 (95% CI 0.58 to 0.70), respectively.

Conclusions: ASA shows promise as a method for detecting depression, though its readiness for clinical application as a standalone tool remains limited. At present, it should be regarded as a complementary method, with potential applications across diverse contexts. Further high-quality, peer-reviewed studies are needed to support the development of robust, generalizable models and to advance this emerging field.

Trial Registration: PROSPERO CRD42023444431; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023444431

(JMIR Ment Health 2025;12:e67802) doi: 10.2196/67802

KEYWORDS

depression; AI; automatic speech analysis; meta-analysis; artificial intelligence; mobile phone

Introduction

Depression represents a significant global health challenge, ranking among the leading causes of disability and premature mortality. Approximately 280 million people worldwide are estimated to endure major depressive disorder, representing 3.8% of the global population [1]. Depression is associated with substantial societal and economic costs [2], loss in productivity [3], and a reduced quality of life [4,5]. Additionally, depression is a key risk factor for a spectrum of chronic health conditions, including arthritis, asthma, cancer, and cardiovascular diseases [6]. Notably, there is evidence that patients who endure both depression and chronic physical conditions tend to have considerably lower quality of life than those enduring a chronic physical condition solely [7], with patients with depression dying 5 to 10 years earlier due to chronic physical conditions [8]. Perhaps most alarmingly, depression is inextricably linked to suicide [9], with people enduring depression being 20 times more likely to commit suicide than the general population [5]. Consequently, the early detection of depression and its timely intervention assume paramount significance.

Current state-of-the-art depression assessment primarily relies on operational diagnostic criteria from the *DSM-5-TR* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition, Text Revision]) [10] and the *ICD-11* (*International Classification of Diseases, 11th Revision*) [11]. When considering screening, the Patient Health Questionnaire-9 [12] is the most commonly used tool for depression due to its brevity, ease of use, and validation across various languages and contexts [13]. However, these assessments are susceptible to a range of biases arising from the interviewer's experience, the quality of the question protocol, and the patient's willingness to communicate their symptoms [14]. Field trials of the *DSM-5* criteria for major depressive disorder have revealed strikingly low interrater reliability [15]. Hence, the acquisition of diagnostic information through these traditional means is not only time-consuming but also demands considerable clinical training and practice to yield reliable results [16]. Adding to these complexities is the stark shortage of mental health professionals worldwide, which represents one of the biggest obstacles to the early detection of depression [17]. For instance, there are approximately 9 psychiatrists per 100,000 people in developed countries [18] and as few as 0.1 for every 1,000,000 in middle- and lower-income countries [19]. Therefore, there is an urgent need to objectively screen and diagnose individuals with depression, particularly those who face geographical, financial, or practical barriers to accessing traditional psychological or psychiatric services [20].

In response to these pressing needs, ongoing efforts have been made to identify efficient and objective biological, physiological, and behavioral biomarkers for depression. A wide range of biological markers, such as low serotonin levels [21], neurotransmitter dysfunction [22], genetic abnormalities [23], electroencephalogram signals [24,25], eye movements [26], gaits [27], and inflammatory biomarkers [28], have been associated with depression. Nevertheless, a specific biomarker with adequate predictive value for diagnosing depression remains elusive.

With recent advancements in the machine learning field, automatic speech analysis (ASA) has gained popularity as an attractive objective biomarker for depression assessment. ASA presents a variety of benefits that make it particularly attractive for both research and clinical applications. First, it is cost-effective, noninvasive, unobtrusive, and suitable for remote monitoring. In contrast to other biological metrics, speech can be easily collected using modern technologies, such as smartphones, tablets, and computers, which are broadly accessible to the population, thereby eliminating the need for expensive wearables or invasive neuroimaging techniques. Second, ASA may offer greater objectivity compared to self-reported measures. While self-report questionnaires are subject to bias and intentional or unintentional symptom masking, speech features, particularly acoustic features, are more difficult to consciously manipulate. This is because depression can disrupt motor and cognitive processes associated with speech production, resulting in subtle acoustic and linguistic changes that reflect underlying physiological and neurocognitive alterations. Third, speech conveys both what is said and how it is said, providing a dual perspective on the speaker's cognitive and emotional states. While the linguistic content serves as a direct manifestation of these states, variations in motor and acoustic features can indirectly reveal underlying neural activity [20]. Finally, speech may be generalized across different languages due to the shared aspects of vocal anatomy, which allow for the comparison and application of speech-based metrics across diverse linguistic contexts (for a more in-depth review of the advantages of speech as a biomarker, see the study by Low et al [20]).

An increasing number of studies have investigated the potential of speech analysis for depression detection, which has led to the publication of several reviews [14,20,29,30]. However, to date, only Liu et al [30] has conducted a meta-analysis to evaluate the diagnostic performance of ASA to detect depression, and their review was limited to studies using deep learning algorithms. This underscores the need for a comprehensive quantitative synthesis of the evidence that includes both machine learning and deep learning approaches. Therefore, this systematic review and meta-analysis aimed to synthesize published data on the performance of ASA for depression detection.

Methods

Overview

This systematic review and meta-analysis were conducted in accordance with the PRISMA-DTA (Preferred Reporting Items

for Systematic Reviews and Meta-Analyses—Extension for Diagnostic Test Accuracy) [31]. The PRISMA-DAT checklist is provided in Multimedia Appendix 1 [32-136]. We registered the review protocol with the PROSPERO (International Prospective Register of Systematic Reviews; ID: CRD42023444431). There was no prior published protocol for the current study.

Search Strategy

We systematically searched 8 electronic databases on April 1, 2025, including MEDLINE (via Ovid), APA PsycInfo (via Ovid), Embase (via Ovid), CINAHL (via EBSCO), IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. For Google Scholar, only the first 100 results were considered. Our search used the following terms as index terms or free-text words: “speech analysis” and depression. The full electronic search strategy used for MEDLINE was the following: ((“speech analysis”[tw] OR “speech processing”[tw] OR “speech feature”[tw] OR “speech features”[tw] OR “speech signal”[tw] OR “speech signal processing”[tw] OR “speech classification”[tw] OR “speech recognition”[tw] OR “speech model”[tw] OR “acoustic analysis”[tw] OR “acoustic processing”[tw] OR “acoustic feature”[tw] OR “acoustic features”[tw] OR “acoustic signal”[tw] OR “acoustic signal processing”[tw] OR “acoustic classification”[tw] OR “acoustic recognition”[tw] OR “acoustic model”[tw] OR “vocal analysis”[tw] OR “vocal processing”[tw] OR “vocal feature”[tw] OR “vocal features”[tw] OR “vocal signal”[tw] OR “vocal signal processing”[tw] OR “voice recognition”[tw] OR “vocal recognition”[tw] OR “voice model”[tw] OR “vocal model”[tw]) AND (exp Depression/ OR depress*[tw] OR exp Depressive Disorder/ OR depressive disorder*[tw])). The full search strategies for all databases are detailed in Table S12 in Multimedia Appendix 2. The screening process involved 2 stages. First, titles and abstracts were reviewed, followed by a thorough examination of full-texts. The entire selection process was independently conducted by 2 reviewers (PLM and AV), and discrepancies were resolved through dialogue and consensus.

Eligibility Criteria and Study Selection

We limited the inclusion to studies that (1) evaluated the accuracy of ASA to detect depression; (2) reported performance metrics such as accuracy, sensitivity, specificity, precision, or confusion matrices; and (3) were published in English since 2013. We excluded studies focusing on the prediction of depression severity or treatment outcomes. Studies that used multimodal data (eg, face features) in addition to speech data were excluded, except those that reported distinct performance values for each separate model. For the publication type, we included journal papers, conference papers, and thesis dissertations, but excluded reviews, preprints, conference abstracts, posters, protocols, editorials, and comments. We did not apply any constraints in terms of setting, reference standard, or country of publication.

Data Extraction

Data extraction covered study metadata, speech variables, artificial intelligence (AI) algorithms, and performance metrics. From studies providing raw data or confusion matrices, we computed accuracy, sensitivity, specificity, and precision. Many studies have conducted multiple experiments to assess distinct speech features, validation methods, and AI techniques, yielding a range of results. To comprehensively represent this variability, we extracted both the lowest and highest values reported for each performance metric across different algorithms. Data extraction was conducted by PLM, while AV and MTA-C verified the extraction for quality assurance.

Risk of Bias and Applicability Appraisal

Two reviewers (AV and MTA-C) independently used a modified version of the Quality Assessment of Studies of Diagnostic Accuracy-Revised (QUADAS-2) by Abd-Alrazaq et al [137] to evaluate the risk of bias and applicability of the included studies. Similar to the original QUADAS-2, the modified version covered 4 domains: participants, index test (AI algorithms), reference standard (ground truth), and analysis [137]. Disagreements were resolved through either consensus or adjudication by a third reviewer (PLM).

Statistical Analysis

Given that many studies conducted more than 1 experiment and reported multiple performance metrics, there was a potential for these studies to disproportionately influence the meta-analysis compared to those reporting a single result. Moreover, experiments from the same study cannot be regarded as independent, which challenges the independence assumption for effect sizes, underlying traditional meta-analytic techniques [138]. Therefore, to calculate the pooled mean of the highest and lowest accuracy, sensitivity, specificity, and precision, we used a 3-level meta-analysis using restricted maximum likelihood [139]. This approach accounts for 3 distinct sources of variance: population differences between study population effects, population differences between effects of experiments from the same study, and, finally, sampling variance [140]. Anticipating significant heterogeneity, we applied a random-effects model.

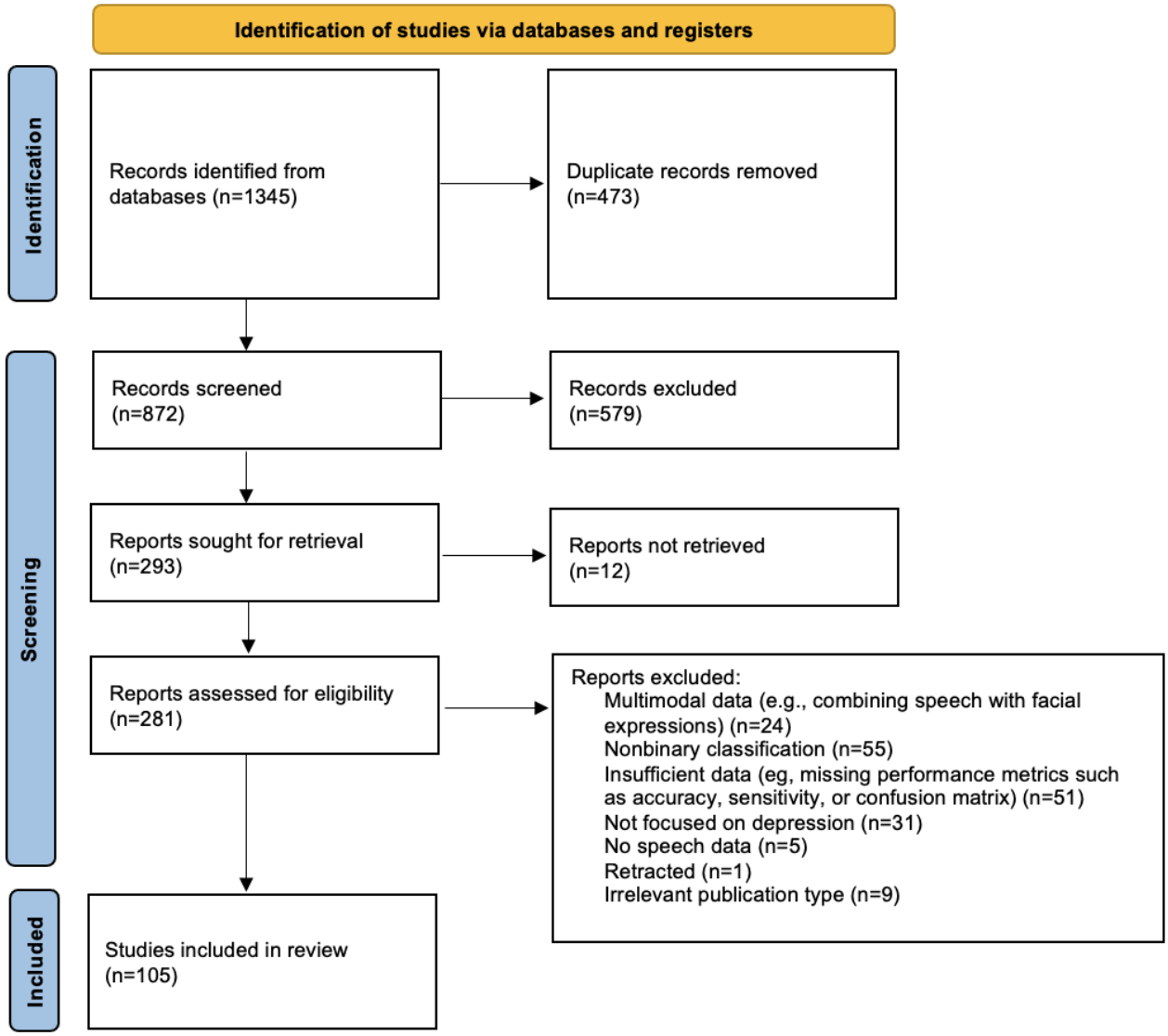
Meta-regression and subgroup analyses were conducted to assess potential variations in ASA performance across different factors, including type of publication, AI algorithms, speech features, speech-eliciting tasks, ground truth assessment, validation approach, dataset, and dataset language, participants’ mean age, and sample size. To ensure the robustness of our findings, we restricted our analyses to subgroups containing 5 or more estimates [141]. To quantify and assess the study heterogeneity, we used the Cochran Q-test and I^2. Cochran Q test determines whether the observed variability exceeds the study’s sampling error alone, and a Cochran P value ≤.05 indicates statistically significant heterogeneity [142]. I^2 statistics estimate the proportion of total variation due to true heterogeneity rather than sampling error, and values of 25%, 50%, and 75% reflect low, moderate, and high heterogeneity, respectively [143]. All analyses were performed using the *metafor* package in R [144,145].

Results

Study Selection

The initial search through 8 databases yielded a total of 1345 studies. After removing 473 duplicates, 872 papers remained. Following titles and abstract screening, a further 579 publications were excluded. Twelve publications were not retrieved. After reading the full text of all the remaining 293 records, a total of 105 studies were ultimately included in the current review. Figure 1 shows the process of the study selection process.

Figure 1. PRISMA flow diagram of the study selection process. This diagram describes the process of identifying, screening, and selecting studies for inclusion. Initially, a total of 1345 records were identified from databases. After the removal of 473 duplicates, 872 records remained for the screening phase. Of these, 579 records were excluded, and 293 reports were sought for retrieval. Further, 12 reports could not be retrieved, resulting in 281 reports assessed for eligibility. Ultimately, 176 reports were excluded based on the predefined inclusion criteria. The final review included 105 studies. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



General Study Characteristics

The 105 studies included in this review were published between 2013 and 2025, with 21 (20%) published in 2022 and 2023. A total of 33 (31.4%) studies were carried out in China, 12 (11.4%) in India, and 10 (9.5%) in Hungary. Regarding the publication type, the majority (55/105, 52.4%) were conference papers, followed by peer-reviewed journal papers (48/105, 45.7%), and theses (2/105, 1.9%). The number of participants included was reported in 98 studies, with sample sizes ranging from 14 to 9337. The age of participants, reported in 34 studies, ranged between 19.0 and 71.7 years, with an average age of 37.47 (SD 10.41) years. The percentage of female participants, reported in 53 studies, varied between 30.8% and 100%. The percentage of depressed participants, as reported in 57 studies, ranged between 21% and 69%, with an average of 45.9% (SD 11.9). Table 1 and Table S2 in Multimedia Appendix 2 detail the specific characteristics of the included studies.

Table 1. Characteristics of the included studies.

Feature	Values	References
Year of publication		
2025	4 (3.8)[a]	[32-35]
2024	13 (12.4)[a]	[36-48]
2023	21 (20)[a]	[49-69]
2022	21 (20)[a]	[70-90]
2021	13 (12.4)[a]	[91-103]
2020	6 (5.7)[a]	[104-109]
2019	6 (5.7)[a]	[110-115]
2018	4 (3.8)[a]	[116-119]
2017	7 (6.7)[a]	[120-126]
2016	5 (4.8)[a]	[127-131]
2015	1 (1)[a]	[132]
2014	2 (1.9)[a]	[133,134]
2013	2 (1.9)[a]	[135,136]
Country of publication		
China	33 (31.4)[a]	[32,37,44,45,48-51,54,58,60,64,65,69-75,78,79,83,91,96,102,103,118,120-122,130,136]
India	12 (11.4)[a]	[42,52,55-57,66,81,94,104,105,114,123]
Hungary	10 (9.5)[a]	[86,88,93,98,99,107,116,124-126]
Australia	6 (5.7)[a]	[111,112,117,131,132,135]
United States	6 (5.7)[a]	[38,40,89,119,129,133]
Malaysia	5 (4.8)[a]	[84,87,101,115,127]
Republic of Korea	4 (3.8)[a]	[34,47,59,97]
United Kingdom	4 (3.8)[a]	[35,43,53,109]
France	3 (2.9)[a]	[80,95,106]
Germany	3 (2.9)[a]	[41,67,76]
Turkey	3 (2.9)[a]	[39,63,108]
Canada	2 (1.9)[a]	[63,128]
Iran	2 (1.9)[a]	[62,77]
Japan	2 (1.9)[a]	[36,82]
Others (Brazil, Denmark, Indonesia, Italy, Philippines, Romania, Saudi Arabia, Singapore, Spain, and Thailand)	1 (each; 0.95)[a]	[33,61,68,85,90,92,100,110,113,134]
Type of publication		
Conference paper	55 (52.4)[a]	[33, 35, 36, 42, 43, 45-47, 50, 52, 55-57, 61, 62, 64-66, 73-75, 83, 87-90, 92-95, 98, 101, 104, 105, 107, 110, 112-116, 119, 121-125, 127-130, 132-136]
Journal paper	48 (45.7)[a]	[32, 34, 37-41, 44, 48, 49, 51, 53, 54, 58-60, 63, 65, 67-72, 76-82, 84-86, 91, 96, 97, 99, 100, 102, 103, 106, 108, 111, 117, 118, 120, 126]

Feature	Values	References
Thesis	2 (1.9)[a]	[109,131]
Number of participants		
Mean (SD)	445.14 (1361.12)	[32-51,53-97,99-101,103,104,107-129,131-133,135,136]
Range	14-9337	[32-51,53-97,99-101,103,104,107-129,131-133,135,136]
Age of participants		
Mean (SD)	37.47 (10.41)	[32,39,41,44,45,48,53,54,58-60,63,67,69-72,74,76,79,80,82-86,96,97,100,107,110,116,122,127,128]
Range	19.0-71.7	[49,51,58,60,64,66,68,70-97,99-101,103,104,107-129,131-133,135,136]
Gender (Female, %)		
Mean (SD)	60.8 (15.6)	[35,36,39,41,42,44,45,48,53,54,56-59,61,63,67,68,76,79,80,82-89,96,97,100,112,118-122,125-129,133]
Range	30.8-1.00	[35,36,39,41,42,44,45,48,53,54,56-59,61,63,67,68,76,79,80,82-89,96,97,100,112,118-122,125-129,133]
Depressed participants (%)		
Mean (SD)	45.9 (11.9)	[33-35, 38-45, 47-51, 53-59, 63-65, 67-76, 78-84, 87-90, 92-94, 96, 99, 100, 103, 104, 107, 110, 112, 116-122, 124, 125, 127-129, 131-133, 135, 136]
Range	21.0-69.0	[33-35, 38-45, 47-51, 53-59, 63-65, 67-76, 78-84, 87-90, 92-94, 96, 99, 100, 103, 104, 107, 110, 112, 116-122, 124, 125, 127-129, 131-133, 135, 136]

[a]Studies, n (%).

Features of ASA Classifiers

In terms of speech features, spectral features (91/105, 86.7%) were the most frequently analyzed, followed by prosodic features (58/105, 55.2%), source features (53/105, 50.5%), format features (39/105, 37.1%), and lexical features (15/105, 14.3%). The most commonly used speech-eliciting task was free speech (76/105, 72.4%). Reading tasks were used in 36.2% (38/105) of studies, while only a few used counting (3/105, 2.9%) or sustained vowels (2/105, 1.9%). Many studies conducted multiple experiments to evaluate various AI algorithms, with support vector machine (SVM) being the most frequently used (43/105, 41%). The most commonly used assessment instruments to establish ground truth were the Patient Health Questionnaire-8/-9 (51/105, 48.6%). The included studies applied 4 different validation methods, the most common of which were hold-out cross-validation (64/105, 61%) and K-fold cross-validation (38/105, 36.2%). More than half of the studies used hand-crafted datasets (56/105, 53.3%), while the DAIC-WOZ (Distress Analysis Interview Corpus–Wizard of Oz) was the most frequently used public database (35/105, 33.3%). The specific details of the models used in the included studies are detailed in Table 2 and Table S2 in Multimedia Appendix 2.

Table 2. Features of automatic speech analysis classifiers.

Feature	Studies, n (%)
Speech features	
Spectral features	91 (86.7)
Prosodic features	58 (55.2)
Source features	53 (50.5)
Format features	39 (37.1)
Lexical features	15 (14.3)
TEO[a]	8 (7.6)
Spectogram	6 (5.7)
Speech-eliciting tasks	
Free speech	76 (72.4)
Reading	38 (36.2)
Counting	3 (2.9)
Sustained vowels	2 (1.9)
Not reported	10 (9.5)
AI[b] algorithms	
Support vector machine	43 (41.0)
Convolutional neural network	15 (14.3)
Logistic regression	14 (13.3)
Random forest	11 (10.5)
Deep neural network	10 (9.5)
Gaussian mixture models	9 (8.6)
Ensemble model	7 (6.7)
K-nearest neighbors	7 (6.7)
Multilayer perceptron	7 (6.7)
Naïve bayes	7 (6.7)
AdaBoost decision tree	3 (2.9)
Artificial neural network	2 (1.9)
Linear discriminant analysis	2 (1.9)
Long-term and short-term memory	3 (2.9)
Recurrent neural network	2 (1.9)
Support vector machine + Gaussian mixture model	2 (1.9)
Others	28 (26.7)
Ground truth assessment	
PHQ-8[c] and PHQ-9[d]	51 (48.6)
BDI[e] and BDI-II[f]	23 (21.9)
HAM-D[g]	10 (9.5)
DSM[h] and *DSM-IV*[i]	10 (9.5)
CIDI[j]	4 (3.8)
MINI[k]	3 (2.9)
Others	9 (8.6)

Feature	Studies, n (%)
Not reported	8 (7.6)
Validation approach	
Hold-out cross-validation	64 (61.0)
K-fold cross-validation	38 (36.2)
Leave-one-out cross-validation	18 (17.1)
Nested cross-validation	2 (1.9)
Not reported	4 (3.8)
Dataset	
Handcrafted	56 (53.3)
DAIC-WOZ[l]	35 (33.3)
AVEC-2013[m], AVEC-2014, AVEC-2017, and AVEC-2019	10 (9.5)
MODMA[n]	10 (9.5)
CONVERGE[o]	3 (2.9)
DEPISDA[p]	3 (2.9)
ORI-DB[q]	2 (1.9)
Others	8 (7.6)

[a]TEO: Teager Energy Operator.
[b]AI: artificial intelligence.
[c]PHQ-8: Patient Health Questionnaire-8.
[d]PHQ-9: Patient Health Questionnaire-9.
[e]BDI: Beck Depression Inventory.
[f]BDI-II: Beck Depression Inventory-II.
[g]HAM-D: Hamilton Depression Rating Scale.
hDSM: Diagnostic and Statistical Manual of Mental Disorders.
iDSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.
[j]CIDI: Composite International Diagnostic Interview.
[k]MINI: Mini-International Neuropsychiatric Interview.
[l]DAIC-WOZ: Distress Analysis Interview Corpus Wizard-of-Oz.
[m]AVEC: Audio or Visual Emotion Challenge.
[n]MODMA: Multimodal Open Dataset for Mental Disorder Analysis.
[o]CONVERGE: China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology.
[p]DEPISDA: Hungarian Depressed Speech Database.
[q]ORI-DB: Oregon Research Institute Database.

Results of Risk of Bias Appraisal

Almost half of the studies (50/105, 47.6%) were classified as high risk in at least 1 domain. Most of the studies (85/105, 81%) used an appropriate consecutive or random sampling method to select eligible participants. A significant portion of the studies (64/105, 61%) avoided inappropriate exclusions, and a balanced representation of depressed versus nondepressed subgroups was maintained in 60% (63/105) of them. Additionally, 66.7% (70/105) included a sufficient sample size. Thus, in the participant domain, almost half (49/105, 46.7%) of the studies had unclear or high risk of bias.

All the studies (105/105, 100%) thoroughly described the AI models used. Additionally, the predictive features (ie, speech features) were clearly outlined in 94.3% (99/105) of the cases, and 93.3% (98/105) of the studies assessed these features consistently for all participants. Nonetheless, in 74.3% (78/105) of the studies, information was insufficient to confirm that feature collection was blinded to outcome data. Risk of bias related to the index test was deemed low in 94.3% (99/105) of the studies.

Most of the studies (96/105, 91.4%) used appropriate reference standards for classifying the outcome (ie, depressed vs nondepressed). Consistent outcome definition was maintained for all participants in 87.6% (92/105) of the studies, and 86.7% (91/105) of the studies determined the outcome without predictor information. However, more than three-quarters (85/105, 81%) of the studies did not provide sufficient data to ensure that an appropriate interval was maintained between the index test and

the reference standard. Accordingly, the risk of bias due to the reference standard was low in 81% (85/105) of the studies.

Only 21.9% (23/76) of the studies included all enrolled participants in the data analysis. In 93.3% (102/105) of the studies, there was insufficient information to confirm proper data preprocessing. The breakdown of the training, validation, and test sets was adequate in 91.4% (96/105) of the studies, and model performance was evaluated using appropriate metrics in 81.9% (86/105) of the studies. Consequently, 43.8% (46/105) of the studies were considered to have a low risk of bias in the analysis domain.

Regarding the evaluation of model applicability, most of the studies, 5 (97/105, 92.4%), were identified as having a low concern of applicability in the participants' domain, and almost all studies (102/105, 97.1%) had a low concern of applicability in the index test domain. In the outcome domain, 92.4% (97/105) of the studies were adjudged to have low concerns regarding outcome definition, timing, or outcome determination. Figure S1 in Multimedia Appendix 2 illustrates the overall risk of bias and applicability concerns assessment. Table S3 in Multimedia Appendix 2 details reviewers' assessment of each study included.

Meta-Analysis of Included Models

Overview

A 3-level meta-analysis summarized the highest and lowest results for accuracy, sensitivity, specificity, and precision. Meta-regressions and subgroup analyses were conducted to examine potential differences in the ASA performance across various factors, including type of publication, AI algorithms, speech features, speech-eliciting tasks, ground truth assessment, validation approach, dataset, dataset language, participants' mean age, and sample size. All meta-regression and subgroup analyses results are shown in Tables S4-S11 in Multimedia Appendix 2.

Accuracy

Accuracy was reported in 86 studies, comprising 148 estimates (N=27,039). The highest accuracy ranged from 0.29 to 0.99, with a pooled mean of 0.81 (95% CI 0.79 to 0.83; Figure 2). There was significant heterogeneity among studies (Cochran $P<.001$; $I^2=96.74\%$). The lowest accuracy estimates, derived from 114 estimates in 65 studies (N=16,394), ranged between 0.23 and 0.94. The pooled mean for lowest accuracy was 0.66 (95% CI 0.63 to 0.69; Figure 3), also showing considerable heterogeneity (Cochran $P<.001$; $I^2=94.44\%$). Meta-regression and subgroup analyses revealed statistically significant differences in the highest accuracy for speech features (Cochran $P=.04$) and algorithms (Cochran $P=.04$) groups (Table S4 in Multimedia Appendix 2). No other statistically significant differences were observed.

Figure 2. Three-level forest plot of the highest accuracy estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 86 studies, comprising 148 estimates [32-34,36,37,39-49,51,53-68,71-74,77-87,89-97,99-101,104,107,109-126,128,129,131-133,136]. The solid squares represent point estimates of the highest accuracy, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the pooled highest accuracy estimates.

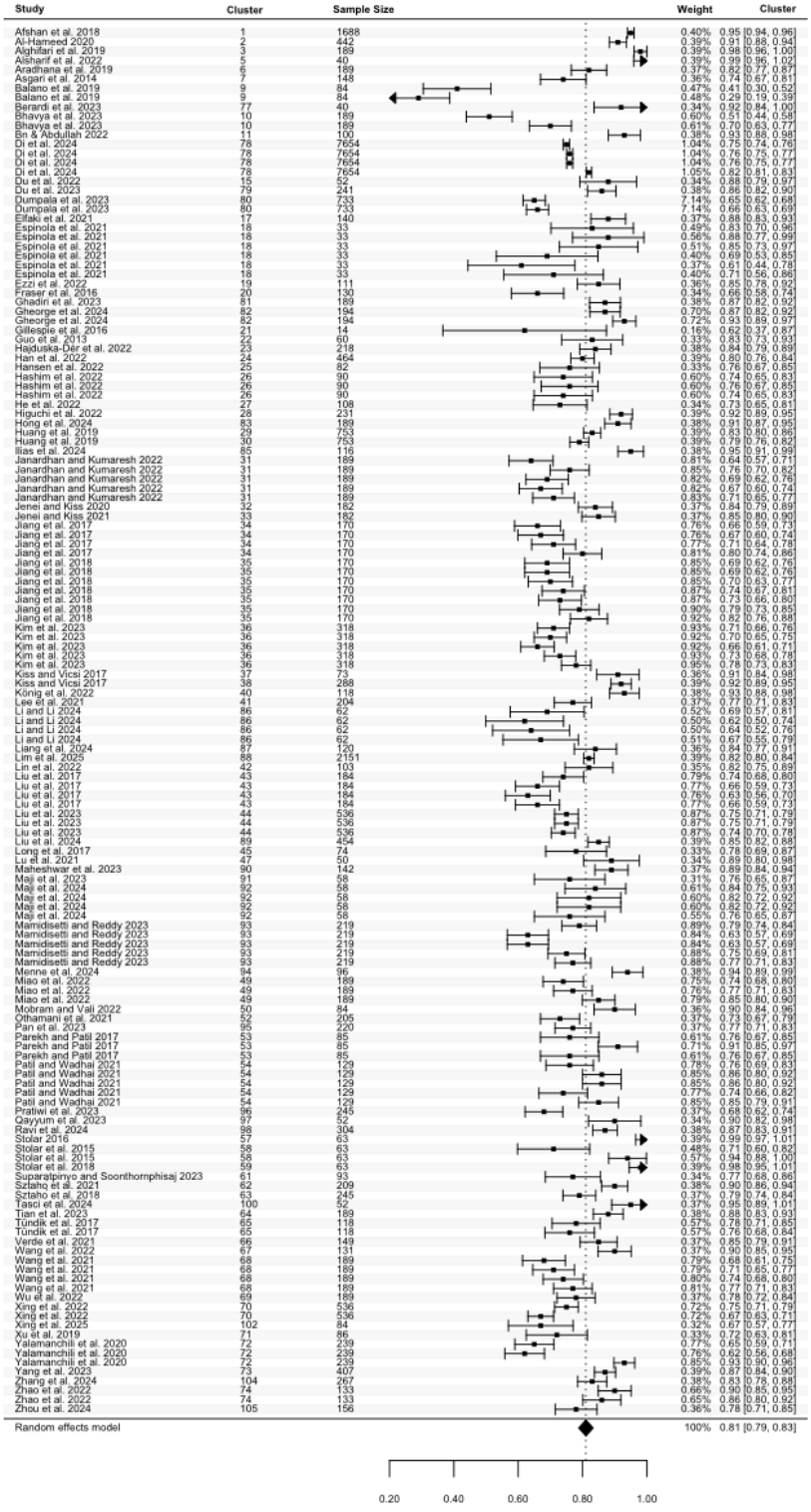
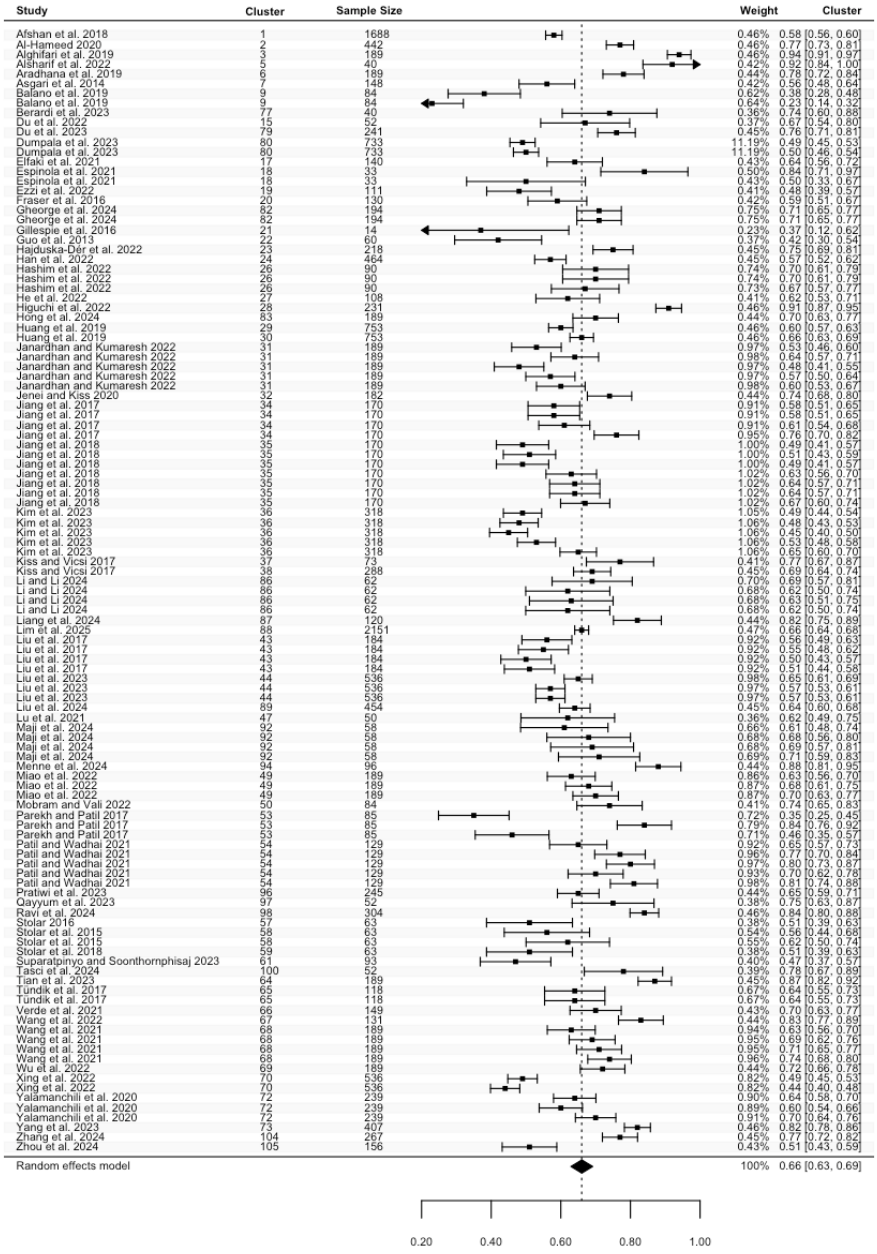


Figure 3. Three-level forest plot of the lowest accuracy estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 114 estimates of the lowest accuracy, reported in 65 studies [33, 34, 36, 37, 39-45, 47, 49, 51, 53, 58-61, 63-65, 67, 68, 72-74, 77, 78, 81-84, 86, 87, 90-92, 94, 96, 100, 101, 104, 107, 109, 111-115, 117-121, 123-126, 128, 129, 131-133, 136]. The solid squares represent point estimates of accuracy, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the pooled lowest accuracy estimates.



Sensitivity

Sensitivity data were reported in 81 studies, including 135 estimates (N=36,096). The highest sensitivity ranged from 0.25 to 1.00, with a pooled mean of 0.84 (95% CI 0.81 to 0.86; Figure 4), and considerable heterogeneity (Cochran P<.001; I^2=99.93%). For the lowest sensitivity, 105 estimates from 64 studies (N=25,913) ranged between 0.00 and 0.98. The pooled mean was 0.63 (95% CI 0.58 to 0.68; Figure 5), with significant heterogeneity (Cochran P<.001; I^2=99.93%). Meta-regression and subgroup analyses revealed no statistically significant differences in sensitivity across groups except for speech features subgroups in the highest sensitivity (P=.05; Table S6 in Multimedia Appendix 2).

Figure 4. Three-level forest plot of the highest sensitivity estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 135 estimates of the highest sensitivity, reported in 81 studies [34-51, 53-62, 64-71, 73-86, 88, 90, 92, 93, 95-97, 99-101, 103, 104, 108-110, 112, 113, 116-120, 123, 125-129, 131, 132, 135]. The solid squares represent point estimates of sensitivity, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled highest sensitivity.

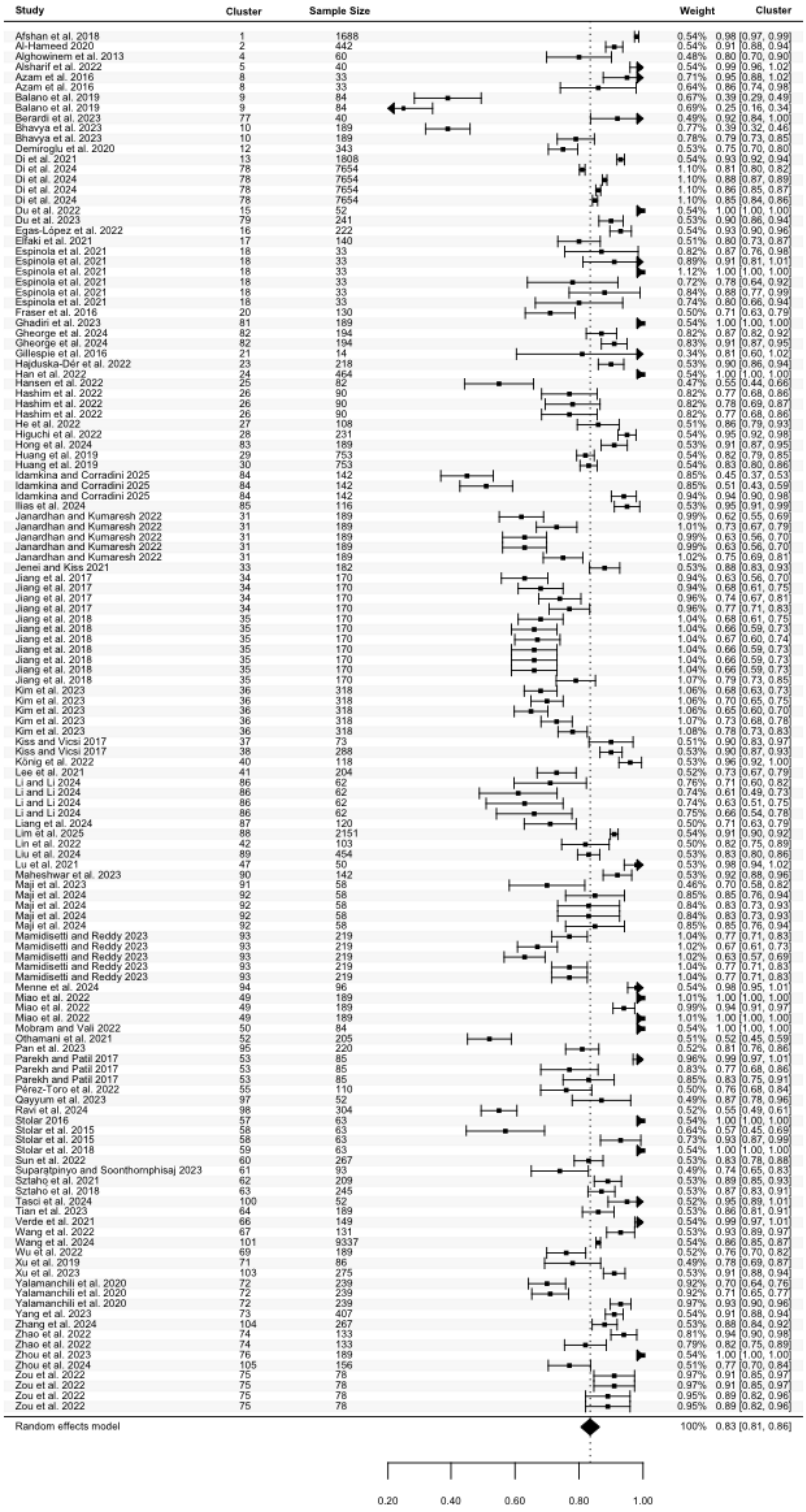
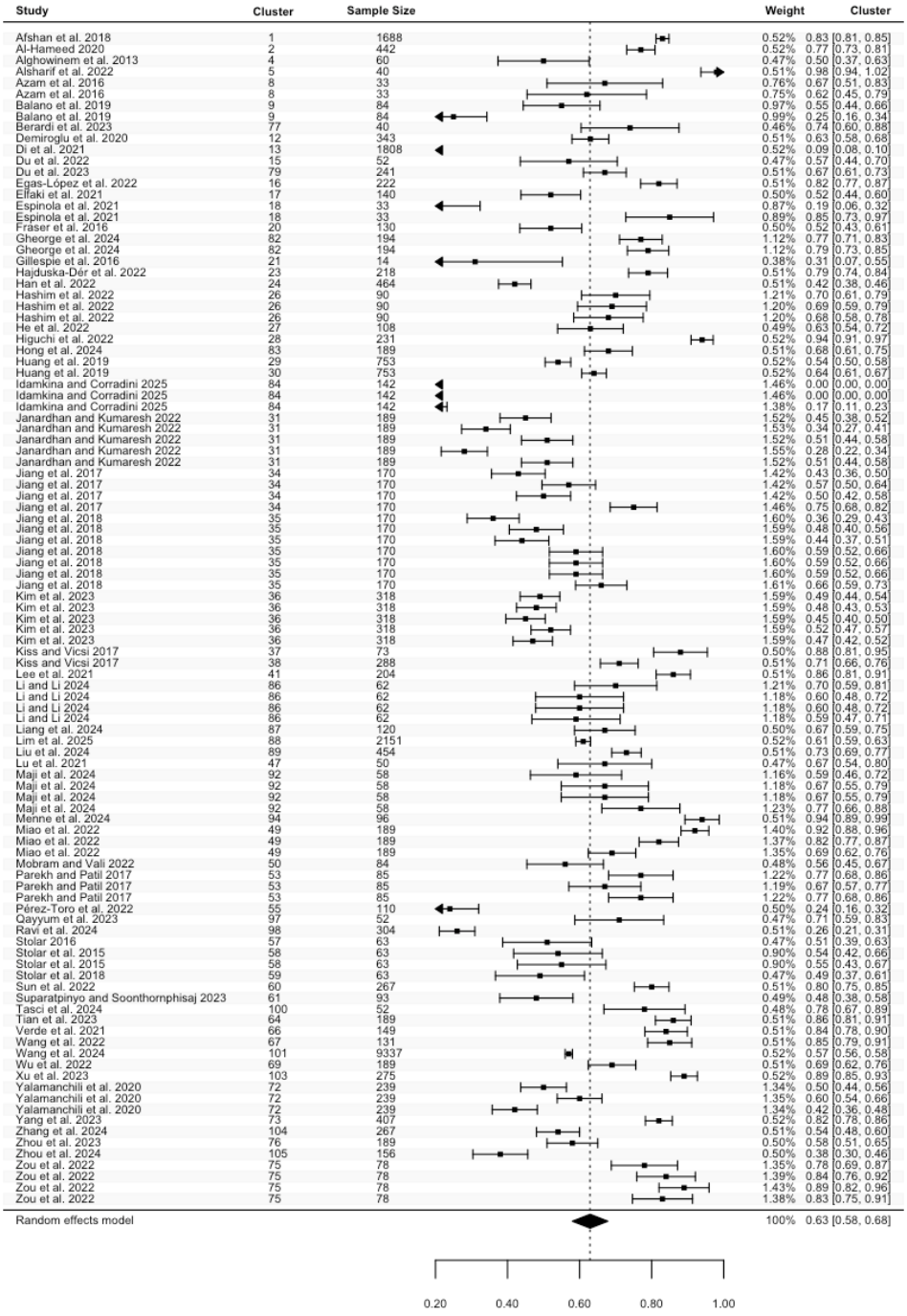


Figure 5. Three-level forest plot of the lowest sensitivity estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 105 estimates of the lowest sensitivity, from 64 studies [34-45, 47, 49-51, 53, 55, 59-61, 64, 65, 67-70, 73-78, 81-84, 86, 88, 90, 92, 96, 97, 100, 101, 103, 104, 108, 109, 112, 113, 117-120, 123, 125-129, 131, 132, 135]. The solid squares represent point estimates of sensitivity, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled lowest sensitivity.



Specificity

Specificity was reported in 47 studies, with a total of 77 estimates (N=20,207). The highest specificity ranged from 0.33 to 1.00, with a pooled mean of 0.83 (95% CI 0.79 to 0.86; Figure 6), and high heterogeneity (Cochran P<.001; I^2=99.81%). The lowest specificity estimates, from 55 estimates in 34 studies (N=10,553), ranged between 0.03 and 0.94. The pooled mean was 0.60 (95% CI 0.55 to 0.66; Figure 7), with significant heterogeneity (Cochran P<.001; I^2=97.81%). Meta-regression and subgroup analyses indicated no statistically significant differences in specificity across groups except for speech features subgroups in the highest specificity (P=.004; Table S8 in Multimedia Appendix 2).

Figure 6. Three-level forest plot of the highest specificity estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 77 estimates of the highest specificity, from 47 studies [34, 36, 37, 40, 41, 43-46, 48, 49, 51, 54, 65, 66, 68, 76, 79-82, 84, 86, 88, 92, 93, 95, 97, 99, 100, 103, 104, 109-112, 116-118, 120, 123, 125-128, 131, 132]. The solid squares represent point estimates of specificity, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled highest specificity.

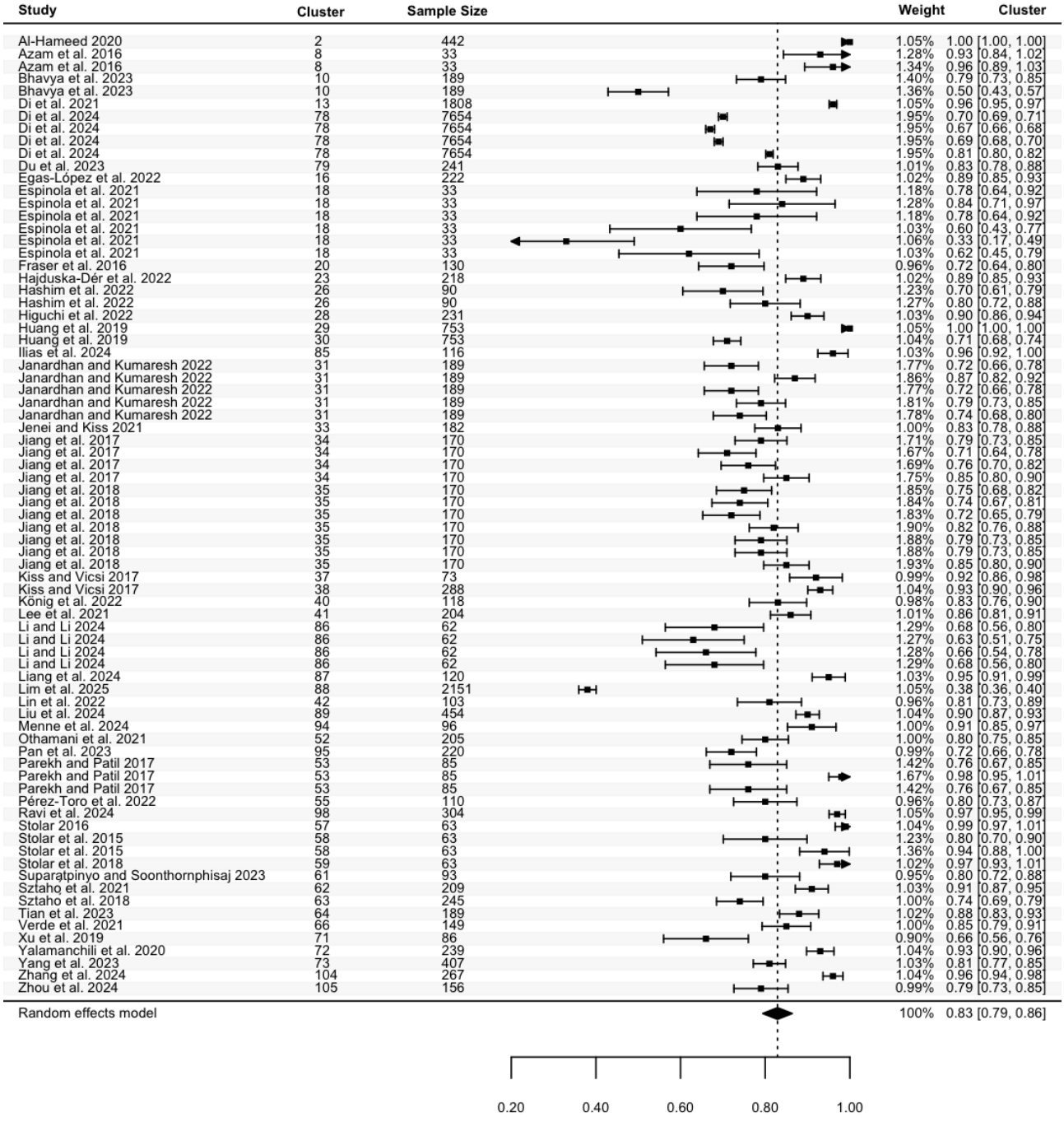
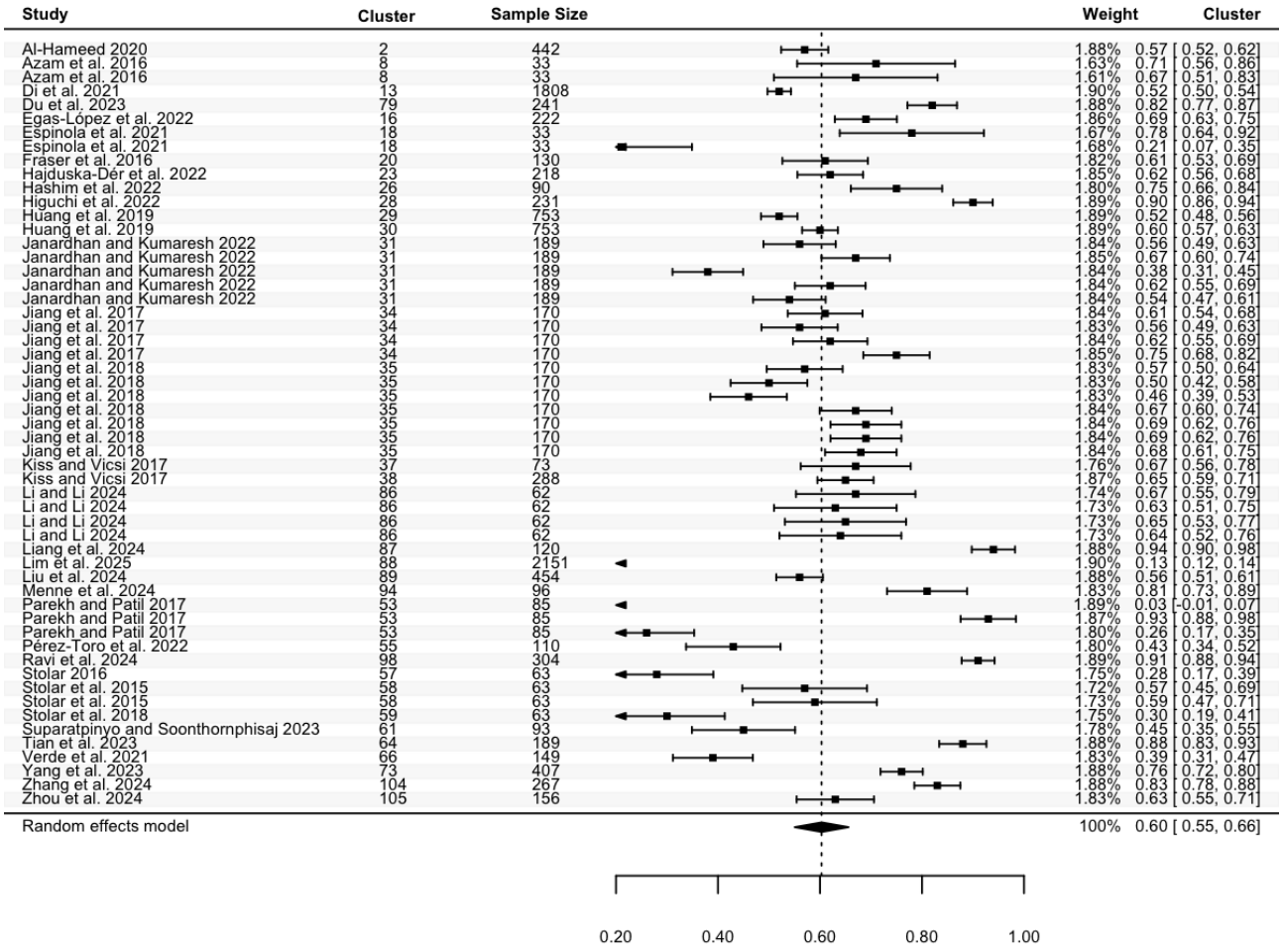


Figure 7. Three-level forest plot of the lowest specificity estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 55 estimates of the lowest specificity, from 34 studies [34, 36, 37, 40, 41, 43-45, 49, 51, 65, 68, 76, 81, 82, 84, 86, 88, 92, 100, 103, 109, 111, 112, 117, 118, 120, 123, 125-128, 131, 132]. The solid squares represent point estimates of specificity, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled lowest specificity.



Precision

Precision was reported in 62 studies, with 95 estimates for the highest precision (N=24,696). The highest precision in these studies ranged from 0.25 to 1.00, with a pooled mean of 0.81 (95% CI 0.77 to 0.84; Figure 8), and considerable heterogeneity (Cochran P<.001; I^2=99.81%). For the lowest precision, 73 estimates from 46 studies (N=22,215) ranged between 0.00 and 0.98. The pooled mean was 0.64 (95% CI 0.58 to 0.70; Figure 9), with considerable heterogeneity (Cochran P<.001; I^2=99.81%). No statistically significant differences in precision were identified across groups.

Figure 8. Three-level forest plot of the highest precision estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 95 estimates of the highest precision, reported in 62 studies [34, 35, 37-40, 42-44, 46, 47, 49-51, 53-57, 59-62, 64-71, 74, 75, 77-85, 90-93, 95, 96, 99, 101, 104, 108-113, 116, 119, 125, 126, 129]. The solid squares represent point estimates of precision, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled highest precision.

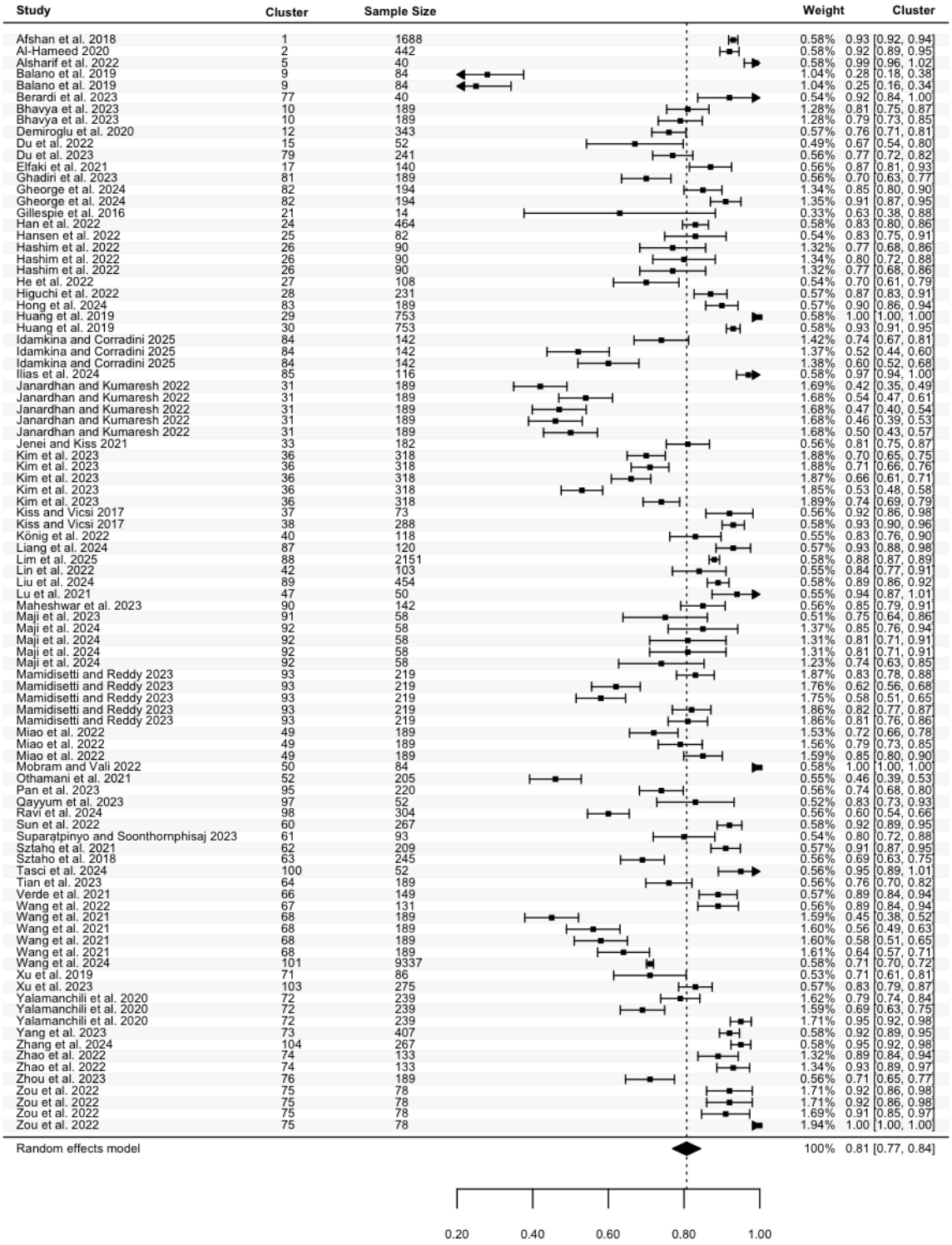
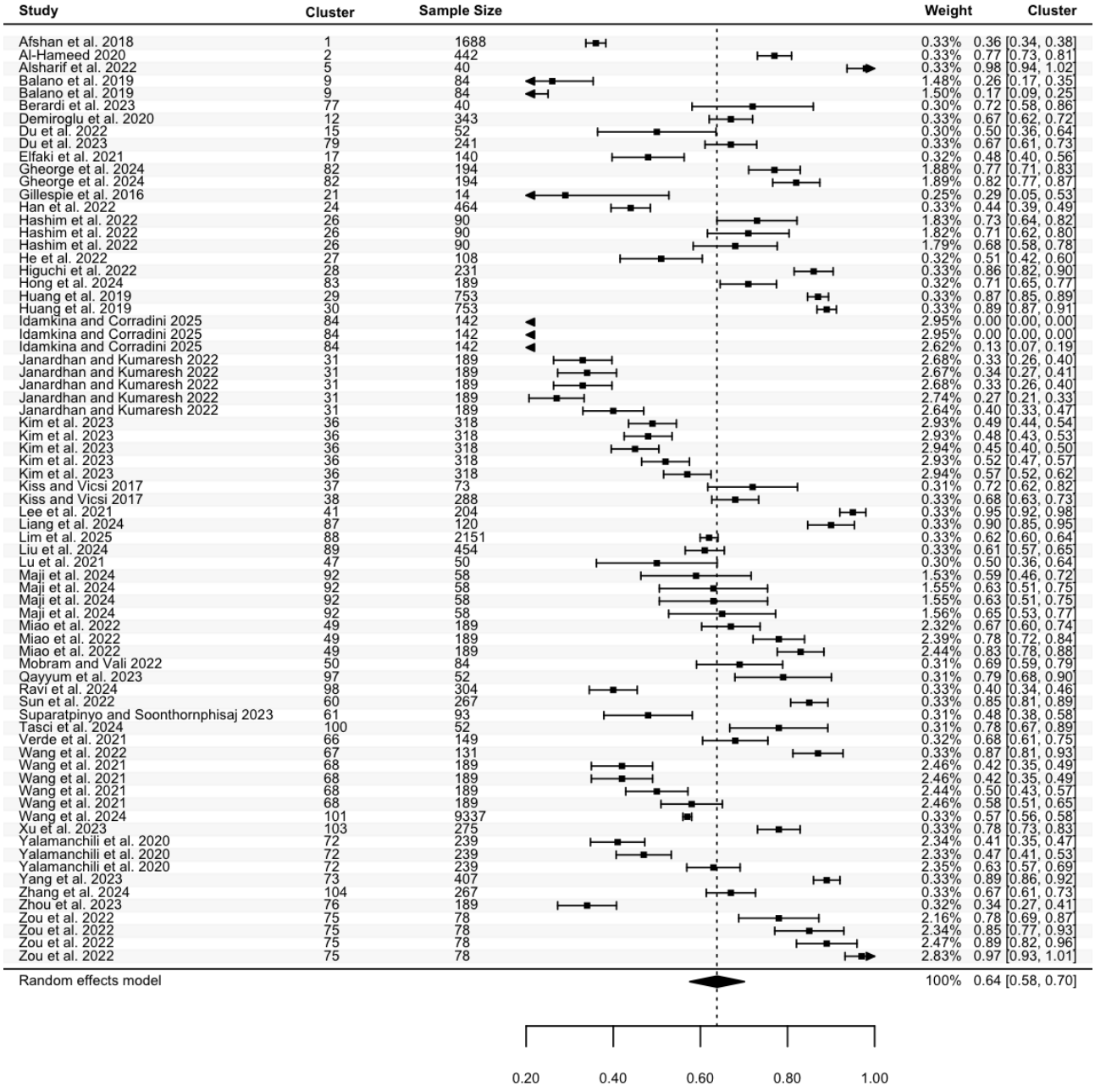


Figure 9. Three-level forest plot of the lowest precision estimates. This forest plot illustrates the results of this 3-level meta-analysis based on 73 estimates of the lowest precision, reported in 46 studies [34, 35, 37-40, 42-44, 47, 49, 50, 53, 59-61, 64, 65, 67-70, 74, 75, 77, 78, 81-84, 90-92, 96, 97, 101, 104, 108, 109, 111-113, 119, 125, 126, 129]. The solid squares represent point estimates of precision, with horizontal lines indicating the 95% CIs. The rhombus at the bottom represents the estimated pooled lowest precision.



Discussion

Principal Findings

To the best of our knowledge, this is the first meta-analysis aimed at assessing the performance of ASA in detecting depression, including both machine learning and deep learning algorithms. Pooled data from 105 studies reveal that ASA demonstrates good, although not optimal, performance in classifying depression. Given that multiple studies conducted multiple experiments, we calculated the pooled mean of both the lowest and highest accuracy, sensitivity, specificity, and precision. Our results indicate that across studies, the pooled accuracy for depression detection was between 65% (ie, pooled mean of the lowest accuracy) and 81% (pooled mean of the highest accuracy), with comparable sensitivity (63%-84%) and specificity (60%-83%). This suggests that the ability of ASA to detect individuals with depression (sensitivity) was relatively consistent with its ability to identify those without depression (specificity). Additionally, pooled precision, which reflects the ability of ASA to correctly classify individuals who truly have depression among all those classified as depressed, ranged between 64% and 81%.

Our results revealed significant heterogeneity among studies. Meta-regression and subgroup analyses indicated a statistically

significant difference between speech features, with Teager Energy Operator (TEO)–based features outperforming others in the pooled mean of the highest accuracy, sensitivity, and specificity. TEO-based features are based on the studies conducted by Teager and Teager [146], which demonstrated that airflow propagation in the vocal tract has a nonlinear character [146]. Moreover, convolutional neural network and deep neural network (DNN), in general, outperformed other algorithms in highest accuracy, while naïve bayes performed the worst. DNNs are often regarded as "black boxes" because their decision-making processes are not easily interpretable [147], which hinders clinicians' trust and willingness to incorporate these models into the routine clinical workflow [148]. To foster trust in AI solutions for the detection of psychiatric conditions, it is important to implement a participatory approach, involving clinicians and other specialists in the model design process to create more interpretable algorithms that can better align with clinical needs and standards, thereby facilitating the incorporation of these potential solutions into clinical practice. Nevertheless, these results should be interpreted with caution as most studies in the current meta-analysis had a small sample size, which might have obscured further potential differences between AI algorithms. ASA performance is likely influenced by a complex interplay of factors that were not fully captured by the subgroup analyses. Beyond the measured variables, cultural and linguistic diversity, confounding factors such as comorbid conditions (eg, fatigue and anxiety), lifestyle influences, or medication effects, as well as differences in study protocols and dataset characteristics, may all contribute to the observed heterogeneity. These findings highlight the urgent need for standardized methodologies, diverse and inclusive datasets, and further research to understand and address these sources of heterogeneity.

Comparison With Prior Work

In the context of AI diagnostic efficacy for depression detection, a recent meta-analysis conducted by Liu et al [30] assessing the diagnostic performance of deep learning algorithms in detecting depression through speech reported a superior accuracy of 0.87 (ie, highest accuracy). Notably, their analysis was restricted to peer-reviewed journal papers that included confusion matrices. In contrast, our review applied broader inclusion criteria, including conference papers, journal papers, and theses, and incorporated studies even if they reported a single performance metric. It is also important to note that their analyses were based on only 8 studies and exclusively focused on deep learning algorithms, whereas our meta-analysis included both machine learning and deep learning approaches. Similarly, Abd-Alrazaq et al [137] performed a meta-analysis focused on AI performance in detecting depression using wearable devices. Based on 38 studies, they reported superior accuracy (70%-93%), superior specificity (73%-93%), and slightly better sensitivity (61%-87%). Notably, their meta-analysis included studies using wearable devices that monitored a range of parameters, including physical activity, sleep patterns, and heart rate. Given the complex and multifaceted clinical profile of depression, integrating ASA into complex statistical models alongside other data sources, such as biological (eg, genetic, inflammatory, and neuroimaging data), psychological (eg, psychometric scales), and clinical variables, could significantly improve the accuracy and reliability of AI tools for depression detection. Another promising avenue is the use of facial expression analysis in combination with speech. However, this approach may raise additional ethical concerns, particularly regarding data privacy, consent, and the risk of algorithmic biases [149].

SVM was the most used algorithm to classify individuals with depression in the present meta-analysis. SVMs are highly regarded in machine learning for their robust ability to handle noisy, interrelated features and process datasets with high-dimensionality efficiently. Convolutional neural network and DNN, in general, were also commonly used. Our findings, in line with those of Liu et al [30], suggest that these architectures hold a significant potential in ASA. Therefore, we recommend that future research efforts go in this direction, particularly in combination with participatory methodologies that involve clinicians throughout the development process. This collaborative approach is essential to ensure the clinical relevance, feasibility, and eventual adoption of such algorithms in mental health care settings. Nonetheless, several challenges remain, such as the large sample sizes typically required to train deep learning models effectively and the considerable computational power needed to support their development and implementation.

DAIC-WOZ [150] was the most commonly used open-source dataset, which includes 189 clinical interviews conducted by Ellie, a virtual interviewer [150]. Although this dataset has facilitated significant advancements in the field, it presents several limitations that warrant attention. First, the participants were volunteers whose depressive symptoms were assessed using the Patient Health Questionnaire-8, rather than through formal clinical diagnosis. Second, the metadata associated with DAIC-WOZ and other datasets is sparse, leaving potential confounding factors unspecified [151]. Third, the dataset endures a significant class imbalance, with nondepressed participants outnumbering depressed ones by a ratio of approximately 4:1 [152]. Furthermore, the dataset exhibits a notable gender bias in depression prevalence, with females having a higher proportion of depressed patients (ratio of approximately 5:8 depressed to nondepressed) compared to males (ratio of about 2:7 depressed to nondepressed) [153]. This imbalance raises concerns about biased model training, as machine learning algorithms may overfit to the majority class [154], thereby reducing their ability to generalize effectively across diverse populations [155]. One promising direction for future research is to integrate principles of fair machine learning [156] as well as prior domain knowledge, such as gender-specific linguistic patterns or balanced sampling strategies, into the design of depression detection models, ensuring that algorithms account for gender and class imbalances while avoiding overfitting to unintended features [154].

In terms of speech features used, most studies included in this review focused exclusively on acoustic features, while only 15 studies incorporated lexical features. Linguistic studies have shown that depression also manifests in language use, specifically in semantic and syntactic patterns that reveal heightened self-focus and pervasive negative affect [157].

Recent advances in deep learning, particularly the development of transformer-based models, such as BERT (Bidirectional Encoder Representations From Transformers) and RoBERTa (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach), have shown significant promise in natural language and speech processing tasks, with some studies reporting accuracy rates of up to 98% in detecting depression [158-160]. BERT is a DNN model that generates bidirectional text representations while preserving contextual and semantic nuances, making it particularly suitable for analyzing language associated with mental health. However, most BERT-based studies to date have focused on text from social media platforms. Future research should explore the integration of such models with speech data to examine whether combining what is said (linguistic content) with how it is said (acoustic features) can offer a more comprehensive understanding of an individual's mental health.

Research and Practical Implications

This review highlights that ASA is an emerging technology with potential for detecting depression, though its readiness for clinical application is still limited. First, the ASA performance is not yet optimal, indicating room for improvement. Second, most included studies had small or modest sample sizes; for robust and reliable estimates, larger samples are required [161]. This is particularly critical as small training sets may lead to inflated accuracy estimates due to overfitting or random effects [162,163]. Third, more than half of the studies assessed had a high risk of bias in at least 1 domain. Fourth, there is a notable absence of large-scale studies that examine the generalizability of ASA across countries, settings, and depression severity levels. Moreover, the clinical and methodological variables that affect this generalization remain unclear. Fifth, more than half of the studies (52.4%) were conference papers, so more peer-reviewed, high-quality studies are warranted. Sixth, considering the diversity in speaker characteristics and individual speaking styles, it is imperative to conduct longitudinal studies to discern whether variations in speech patterns are indicative of depression symptoms or merely reflect inherent personal differences or related conditions such as increased anxiety or fatigue [14,29]. Such studies are essential for validating the diagnostic precision of ASA and ensuring its reliability across different individuals and contexts. Finally, most of the studies included in this review used self-reported questionnaires for depressive symptoms and applied commonly used cutoff scores to indicate the presence of depression. Only a small percentage used structured clinical interviews, such as the *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders* [Fourth Edition]) criteria, the Composite International Diagnostic Interview, or the Mini-International Neuropsychiatric Interview, to establish a formal diagnosis. This methodological discrepancy has important implications for interpreting the findings. Self-report measures primarily capture symptom severity, and while useful for screening purposes, they do not equate to a clinical diagnosis of depression. As such, when relying on these instruments, the objective of the study shifts from predicting a clinical diagnosis to predicting questionnaire scores, which may not fully align with diagnostic criteria [103]. To determine whether ASA could serve as a viable method for diagnosing depression, future studies should incorporate clinical data based on a formal diagnosis of depression.

Early detection is paramount, particularly for conditions such as depression, where delays in diagnosis can significantly worsen outcomes [164,165]. Although ASA offers a potential cost-effective, clinically relevant solution that could make depression screening more feasible and accessible, these methods should still be considered complementary tools rather than a replacement for established diagnostic methods. Moreover, ensuring its equitable, effective, and sustainable deployment requires addressing critical ethical and implementation challenges. The Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability framework provides valuable guidance in this effort, offering an evidence-based approach for studying the nonadoption and abandonment of technologies by individuals and the challenges to scale-up, spread, and sustainability of such technologies in health care organizations and systems [166]. Specifically, future studies should explore how ASA interacts with organizational capacity, user adoption, decision-making processes, and broader policy contexts. For instance, disruptions to clinicians' workflows [167] or gaps in organizational readiness [168] can significantly impede adoption. Furthermore, the explainability of AI algorithms and reproducibility of results are critical to building trust among health care professionals and patients [169]. The adoption of explainable artificial intelligence techniques, such as SHAP (Shapley Additive Explanations) [170] and Local Interpretable Model-Agnostic Explanations [171], can help elucidate the influence of individual acoustic features on AI model predictions. In our review, only Verde et al [172] used Local Interpretable Model-Agnostic Explanations to assess the relevance of acoustic features in the best-performing machine learning models, while Lin et al [79] applied SHAP. Notably, no other included studies used such explainable artificial intelligence techniques. Additionally, comprehensive frameworks for accountability and liability are necessary to clarify stakeholder responsibilities during the deployment process [169]. Finally, compliance with data protection laws must remain a priority to safeguard patient privacy [169]. For example, speech features could be extracted in a manner that prevents raw speech signal reconstruction [152,173], or processed and encrypted on local devices before being securely transmitted to servers for further analysis [174].

Limitations

While this meta-analysis represents the first comprehensive review summarizing the performance of ASA in detecting depression, several limitations warrant mention. First, the analysis was confined to English-language papers, which may have excluded relevant research conducted in other languages. Second, we included only those studies that provide the performance metrics under examination. However, in cases where studies lacked complete information, we did not contact the authors to obtain the missing data. Third, the included studies encompassed diverse patient populations, which adds complexity to the interpretation of our findings due to the variability in demographic and clinical characteristics. For example, patients with chronic depression might exhibit different speech features compared to those experiencing a first episode of depression.

This heterogeneity may influence the diagnostic performance of ASA, potentially resulting in greater accuracy in identifying more severe cases of depression while missing milder cases. However, the assessment of ASA performance by severity level was beyond the scope of the present meta-analysis, as we excluded studies specifically focused on the prediction of depression severity. Moreover, most included studies did not provide sufficient information on the severity of depression within their samples. Future research should therefore examine the performance of ASA across varying levels of depression severity, with particular emphasis on mild cases, to better evaluate its potential as a screening tool for early detection and intervention. Fourth, while we used a modified version of QUADAS-2 as proposed by Abd-Alrazaq et al [137], we acknowledge that this tool still has notable limitations when applied to AI-based research. QUADAS-2 was originally developed for conventional diagnostic accuracy studies and does not adequately address AI-specific sources of bias, such as algorithm and input data quality, real-world clinical applicability, and algorithm generalizability, among others [175]. Additionally, the "flow and timing" domain in QUADAS-2 is often weakly applied in this context, as speech data collection and depression assessments are often conducted simultaneously, limiting its relevance for evaluating temporal relationships or diagnostic latency. Recent initiatives have proposed comprehensive frameworks to support the development and evaluation of trustworthy AI in health care based on 6 guiding principles, including fairness, universality, traceability, usability, robustness, and explainability [176]. Building on these principles, a dedicated, standardized checklist tailored to AI-based diagnostic studies is urgently needed. Such a tool would enhance the rigor of bias assessment, improve reporting practices, and support the effective translation of AI solutions into clinical settings. Fifth, while we reported accuracy, sensitivity, specificity, and precision, these metrics alone may not fully capture model performance in the context of imbalanced datasets, which are common in depression detection. Metrics such as the F_1-score, the area under the curve of the receiver operating characteristic curve, and the Matthews correlation coefficient are better suited for evaluating performance under class imbalance and should be considered in future research to provide a more comprehensive assessment of model effectiveness. Finally, our approach to pooling both the highest and lowest performance metrics across studies deviates from conventional meta-analytic practices and may limit direct comparability with prior work. However, we argue that this method offers specific advantages that justify its use in this context. Previous meta-analyses, such as that conducted by Liu et al [30], have typically reported only the highest performance, which may inadvertently overestimate model effectiveness, particularly in studies where multiple experiments are conducted and only the highest performance metrics are selected. By including both the highest and lowest reported outcomes, we sought to capture the inherent variability in AI-based research and mitigate potential reporting bias. This strategy is supported by prior work from Abd-Alrazaq et al [137], and aligns with ongoing calls for greater methodological transparency in AI research. For researchers seeking direct comparisons with more conventional meta-analyses, the highest pooled estimates in our analysis can serve as a benchmark.

Conclusions

In sum, this study showed that ASA is a promising method for detecting depression, though its readiness for clinical application as a standalone tool remains limited. More peer-reviewed, high-quality studies are warranted to further advance this emerging field. At present, ASA should be considered as a complementary method, with potential application across various settings, including the general population, clinical field, primary care, or environments where stigma still presents a significant barrier to care. Future studies should focus on exploring how ASA can generalize across languages and cultures, and how it might be integrated into complex statistical models, alongside other data sources, to significantly improve depression detection within the stratified psychiatry framework. Additionally, successful clinical implementation of ASA will require addressing critical challenges, including ensuring data reproducibility, improving the explainability of AI algorithms, among other ethical and legal considerations.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All data generated or analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

PLM, MDB, and AR-U contributed to the conception and design of this study. PLM searched the electronic databases. PLM, AV, and MTA-C contributed to the screening, data extraction, and risk of bias assessment. PLM conducted the data synthesis. PLM, AV, and MTA-C drafted the initial paper. PLM, MDB, AR-U, AV, MTA-C, JV-S, and JAR-Q participated in the writing, reviewing, and editing of this paper, and have read and approved the published version of this paper.

Conflicts of Interest

JV-S has received travel awards (air tickets + hotel) for taking part in annual psychiatric meetings from Lundbeck and Janssen-Cilag, and was on the speakers' bureau and acted as a consultant for Janssen Cilag. JAR-Q was on the speakers' bureau and acted as a consultant for Biogen, Idorsia, Casen-Recordati, Janssen-Cilag, Novartis, Takeda, Bial, Sincrolab, Neuraxpharm, Novartis, Bristol

Myers Squibb, Medice, Rubió, Uriach, Technofarma, and Raffo in the last 3 years. He also received travel awards (air tickets + hotel) for taking part in psychiatric meetings from Idorsia, Janssen-Cilag, Rubió, Takeda, Bial, and Medice. The Department of Psychiatry, chaired by him, received unrestricted educational and research support from the following companies in the last 3 years: Exeltis, Idorsia, Janssen-Cilag, Neuraxpharm, Oryzon, Roche, Probitas, and Rubió. AR-U acted as a consultant for Danone, and she has collaborated scientifically with Janssen-Cilag, Pileje, Farmasierra, and Organon. She has also received travel awards (air tickets and hotel) for taking part in annual psychiatric meetings from Lundbeck. All other authors declare no financial or nonfinancial competing interests.

Multimedia Appendix 1

PRISMA-DTA checklist.
[DOCX File , 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Study characteristics, risk of bias, meta-regression and subgroup analyses, and search string.
[DOCX File , 223 KB-Multimedia Appendix 2]

References

1. The global burden of disease: 2004 update. World Health Organization. 2008. URL: https://www.who.int/publications/i/item/9789241563710 [accessed 2025-09-20]
2. Greenberg P, Birnbaum H. The economic burden of depression in the US: societal and patient perspectives. Expert Opin Pharmacother. 2005;6(3):369-376. [doi: 10.1002/9783527619672.ch3]
3. Evans-Lacko S, Knapp M. Global patterns of workplace productivity for people with depression: absenteeism and presenteeism costs across eight diverse countries. Soc Psychiatry Psychiatr Epidemiol. 2016;51(11):1525-1537. [FREE Full text] [doi: 10.1007/s00127-016-1278-4] [Medline: 27667656]
4. Saarni SI, Suvisaari J, Sintonen H, Pirkola S, Koskinen S, Aromaa A, et al. Impact of psychiatric disorders on health-related quality of life: general population survey. Br J Psychiatry. 2007;190:326-332. [doi: 10.1192/bjp.bp.106.025106] [Medline: 17401039]
5. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat. 2011;7(Suppl 1):3-7. [FREE Full text] [doi: 10.2147/NDT.S19617] [Medline: 21750622]
6. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers. 2020;6(1):69. [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]
7. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet. 2007;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
8. Chang C, Hayes RD, Broadbent M, Fernandes AC, Lee W, Hotopf M, et al. All-cause mortality among people with serious mental illness (SMI), substance use disorders, and depressive disorders in southeast London: a cohort study. BMC Psychiatry. 2010;10(1):77. [FREE Full text] [doi: 10.1186/1471-244X-10-77] [Medline: 20920287]
9. Hawton K, Casañas I Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affective Disord. 2013;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
10. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Arlington, VA. American Psychiatric Publishing, Inc; 2022.
11. Organization WH. International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index. Geneva. World Health Organization; 2004.
12. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613. [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
13. Cheung R. Patient health questionnaire-9 (PHQ-9). In: Medvedev ON, Krägeloh CU, Siegert RJ, editors. Handbook of Assessment in Mindfulness Research. Cham. Springer International Publishing; 2022:1-11.
14. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri T. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10-49. [FREE Full text] [doi: 10.1016/j.specom.2015.03.004]
15. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013;170(1):59-70. [doi: 10.1176/appi.ajp.2012.12070999] [Medline: 23111466]
16. Mundt J, Snyder P, Cannizzaro M, Chappie K, Geralts D. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics. 2007;20(1):50-64. [FREE Full text] [doi: 10.1016/j.jneuroling.2006.04.001] [Medline: 21253440]

17. Wainberg ML, Scorza P, Shultz JM, Helpman L, Mootz JJ, Johnson KA, et al. Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep. 2017;19(5):28. [FREE Full text] [doi: 10.1007/s11920-017-0780-z] [Medline: 28425023]
18. Murray CJL, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. Lancet. 2012;380(9859):2197-2223. [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
19. Oladeji BD, Gureje O. Brain drain: a challenge to global mental health. BJPsych Int. 2016;13(3):61-63. [FREE Full text] [doi: 10.1192/s2056474000001240] [Medline: 29093905]
20. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol. 2020;5(1):96-116. [FREE Full text] [doi: 10.1002/lio2.354] [Medline: 32128436]
21. Sharp T, Cowen P. 5-HT and depression: is the glass half-full? Curr Opin Pharmacol. 2011;11(1):45-51. [doi: 10.1016/j.coph.2011.02.003] [Medline: 21377932]
22. Luscher B, Shen Q, Sahir N. The GABAergic deficit hypothesis of major depressive disorder. Mol Psychiatry. 2011;16(4):383-406. [FREE Full text] [doi: 10.1038/mp.2010.120] [Medline: 21079608]
23. Gatt JM, Nemeroff CB, Dobson-Stone C, Paul RH, Bryant RA, Schofield PR, et al. Interactions between BDNF Val66Met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety. Mol Psychiatry. 2009;14(7):681-695. [doi: 10.1038/mp.2008.143] [Medline: 19153574]
24. Tasci G, Loh H, Barua P, Baygin M, Tasci B, Dogan S, et al. Automated accurate detection of depression using twin Pascal's triangles lattice pattern with EEG signals. Knowl-Based Syst. Jan 2023;260:110190. [doi: 10.1016/j.knosys.2022.110190]
25. Loh HW, Ooi CP, Aydemir E, Tuncer T, Dogan S, Acharya UR. Decision support system for major depression detection using spectrogram and convolution neural network with EEG signals. Expert Syst. 2021;39(3):e12773. [doi: 10.1111/exsy.12773]
26. Shen R, Zhan Q, Wang Y, Ma H. Depression detection by analysing eye movements on emotional images. IEEE; 2021. Presented at: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 06-11, 2021; Toronto, Ontario, Canada. [doi: 10.1109/icassp39728.2021.9414663]
27. Wang T, Li C, Wu C, Zhao C, Sun J, Peng H, et al. A gait assessment framework for depression detection using kinect sensors. IEEE Sensors J. 2021;21(3):3260-3270. [doi: 10.1109/jsen.2020.3022374]
28. Osimo EF, Pillinger T, Rodriguez IM, Khandaker GM, Pariante CM, Howes OD. Inflammatory markers in depression: a meta-analysis of mean differences and variability in 5,166 patients and 5,083 controls. Brain Behav Immun. 2020;87:901-909. [FREE Full text] [doi: 10.1016/j.bbi.2020.02.010] [Medline: 32113908]
29. Almaghrabi S, Clark S, Baumert M. Bio-acoustic features of depression: a review. Biomed Signal Process Control. 2023;85:105020. [FREE Full text] [doi: 10.1016/j.bspc.2023.105020]
30. Liu L, Liu L, Wafa H, Tydeman F, Xie W, Wang Y. Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis. J Am Med Inform Assoc. 2024;31(10):2394-2404. [doi: 10.1093/jamia/ocae189] [Medline: 39013193]
31. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319(4):388-396. [FREE Full text] [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
32. Xing Y, He R, Zhang C, Tan P. Hierarchical multi-task learning based on interactive multi-head attention feature fusion for speech depression recognition. IEEE Access. 2025;13:51208-51219. [doi: 10.1109/access.2025.3551549]
33. Pratiwi M, Sanjaya S. Vision transformer for audio-based depression detection on multi-lingual audio data. Association for Computing Machinery; 2024. Presented at: DMIP '24: Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing; November 8-11, 2024:35-41; Osaka Japan. [doi: 10.1145/3705927.3705934]
34. Lim E, Jhon M, Kim J, Kim S, Kim S, Yang H. A lightweight approach based on cross-modality for depression detection. Comput Biol Med. 2025;186:109618. [FREE Full text] [doi: 10.1016/j.compbiomed.2024.109618] [Medline: 39765105]
35. Idamkina M, Corradini A. Detecting depression from audio data. ACM; 2024. Presented at: International Conference on Speech and Computer; November 25, 2024:336-351; Belgrade, Serbia. [doi: 10.1007/978-3-031-77961-9_25]
36. Zhou D, Mizuguchi D, Yamamoto T, Omiya Y. A study on depression detection through explainable features of speech. IEEE; 2024. Presented at: IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP); November 07-10, 2024; Beijing, China. [doi: 10.1109/iscslp63861.2024.10800164]
37. Zhang X, Zhang X, Chen W, Li C, Yu C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Sci Rep. 2024;14(1):9543. [FREE Full text] [doi: 10.1038/s41598-024-60278-1] [Medline: 38664511]
38. Wang J, Ravi V, Flint J, Alwan A. Speechformer-CTC: sequential modeling of depression detection with speech temporal classification. Speech Commun. 2024;163:103106. [FREE Full text] [doi: 10.1016/j.specom.2024.103106] [Medline: 39364289]
39. Taşcı B. Multilevel hybrid handcrafted feature extraction based depression recognition method using speech. J Affective Disord. 2024;364:9-19. [doi: 10.1016/j.jad.2024.08.002] [Medline: 39127304]

40. Ravi V, Wang J, Flint J, Alwan A. Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. Comput Speech Lang. 2024;86:101605. [FREE Full text] [doi: 10.1016/j.csl.2023.101605] [Medline: 38313320]
41. Menne F, Dörr F, Schräder J, Tröger J, Habel U, König A, et al. The voice of depression: speech features as biomarkers for major depressive disorder. BMC Psychiatry. 2024;24(1):794. [FREE Full text] [doi: 10.1186/s12888-024-06253-6] [Medline: 39533239]
42. Maji B, Nasreen S, Guha R, Routray A, Majumdar D, Poonam K. Exploring self-supervised models for depressive disorder detection: a study on speech corpora. IEEE; 2024. Presented at: 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 15-19, 2024; Orlando, FL. [doi: 10.1109/embc53108.2024.10781765]
43. Liu L, Tydeman F, Xie W, Wang Y. Multilingual depression detection based on speech signals and deep learning. IEEE; 2024. Presented at: IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService); July 15-18, 2024; Shanghai, China. [doi: 10.1109/bigdataservice62917.2024.00024]
44. Liang L, Wang Y, Ma H, Zhang R, Liu R, Zhu R, et al. Enhanced classification and severity prediction of major depressive disorder using acoustic features and machine learning. Front Psychiatry. 2024;15:1422020. [FREE Full text] [doi: 10.3389/fpsyt.2024.1422020] [Medline: 39355380]
45. Li J, Li Y. Recognition of mild-to-moderate depression based on facial expression and speech. Association for Computing Machinery; 2024. Presented at: CNIOT 2024: 2024 5th International Conference on Computing, Networks and Internet of Things; May 24-26, 2024:26-31; Tokyo, Japan. [doi: 10.1145/3670105.3670110]
46. Ilias L, Askounis D. A cross-attention layer coupled with multimodal fusion methods for recognizing depression from spontaneous speech. 2024. Presented at: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; 2024; September 1-5, 2024; Kos Island, Greece. [doi: 10.21437/interspeech.2024-188]
47. Hong J, Lee J, Cho D, Jung J. Depression classification algorithm based on voice signals using MFCC and CNN autoencoders. 2024. Presented at: International Conference on Machine Learning and Applications (ICMLA); December 16-19, 2024:18-20; Boca Raton, Florida. [doi: 10.1109/icmla61862.2024.00213]
48. Di Y, Rahmani E, Mefford J, Wang J, Ravi V, Gorla A. Unraveling the associations between voice pitch and major depressive disorder: a multisite genetic study. Res Square. Preprint posted online on April 03, 2024. [doi: 10.21203/rs.3.rs-4135145/v1]
49. Yang W, Liu J, Cao P, Zhu R, Wang Y, Liu JK, et al. Attention guided learnable time-domain filterbanks for speech depression detection. Neural Netw. 2023;165:135-149. [doi: 10.1016/j.neunet.2023.05.041] [Medline: 37285730]
50. Xu X, Zhang G, Lu Q, Mao X. Multimodal depression recognition that integrates audio and text. IEEE; 2023. Presented at: 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC); Nanjing, China:18-20; August 8-20, 2023. [doi: 10.1109/isceic59030.2023.10271158]
51. Tian H, Zhu Z, Jing X. Deep learning for depression recognition from speech. Mobile Netw Appl. 2023;29(4):1212-1227. [doi: 10.1007/s11036-022-02086-3]
52. Srinivasan J, Vishnu AJS, Pragna N, Polavarapu R. Unleashing the potential of convolutional neural networks for automated depression detection using audio modality. IEEE; 2023. Presented at: 7th International Conference On Computing, Communication, Control and Automation (ICCUBEA); August 18-19, 2023:18-19; Pune, India. [doi: 10.1109/iccubea58933.2023.10391952]
53. Qayyum A, Razzak I, Tanveer M, Mazher M, Alhaqbani B. High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(4):2587-2597. [doi: 10.1109/TCBB.2023.3257175] [Medline: 37028339]
54. Pan W, Deng F, Wang X, Hang B, Zhou W, Zhu T. Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls. Front Psychiatry. 2023;14:1079448. [FREE Full text] [doi: 10.3389/fpsyt.2023.1079448] [Medline: 37575564]
55. Mamidisetti S, Reddy AM. A stacking-based ensemble framework for automatic depression detection using audio signals. IJACSA. 2023;14(7):603-612. [doi: 10.14569/ijacsa.2023.0140767]
56. Maji B, Roy AK, Nasreen S, Guha R, Routray A, Majumdar D. A novel technique for detecting depressive disorder: a speech database-based approach. IEEE; 2023. Presented at: 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); July 24-27, 2023; Sydney, Australia. [doi: 10.1109/embc40787.2023.10341118]
57. Maheshwar V, Venu GN, Naveen KV, Pranavi D, Padma SY. Development of an SVM-based depression detection model using MFCC feature extraction. 2023. Presented at: 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS); June 14-16, 2023; Coimbatore, India. [doi: 10.1109/icscss57650.2023.10169770]
58. Liu Z, Yu H, Li G, Chen Q, Ding Z, Feng L, et al. Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection. Front Neurosci. 2023;17:1141621. [FREE Full text] [doi: 10.3389/fnins.2023.1141621] [Medline: 37034153]
59. Kim AY, Jang EH, Lee S, Choi K, Park JG, Shin H. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. J Med Internet Res. 2023;25:e34474. [FREE Full text] [doi: 10.2196/34474] [Medline: 36696160]
60. Han Z, Shang Y, Shao Z, Liu J, Guo G, Liu T, et al. Spatial-temporal feature network for speech-based depression recognition. IEEE Trans Cognit Dev Syst. 2024;16(1):308-318. [doi: 10.1109/tcds.2023.3273614]

61. Gheorghe M, Mihalache S, Burileanu D. Using deep neural networks for detecting depression from speech. IEEE; 2023. Presented at: 31st European Signal Processing Conference (EUSIPCO); September 04-08, 2023; Helsinki, Finland. [doi: 10.23919/eusipco58844.2023.10289973]
62. Ghadiri N, Samani R, Shahrokh F. Integration of text and graph-based features for depression detection using visibility graph. 2023. Presented at: 22nd International Conference on Intelligent Systems Design and Applications (ISDA 2022); December 12-14, 2022:332-341; 22nd International Conference on Intelligent Systems Design and Applications (ISDA 2022). [doi: 10.1007/978-3-031-27440-4_32]
63. Dumpala SH, Dikaios K, Rodriguez S, Langley R, Rempel S, Uher R, et al. Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. Sci Rep. 2023;13(1):11155. [FREE Full text] [doi: 10.1038/s41598-023-35184-7] [Medline: 37429935]
64. Du M, Zhang W, Wang T, Liu S, Ming D. An automatic depression recognition method from spontaneous pronunciation using machine learning. Association for Computing Machinery; 2022. Presented at: ICBBE '22: Proceedings of the 2022 9th International Conference on Biomedical and Bioinformatics Engineering; November 10-13, 2022:133-139; Kyoto Japan. [doi: 10.1145/3574198.3574219]
65. Du M, Liu S, Wang T, Zhang W, Ke Y, Chen L, et al. Depression recognition using a proposed speech chain model fusing speech production and perception features. J Affective Disord. 2023;323:299-308. [FREE Full text] [doi: 10.1016/j.jad.2022.11.060] [Medline: 2775765243]
66. Bhavya S, Nayak D, Dmello R, Nayak A, Bangera S. Machine learning applied to speech emotion analysis for depression recognition. IEEE; 2023. Presented at: International Conference for Advancement in Technology (ICONAT); January 24-26, 2023:24-26; Goa, India. [doi: 10.1109/iconat57137.2023.10080060]
67. Berardi M, Brosch K, Pfarr J, Schneider K, Sültmann A, Thomas-Odenthal F, et al. Relative importance of speech and voice features in the classification of schizophrenia and depression. Transl Psychiatry. 2023;13(1):298. [FREE Full text] [doi: 10.1038/s41398-023-02594-0] [Medline: 37726285]
68. Suparatpinyo S, Soonthornphisaj N. Smart voice recognition based on deep learning for depression diagnosis. Artif Life Rob. 2023;28(2):332-342. [doi: 10.1007/s10015-023-00852-4]
69. Zhou Z, Guo Y, Hao S, Hong R. Hierarchical multifeature fusion via audio-response-level modeling for depression detection. IEEE Trans Comput Soc Syst. 2023;10(5):2797-2805. [doi: 10.1109/tcss.2022.3202294]
70. Zou B, Han J, Wang Y, Liu R, Zhao S, Feng L, et al. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans Affective Comput. 2023;14(4):2823-2838. [doi: 10.1109/taffc.2022.3181210]
71. Zhao Q, Fan H, Li Y, Liu L, Wu Y, Zhao Y, et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. Front Psychiatry. 2022;13:815678. [FREE Full text] [doi: 10.3389/fpsyt.2022.815678] [Medline: 35573349]
72. Xing Y, Liu Z, Li G, Ding Z, Hu B. 2-level hierarchical depression recognition method based on task-stimulated and integrated speech features. Biomedical Signal Processing and Control. 2022;72:103287. [doi: 10.1016/j.bspc.2021.103287]
73. Wu H, Hu W, Fu D. Autoencoder based on VMD and mutual information to detect depression from speech. 2022. Presented at: RICAI '22: Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence; December 16-18, 2022:424-428; Dongguan, China. [doi: 10.1145/3584376.3584453]
74. Wang J, Ravi V, Flint J, Alwan A. Unsupervised instance discriminative learning for depression detection from speech signals. Interspeech. 2022;2022:2018-2022. [FREE Full text] [doi: 10.21437/interspeech.2022-10814] [Medline: 36341466]
75. Sun G, Zhao S, Zou B, An Y. Speech-based depression detection using unsupervised autoencoder. IEEE; 2022. Presented at: 7th International Conference on Signal and Image Processing (ICSIP); July 20-22, 2022:20-22; Suzhou, China. [doi: 10.1109/icsip55141.2022.9886372]
76. Pérez-Toro PA, Arias-Vergara T, Klumpp P, Vásquez-Correa JC, Schuster M, Nöth E, et al. Depression assessment in people with Parkinson's disease: the combination of acoustic features and natural language processing. Speech Commun. 2022;145:10-20. [doi: 10.1016/j.specom.2022.09.001]
77. Mobram S, Vali M. Depression detection based on linear and nonlinear speech features in I-vector/SVDA framework. Comput Biol Med. 2022;149:105926. [doi: 10.1016/j.compbiomed.2022.105926] [Medline: 36037628]
78. Miao X, Li Y, Wen M, Liu Y, Julian IN, Guo H. Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Commun. 2022;143:46-56. [doi: 10.1016/j.specom.2022.07.006]
79. Lin Y, Liyanage BN, Sun Y, Lu T, Zhu Z, Liao Y, et al. A deep learning-based model for detecting depression in senior population. Front Psychiatry. 2022;13:1016676. [FREE Full text] [doi: 10.3389/fpsyt.2022.1016676] [Medline: 36419976]
80. König A, Tröger J, Mallick E, Mina M, Linz N, Wagnon C, et al. Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. BMC Psychiatry. 2022;22(1):830. [FREE Full text] [doi: 10.1186/s12888-022-04475-0] [Medline: 36575442]
81. Janardhan N, Kumaresh N. Improving depression prediction accuracy using fisher score-based feature selection and dynamic ensemble selection approach based on acoustic features of speech. TS. 2022;39(1):87-107. [doi: 10.18280/ts.390109]

82. Higuchi M, Nakamura M, Shinohara S, Omiya Y, Takano T, Mizuguchi D, et al. Detection of major depressive disorder based on a combination of voice features: an exploratory approach. Int J Environ Res Public Health. 2022;19(18):11397. [FREE Full text] [doi: 10.3390/ijerph191811397] [Medline: 36141675]
83. He Y, Lu X, Yuan J, Pan T, Wang Y. Depressive tendency recognition by fusing speech and text features: a comparative analysis. IEEE; 2022. Presented at: 13th International Symposium on Chinese Spoken Language Processing (ISCSLP); December 11-14, 2022:11-14; Singapore, Singapore. [doi: 10.1109/iscslp57327.2022.10038078]
84. Nik Hashim NNW, Basri NA, Ahmad Ezzi MA, Nik Hashim NMH. Comparison of classifiers using robust features for depression detection on Bahasa Malaysia speech. IJ-AI. 2022;11(1):238. [doi: 10.11591/ijai.v11.i1.pp238-253]
85. Hansen L, Zhang Y, Wolf D, Sechidis K, Ladegaard N, Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatr Scand. 2022;145(2):186-199. [doi: 10.1111/acps.13388] [Medline: 34850386]
86. Hajduska-Dér B, Kiss G, Sztahó D, Vicsi K, Simon L. The applicability of the Beck Depression Inventory and Hamilton Depression Scale in the automatic recognition of depression based on speech signal processing. Front Psychiatry. 2022;13:879896. [FREE Full text] [doi: 10.3389/fpsyt.2022.879896] [Medline: 35990073]
87. Ezzi MAEA, Hashim NNWN, Basri NA. Microphone-independent speech features for automatic depression detection using recurrent neural network. Springer; 2022. Presented at: Proceedings of the 8th International Conference on Computational Science and Technology; August 28–29, 2021:711-724; Labuan, Malaysia. [doi: 10.1007/978-981-16-8515-6_54]
88. Egas-López J, Kiss G, Sztahó D, Gosztolya G. Automatic assessment of the degree of clinical depression from speech using X-vectors. IEEE; 2022. Presented at: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore, Singapore. [doi: 10.1109/icassp43922.2022.9746068]
89. Bn S, Abdullah S. Privacy sensitive speech analysis using federated learning to assess depression. IEEE; 2022. Presented at: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore, Singapore. [doi: 10.1109/icassp43922.2022.9746827]
90. Alsharif Z, Elhag S, Alfakeh S. Depression detection in Arabic using speech language recognition. IEEE; 2022. Presented at: 7th International Conference on Data Science and Machine Learning Applications (CDMA); March 01-03, 2022; Riyadh, Saudi Arabia.
91. Wang H, Liu Y, Zhen X, Tu X. Depression speech recognition with a three-dimensional convolutional network. Front Hum Neurosci. 2021;15:713823. [FREE Full text] [doi: 10.3389/fnhum.2021.713823] [Medline: 34658815]
92. Verde L, Raimo G, Vitale F, Carbonaro B, Cordasco G, Marrone S. A lightweight machine learning approach to detect depression from speech analysis. IEEE; 2021. Presented at: IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI); November 01-03, 2021; Washington, DC. [doi: 10.1109/ictai52525.2021.00054]
93. Sztahó D, Gábor K, Gábriel T. Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning. 2021. Presented at: 14th International Conference on Bio-Inspired Systems and Signal Processing; February 11-13, 2021; Vienna, Austria. [doi: 10.5220/0010193101350141]
94. Patil M, Wadhai V. Selection of classifiers for depression detection using acoustic features. IEEE; 2021. Presented at: International Conference on Computational Intelligence and Computing Applications (ICCICA); November 26-27, 2021; Nagpur, India. [doi: 10.1109/iccica52458.2021.9697240]
95. Othmani A, Kadoch D, Bentounes K, Rejaibi E, Alfred R, Hadid A. Towards robust deep neural networks for affect and depression recognition from speech. 2021. Presented at: Pattern Recognition. ICPR International Workshops and Challenges; January 10-15, 2021:5-19; Virtual Event. [doi: 10.1007/978-3-030-68790-8_1]
96. Lu X, Shi D, Liu Y, Yuan J. Speech depression recognition based on attentional residual network. Front Biosci (Landmark Ed). Dec 30, 2021;26(12):1746-1759. [FREE Full text] [doi: 10.52586/5066] [Medline: 34994187]
97. Lee S, Suh SW, Kim T, Kim K, Lee KH, Lee JR, et al. Screening major depressive disorder using vocal acoustic features in the elderly by sex. J Affective Disord. 2021;291:15-23. [doi: 10.1016/j.jad.2021.04.098] [Medline: 34022551]
98. Kiss G, Sztahó D, Tulics MG. Application for detecting depression, Parkinson's disease and dysphonic speech. 2021. Presented at: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; February 5, 2021; Czechia.
99. Jenei AZ, Kiss G. Severity estimation of depression using convolutional neural network. Period Polytech Elec Eng Comp Sci. 2021;65(3):227-234. [doi: 10.3311/ppee.15958]
100. Espinola CW, Gomes JC, Pereira JMS, dos Santos WP. Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. Res Biomed Eng. 2021;37(1):53-64. [doi: 10.1007/s42600-020-00100-9]
101. Elfaki A, Asnawi A, Jusoh A, Ismail A, Ibrahim S, Azmin N. Using the short-time fourier transform and resNet to diagnose depression from speech data. 2021. Presented at: Using the Short-Time Fourier Transform and ResNet to Diagnose Depression from Speech Data. 2021 IEEE International Conference on Computing (ICOCO); 2021 Nov. 2021; November 17-19, 2021:17-19; Malaysia. [doi: 10.1109/icoco53166.2021.9673562]
102. Dong Y, Yang X. A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing. 2021;441:279-290. [doi: 10.1016/j.neucom.2021.02.019]

103. Di Y, Wang J, Li W, Zhu T. Using i-vectors from voice features to identify major depressive disorder. J Affective Disord. 2021;288:161-166. [doi: 10.1016/j.jad.2021.04.004] [Medline: 33895418]
104. Yalamanchili B, Kota N, Abbaraju M, Nadella V, Alluri S. Real-time acoustic based depression detection using machine learning techniques. IEEE; 2020. Presented at: International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE); February 24-25, 2020; Vellore, India. [doi: 10.1109/ic-etite47903.2020.394]
105. Shukla DM, Sharma K, Gupta S. Identifying depression in a person using speech signals by extracting energy and statistical features. IEEE; 2020. Presented at: IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS); February 22-23, 2020; Bhopal, India. [doi: 10.1109/sceecs48394.2020.60]
106. Muzammel M, Salam H, Hoffmann Y, Chetouani M, Othmani A. AudVowelConsNet: a phoneme-level based deep CNN architecture for clinical depression diagnosis. Mach Learn Appl. 2020;2:100005. [doi: 10.1016/j.mlwa.2020.100005]
107. Jenei AZ, Kiss G. Possibilities of recognizing depression with convolutional networks applied in correlation structure. IEEE; 2020. Presented at: 43rd International Conference on Telecommunications and Signal Processing (TSP); July 07-09, 2020:7-9; Milan, Italy. [doi: 10.1109/tsp49548.2020.9163547]
108. Demiroglu C, Beşirli A, Ozkanca Y, Çelik S. Depression-level assessment from multi-lingual conversational speech data using acoustic and text features. J Audio Speech Music Process. 2020;2020(1):17. [doi: 10.1186/s13636-020-00182-4]
109. Al-Hameed S. Audio Based Signal Processing and Computational Models for Early Detection and Prediction of Dementia and Mood Disorders. United Kingdom. University of Sheffield; 2020.
110. Xu S, Yang Z, Chakraborty D, Victoria CY, Dauwels J, Thalmann D. Automated verbal and non-verbal speech analysis of interviews of individuals with schizophrenia and depression. IEEE; 2019. Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany. [doi: 10.1109/embc.2019.8857071]
111. Huang Z, Epps J, Joachim D, Sethu V. Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. IEEE J Sel Top Signal Process. 2020;14(2):435-448. [doi: 10.1109/jstsp.2019.2949419]
112. Huang Z, Epps J, Joachim D. Speech landmark bigrams for depression detection from naturalistic smartphone speech. IEEE; 2019. Presented at: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 12-17, 2019; Brighton, UK. [doi: 10.1109/icassp.2019.8682916]
113. Balano JB, Huerto VL, Sanchez S, Saharkhiz A, De GJ. Determining the level of depression using BDI-II through voice recognition. 2019. Presented at: IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA); April 12-15, 2019; Tokyo, Japan. [doi: 10.1109/iea.2019.8715187]
114. Aradhana M, Chander S, Krishna B, Amritha S, Roy R. Diagnosing clinical depression from voice: using signal processing and neural network algorithms to build a mental wellness monitor. IEEE; 2019. Presented at: International Conference on Advances in Computing, Communication and Control (ICAC3); December 20-21, 2019; Mumbai, India. [doi: 10.1109/icac347590.2019.9036838]
115. Alghifari M, Gunawan T, Nordin M, Kartiwi M, Borhan L. On the optimum speech segment length for depression detection. 2019. Presented at: IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA); August 27-29, 2019:27-29; Kuala Lumpur, Malaysia. [doi: 10.1109/icsima47653.2019.9057319]
116. Sztahó D, Kiss G, Tulics MG, Vicsi K. Automatic separation of various disease types by correlation structure of time shifted speech features. IEEE; 2018. Presented at: 41st International Conference on Telecommunications and Signal Processing (TSP); July 04-06, 2018; Athens, Greece. [doi: 10.1109/tsp.2018.8441395]
117. Lech M. Detection of adolescent depression from speech using optimised spectral roll-off parameters. BJSTR. 2018;5(1):4350-4359. [doi: 10.26717/bjstr.2018.05.0001156]
118. Jiang H, Hu B, Liu Z, Wang G, Zhang L, Li X, et al. Detecting depression using an ensemble logistic regression model based on multiple speech features. Comput Math Methods Med. 2018;2018:6508319. [FREE Full text] [doi: 10.1155/2018/6508319] [Medline: 30344616]
119. Afshan A, Guo J, Park SJ, Ravi V, Flint J, Alwan A. Effectiveness of voice quality features in detecting depression. 2018. Presented at: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; August 17-21, 2018; Rotterdam, The Netherlands. [doi: 10.21437/interspeech.2018-1399]
120. Jiang H, Hu B, Liu Z, Yan L, Wang T, Liu F, et al. Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Commun. 2017;90:39-46. [doi: 10.1016/j.specom.2017.04.001]
121. Liu Z, Li C, Gao X, Wang G, Yang J. Ensemble-based depression detection in speech. IEEE; 2017. Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 13-16, 2017:13-16; Kansas City, MO. [doi: 10.1109/bibm.2017.8217789]
122. Long H, Guo Z, Wu X, Hu B, Liu Z, Cai H. Detecting depression in speech: comparison and combination between different speech types. IEEE; 2017. Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 13-16, 2017:1052-1058; Kansas City, MO. [doi: 10.1109/bibm.2017.8217802]
123. Parekh P, Patil M. Clinical depression detection for adolescent by speech features. IEEE; 2017. Presented at: International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS); August 01-02, 2017; Chennai, India. [doi: 10.1109/icecds.2017.8390102]

124. Tundik M, Kiss G, Sztaho D, Szaszak G. Assessment of pathological speech prosody based on automatic stress detection and phrasing approaches. IEEE; 2017. Presented at: 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); September 11-14, 2017; Debrecen, Hungary. [doi: 10.1109/coginfocom.2017.8268218]
125. Kiss G, Vicsi K. Comparison of read and spontaneous speech in case of automatic detection of depression. IEEE; 2017. Presented at: 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); September 11-14, 2017:11-14; Debrecen, Hungary. [doi: 10.1109/coginfocom.2017.8268223]
126. Kiss G, Vicsi K. Mono- and multi-lingual depression prediction based on speech processing. Int J Speech Technol. 2017;20(4):919-935. [doi: 10.1007/s10772-017-9455-8]
127. Azam H, Hashim NNWN, Sediono W, Mukhtar F, Ibrahim N, Mokhtar S. Classifications of clinical depression detection using acoustic measures in Malay speakers. IEEE; 2016. Presented at: IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES); December 04-08, 2016:4-8; Kuala Lumpur, Malaysia. [doi: 10.1109/iecbes.2016.7843521]
128. Fraser K, Rudzicz F, Hirst G. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. 2016. Presented at: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology; September 15, 2025:1-11; San Diego, CA. [doi: 10.18653/v1/w16-0301]
129. Gillespie S, Moore E, Laures-Gore J, Farina M. Exploratory analysis of speech features related to depression in adults with Aphasia. IEEE; 2016. Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 20-25, 2016; Shanghai, China. [doi: 10.1109/icassp.2016.7472792]
130. Ma X, Yang H, Chen Q, Huang D, Wang Y. Depaudionet: an efficient deep model for audio based depression classification. ACM; 2016. Presented at: AVEC '16: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge; October 16, 2016:35-42; Amsterdam The Netherlands. [doi: 10.1145/2988257.2988267]
131. Stolar M. Acoustic and Conversational Speech Analysis of Depressed Adolescents and Their Parents. Australia. RMIT University; 2016.
132. Stolar MN, Lech M, Allen NB. Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations. IEEE; 2015. Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 19-24, 2015; South Brisbane, QLD, Australia. [doi: 10.1109/icassp.2015.7178117]
133. Asgari M, Shafran I, Sheeber L. Inferring clinical depression from speech and spoken utterances. IEEE; 2014. Presented at: IEEE International Workshop on Machine Learning for Signal Processing (MLSP); September 21-24, 2014; Reims, France. [doi: 10.1109/mlsp.2014.6958856]
134. Lopez-Otero P, Docio-Fernandez L, Garcia-Mateo C. A study of acoustic features for the classification of depressed speech. IEEE; 2014. Presented at: 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); May 26-30, 2014; Opatija, Croatia. [doi: 10.1109/mipro.2014.6859774]
135. Alghowinem S, Goecke R, Wagner M, Epps J, Parker G, Breakspear M. Characterising depressed speech for classification. 2013. Presented at: 14th Annual Conference of the International Speech Communication Association (InterSpeech); August 25-29, 2013; Lyon, France. [doi: 10.21437/interspeech.2013-571]
136. Guo M, Wang J, Li D, Chang L. Depression detection using the derivative features of group delay and Delta phase spectrum. IEEE; 2013. Presented at: 3rd IEEE International Advance Computing Conference (IACC); February 22-23, 2013:22-23; Ghaziabad, India. [doi: 10.1109/iadcc.2013.6514411]
137. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med. 2023;6(1):84. [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
138. Wilson D, Lipsey M. The role of method in treatment effectiveness research: evidence from meta-analysis. Psychol Methods. 2001;6(4):413. [doi: 10.1037//1082-989x.6.4.413-429]
139. Assink M, Wibbelink CJM. Fitting three-level meta-analytic models in R: a step-by-step tutorial. TQMP. 2016;12(3):154-174. [doi: 10.20982/tqmp.12.3.p154]
140. Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. Behav Res Methods. 2013;45(2):576-594. [doi: 10.3758/s13428-012-0261-6] [Medline: 23055166]
141. Borenstein M, Hedges L, Higgins J, Rothstein H. Introduction to Meta-Analysis. New Jersey. John Wiley & Sons; 2021:87.
142. Jonathan JDJ, Douglas G, Altman. Analysing data and undertaking meta-analyses. Cochrane Statistical Methods Group. 2019. URL: https://training.cochrane.org/handbook/current/chapter-10 [accessed 2025-09-20]
143. Borenstein M, Cooper H, Hedges L, Valentine J. Effect sizes for continuous data. In: The Handbook of Research Synthesis and Meta-Analysis. New York. Russell Sage Foundation; 2019:221-235.
144. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Software. 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
145. R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing. 2023. URL: https://www.r-project.org/ [accessed 2025-09-20]
146. Teager H, Teager S. Evidence for nonlinear sound production mechanisms in the vocal tract. In: Speech Production and Speech Modelling. Cham. Springer; 1990:241-261.
147. Sheu Y. Illuminating the black box: interpreting deep neural network models for psychiatric research. Front Psychiatry. 2020;11:551299. [FREE Full text] [doi: 10.3389/fpsyt.2020.551299] [Medline: 33192663]

148. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med. 2022;140:105111. [FREE Full text] [doi: 10.1016/j.compbiomed.2021.105111] [Medline: 34891095]
149. Katirai A. Ethical considerations in emotion recognition technologies: a review of the literature. AI Ethics. Jun 20, 2023;4(4):927-948. [doi: 10.1007/s43681-023-00307-3]
150. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A. The distress analysis interview corpus of human and computer interviews. 2014. Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); September 15, 2025:3123-3128; Reykjavik, Iceland.
151. Cummins N, Dineley J, Conde P, Matcham F, Siddi S, Lamers F, et al. RADAR-CNS Consortium. Multilingual markers of depression in remotely collected speech samples: a preliminary analysis. J Affective Disord. 2023;341:128-136. [FREE Full text] [doi: 10.1016/j.jad.2023.08.097] [Medline: 37598722]
152. Vázquez-Romero A, Gallardo-Antolín A. Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy (Basel). 2020;22(6):688. [FREE Full text] [doi: 10.3390/e22060688] [Medline: 33286460]
153. Bailey A, Plumbley MD. Gender bias in depression detection using audio features. 2021. Presented at: 29th European Signal Processing Conference (EUSIPCO); August 23-27, 2021:23-27; Dublin, Ireland. [doi: 10.23919/eusipco54536.2021.9615933]
154. Zuo L, Mak M. Avoiding dominance of speaker features in speech-based depression detection. Pattern Recognit Lett. 2023;173:50-56. [FREE Full text] [doi: 10.1016/j.patrec.2023.07.016]
155. Ndaba S. Class imbalance handling techniques used in depression prediction and detection. IJDKP. 2023;13(1/2):17-33. [doi: 10.5121/ijdkp.2023.13202]
156. Kearns M, Neel S, Roth A, Wu ZS. An empirical study of rich subgroup fairness for machine learning. 2019. Presented at: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency; January 29-31, 2019:100-109; Atlanta, GA. [doi: 10.1145/3287560.3287592]
157. Davey CG, Harrison BJ. The self on its axis: a framework for understanding depression. Transl Psychiatry. 2022;12(1):23. [FREE Full text] [doi: 10.1038/s41398-022-01790-8] [Medline: 35042843]
158. Kurniadi F, Paramita N, Sihotang E, Anggreainy M, Zhang R. BERT and RoBERTa models for enhanced detection of depression in social media text. Procedia Comput Sci. 2024;245:202-209. [FREE Full text] [doi: 10.1016/j.procs.2024.10.244]
159. Bokolo BG, Liu Q. Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques. Electronics. 2023;12(21):4396. [doi: 10.3390/electronics12214396]
160. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affective Disord. 2025;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
161. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry. 2016;7:50. [FREE Full text] [doi: 10.3389/fpsyt.2016.00050] [Medline: 27064972]
162. Shaikhina T, Khovanova NA. Handling limited datasets with neural networks in medical applications: a small-data approach. Artif Intell Med. 2017;75:51-63. [FREE Full text] [doi: 10.1016/j.artmed.2016.12.003] [Medline: 28363456]
163. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLOS One. 2019;14(11):e0224365. [FREE Full text] [doi: 10.1371/journal.pone.0224365] [Medline: 31697686]
164. Pearson SD, Katzelnick DJ, Simon GE, Manning WG, Helstad CP, Henk HJ. Depression among high utilizers of medical care. J Gen Intern Med. 1999;14(8):461-468. [FREE Full text] [doi: 10.1046/j.1525-1497.1999.06278.x] [Medline: 10491229]
165. Park LT, Zarate CA. Depression in the primary care setting. N Engl J Med. 2019;380(6):559-568. [FREE Full text] [doi: 10.1056/NEJMcp1712493] [Medline: 30726688]
166. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11):e367. [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
167. Fitzgerald L, McDermott A. Challenging Perspectives on Organizational Change in Health Care. Routledge Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK. Taylor & Group; 2017:1138914495.
168. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q. 2004;82(4):581-629. [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
169. Goh S, Goh RSJ, Chong B, Ng QX, Koh GCH, Ngiam KY, et al. Challenges in Implementing Artificial Intelligence in Breast Cancer Screening Programs: Systematic Review and Framework for Safe Adoption. J Med Internet Res. May 15, 2025;27:e62941-e62929. [doi: 10.2196/62941] [Medline: 40373301]
170. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017:4768-4777; Long Beach California USA.
171. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. Association for Computing Machinery; 2016. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016:1135-1144; San Francisco, CA. [doi: 10.1145/2939672.2939778]

172. Verde L, Marulli F, De Fazio R, Campanile L, Marrone S. HEAR set: a lightweight acoustic parameters set to assess mental health from voice analysis. Comput Biol Med. 2024;182:109021. [FREE Full text] [doi: 10.1016/j.compbiomed.2024.109021] [Medline: 39236660]
173. Little B, Alshabrawy O, Stow D, Ferrier IN, McNaney R, Jackson DG, et al. Deep learning-based automated speech detection as a marker of social functioning in late-life depression. Psychol Med. 2021;51(9):1441-1450. [doi: 10.1017/S0033291719003994] [Medline: 31944174]
174. Faurholt-Jepsen M, Busk J, Frost M, Vinberg M, Christensen EM, Winther O, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry. 2016;6(7):e856. [FREE Full text] [doi: 10.1038/tp.2016.123] [Medline: 27434490]
175. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digit Med. 2022;5(1):11. [FREE Full text] [doi: 10.1038/s41746-021-00544-y] [Medline: 35087178]
176. No author listed. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:r340. [FREE Full text] [doi: 10.1136/bmj.r340] [Medline: 39961614]

Abbreviations

AI: artificial intelligence
ASA: automatic speech analysis
BERT: Bidirectional Encoder Representations From Transformers
DAIC-WOZ: Distress Analysis Interview Corpus–Wizard of Oz
DNN: deep neural network
DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition, Text Revision)
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition)
ICD-11: International Classification of Diseases, 11th Revision
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Extension for Diagnostic Test Accuracy
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Studies of Diagnostic Accuracy-Revised
RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach
SHAP: Shapley Additive Explanations
SVM: support vector machine
TEO: Teager Energy Operator

Edited by J Torous; submitted 23.Oct.2024; peer-reviewed by Y Khan, Q Ng; comments to author 09.Dec.2024; revised version received 05.Feb.2025; accepted 06.Aug.2025; published 22.Oct.2025

Please cite as:
Maran PL, Braquehais MD, Vlaic A, Alonzo-Castillo MT, Vendrell-Serres J, Ramos-Quiroga JA, Rodríguez-Urrutia A
Performance of Automatic Speech Analysis in Detecting Depression: Systematic Review and Meta-Analysis
JMIR Ment Health 2025;12:e67802
URL: https://mental.jmir.org/2025/1/e67802
doi: 10.2196/67802
PMID: