Review Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–14

DOI: 10.1177/20552076241261920
journals.sagepub.com/home/dhj

Sage

Use of voice features from smartphones for monitoring depressive disorders: Scoping review

Jaeeun Shin[1] and Sung Man Bae[2,3]

Abstract

Object: This review evaluates the use of smartphone-based voice data for predicting and monitoring depression.

Methods: A scoping review was conducted, examining 14 studies from Medline, Scopus, and Web of Science (2010–2023) on voice data collection methods and the use of voice features for minitoring depression.

Results: Voice data, especially prosodic features like fundamental frequency and pitch, show promise for predicting depression, though their sole predictive power requires further validation. Integrating voice with multimodal sensor data has been shown to improve accuracy significantly.

Conclusion: Smartphone-based voice monitoring offers a promising, noninvasive, and cost-effective approach to depression management. The integration of machine learning with sensor data could significantly enhance mental health monitoring, necessitating further research and longitudinal studies for validation.

Keywords

Voice features, smartphone, depressive disorders, digital health

Submission date: 9 March 2024; Acceptance date: 29 May 2024

Introduction

Depression is a common mental disorder affecting approximately 3.8% of the global population, including 5% of adults (4% of males and 6% of females), according to the World Health Organization.[1] Individuals with depression experience persistent feelings of sadness and a loss of interest or pleasure in their activities.[2] If left untreated, depression can lead to functional impairment affecting various aspects of life, including social relationships, family, work, and school. Early and objective identification of and intervention for depression can help mitigate its negative impact on individuals' overall well-being and prevent adverse effects on communities and societies.

Current mental health assessment methods rely primarily on self-reporting based on individual retrospective memory and clinical interviews. However, these methods often lead to delayed intervention, as depressive symptoms may reach a clinically significant level before therapeutic approaches are initiated, thereby hindering early intervention. Moreover, relying on retrospective memory can introduce bias, and the ecological validity of such assessments is limited.[3,4] Furthermore, among individuals undergoing treatment, there is little follow-up after sessions for psychotherapy and medication therapy.[5] As a result, patients are not monitored in a way that could help detect symptom exacerbation or prevent relapse.

Recent advancements in digital technology, such as smartphones and smartwatches, have led to increased utilization of sensor-data-based monitoring approaches in the field of mental health. By integrating the functionalities of

[1]Department of Psychology, Chung-Ang University, Seoul, Republic of Korea
[2]Department of Psychology and Psychotherapy, Dankook University, Cheonan, Republic of Korea
[3]Department of Psychology, Graduate School, Dankook University, Cheonan, Republic of Korea

Corresponding author:
Sung Man Bae, Department of Psychology and Psychotherapy, Dankook University, 119 Dandae-ro, Dongnam-gu, Cheonan, Chungnam, Republic of Korea.
Email: spirit73@hanmail.net

sensors embedded in smartphones, objective behavioral markers related to individual activities and functioning can be extracted and collected ecologically. Studies conducted thus far have demonstrated associations between depressive symptoms and levels of individual mobility, activity, and social interaction features extracted through sensors such as GPS, accelerometers, message logs, and phone call logs. These features significantly predict depressive symptoms as behavioral markers.[3,6–8]

Recently, voice features have emerged as key characteristics for distinguishing and diagnosing mental health disorders. Several studies have shown differences in linguistic patterns between individuals with and without depression, highlighting the potential of utilizing smartphones' built-in microphones as sensors to collect speech and voice patterns and investigate their relationship with depression.[9]

Previous reviews, such as those conducted by Flanagan et al.[10] and Or et al.,[11] primarily focused on the voice characteristics of patients with bipolar disorder and predicted bipolar symptoms based on these characteristics. Flanagan et al.[10] broadened their focus to mood disorders and found compelling evidence demonstrating the high potential of using smartphone voice data to monitor and detect mood disorders in real time. However, among the 13 studies included in the review by Flanagan et al.,[10] only one targeted depression. Therefore, there is a lack of review studies that consolidate information on the use of voice features to predict depression.

This literature review focuses on studies exploring the association between voice features and depression using smartphones. Specifically, the objectives of this study are to: (1) summarize the characteristics of studies using voice data to diagnose, monitor, or understand the relevance of depression using smartphone devices and (2) identify specific methods of voice data collection and extracted variables to provide objective indicators of the utility and validity of voice feature information for predicting depression.

Methods

Design

Search strategies. A scoping review is a methodology used to explore the major concepts, materials, and evidence within a specific research field and to identify gaps in research, allowing for a comprehensive overview of a broader scope of studies.[12] Our research topic is a rapidly developing area. Thus, a scoping review is appropriate to identify the key research topics and unexplored areas within this field. Furthermore, adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[13] ensures the quality of the literature review and enhances the transparency and reproducibility of the research.

The search strategies employed Medline, Scopus, and Web of Science to identify studies on depression and smartphone-based voice features, published between 2010 and October 18, 2023. The search was limited to academic sources, excluding grey literature and nonacademic databases to ensure the integrity of the search. Keywords including "depression," "depressive disorder," "smartphone," "mobile," "voice," "audio," "speech," "monitor," and "predict" were used. These keywords were expanded using the Boolean operator "OR" and the wildcard "*" where needed, and were combined with "AND" for a comprehensive and targeted search. This approach captured all relevant studies, enhancing the thoroughness and relevance of the review. The search was confined to studies published in English. Textbox 1 describes the inclusion and exclusion criteria.

Search outcomes

The study selection process is described using the PRISMA flowchart (Figure 1). After removing duplicates, decisions were made regarding inclusion or exclusion from the review. The lead author (JS) performed the screening of all titles and abstracts for 648 articles. Following a review of the titles, abstracts, and keywords of the articles, 43 studies underwent the first round of screening. Two reviewers (JS and SMB) independently engaged in the review process, retrieving and reading the full texts of 43 studies selected for in-depth review. Additionally, relevant studies that were not identified through the database searches were manually reviewed, and papers within the systematic literature review that were identified during the search process were selected to extract relevant studies.

Data extraction and analysis

The data was summarized by JS and entered into an Excel spreadsheet format. To ensure the accuracy of the process,

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria
- Use of smartphone
- Focus on diagnosing/monitoring depression
- Voice data
- English language

Exclusion criteria
- Not using a smartphone
- Focus on other mental health conditions rather than depression (e.g. bipolar disorder, neurocognitive disorder)
- Focus on depression intervention rather than prediction/monitoring
- Does not capture voice data
- Non-English language publications

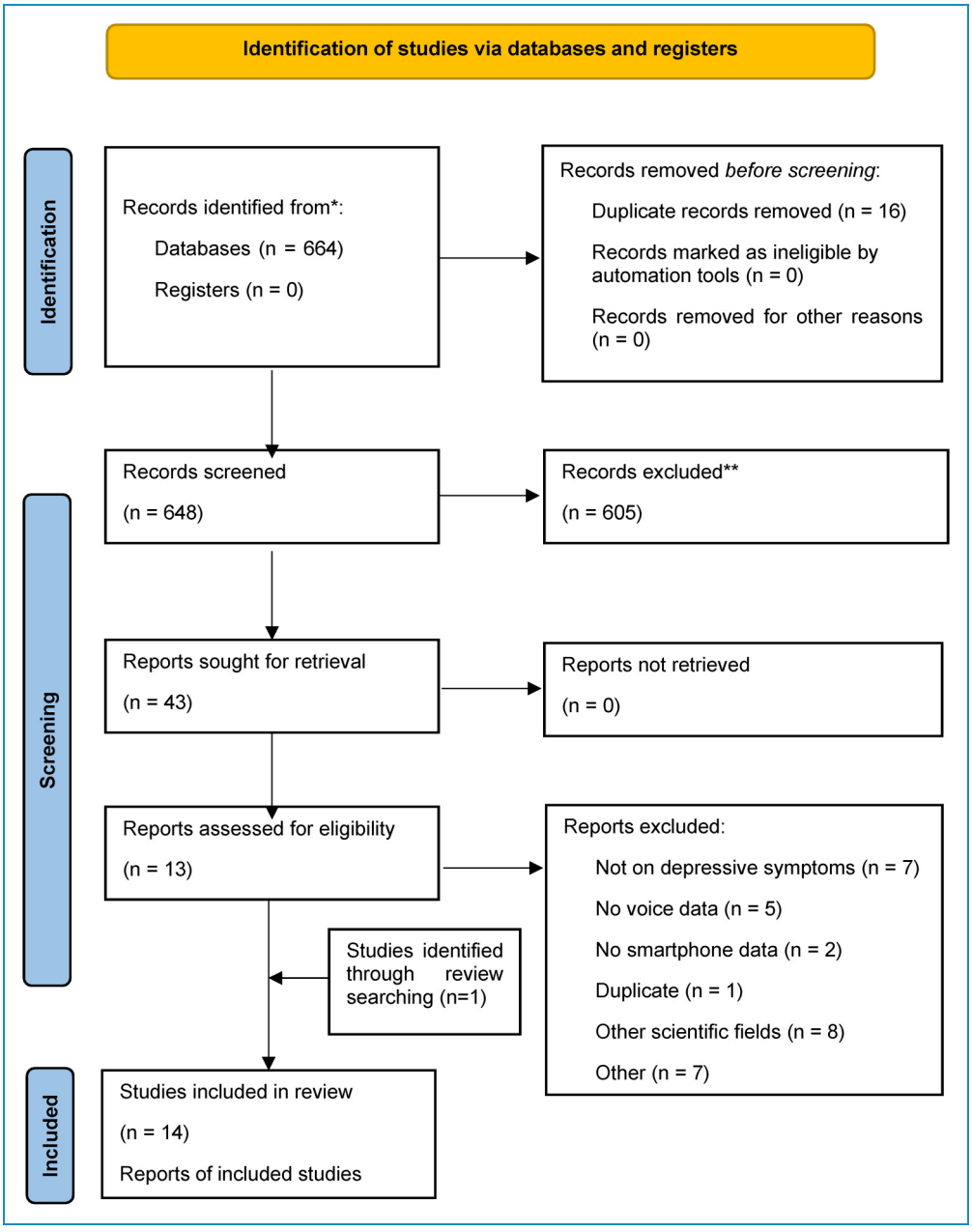

Figure 1. Flow diagram for screening and inclusion of relevant articles.

the information entered was verified by the reviewer, SMB, and no significant discrepancies or errors were found. The content of the study was systematically categorized into general characteristics (year, design, sample size, study aim), methods (duration, type of data collected, audio data processing techniques), and results (assessment tools, key findings). This categorization allowed for a detailed description of the characteristics of the studies included in the review and prepared the groundwork for comprehensive analysis. A narrative synthesis method was employed to examine and analyze the characteristics of each study, explore the methods used to analyze voice features, and evaluate the predictive efficacy of these voice features for predicting depression.

Results

Characteristics of included studies

Of the initially identified 664 studies, 13 were included in the literature review. One study was added based on a suitability judgment during the literature review process. Most excluded studies were not included because they did not involve the use of smartphone devices or did not focus on monitoring or

predicting depression. The publication years of the included studies were targeted at the past 10 years to consider recent advancements in the field. Finally, the studies from 2015 to 2022 were included. According to the Global System for Mobile Communications, the worldwide smartphone penetration rate is expected to exceed 70% in 2020, indicating a gradual increase in smartphone use. This rapid expansion is reflected in the growing number of smartphone sensor-based research studies within the last decade. Furthermore, high-quality microphones are now equipped in all smartphones, making it easier to collect voice data.[14]

General information (year published, sample size, diagnosis), study aim, data capture methods, study duration, clinical evaluations, voice data captured, and the key findings of each study are summarized in Table 1.

Characteristics of data capture

Length of voice data capture. The duration of voice data capture ranged from two weeks[7,15] to 24 months.[16]

Methods of data capture. The methods for capturing voice data can be broadly classified into three categories: First, manual capture involves activating the microphone at regular intervals (e.g. every 2 or 5 min)[6,7] to record ambient sounds or automatically record voice data during phone calls.[16–18]

Second, the method involves recording predetermined speech, which includes tasks such as counting simple text and number sequences (e.g. counting from 1 to 40) and reading emotionally neutral or stimulating passages.[15,19] In large-scale datasets collected by Sonde Health, both predetermined and spontaneous speech were included.[20–22]

Third, capturing intentional (nonpassive) spontaneous speech involves recording audio while responding to prompts. Studies employing this method required participants to speak for at least 10 s about the presented images (three positive images, three negative images, and seven neutral images) as a response to prompts.[23] Additionally, in studies in which participants were asked to provide audio samples of fixed lengths on a weekly basis, they primarily recorded audio diaries about their mental health status[24,25] or engaged in conversations via applications on topics related to health, work, and life.[26]

In summary, among the 14 studies, five employed manual capture methods, two utilized predetermined speech, and four involved spontaneous speech. Additionally, three studies used a combination of predetermined and spontaneous speech datasets.

Timing of data capture. Data can be categorized into studies that use voice data as markers to track symptoms in clinical populations and studies that use voice data as markers to monitor depressive symptoms before clinical symptom manifestation. Among the studies targeting clinical populations, one explored the response to medication in individuals diagnosed with major depressive disorder,[23] and another tracked symptoms in individuals diagnosed with depression and posttraumatic stress disorder.[25]

In studies monitoring depressive symptoms, participants included perinatal women at risk of depressive symptoms,[15] adolescents and young adults diagnosed with cancer,[26] and cohorts including individuals experiencing exam stress or unemployment.[19]

Studies have targeted both college students and general adults,[6,7,16–22] as well as a clinical population consisting of general adults and individuals with major depressive disorder, bipolar disorder, anxiety disorders, and psychiatric patients simultaneously.[24]

Voice data captured. Numerous studies have utilized prosodic features of speech. Prosodic features, such as pitch, speed, timing, and tone, have been successfully used in previous studies to distinguish between patients with and without depression. These features are typically extracted using open-source signal processing and feature extraction tools such as openSMILE and Praat. Extracting multiple features from speech data allows for their quantification and machine learning, providing insights into how accurately speech features can be used to monitor and predict depressive symptoms.[15,17–19,21,25]

Studies utilizing prosodic features have pointed out a limitation where the analysis treats all frames equally by dividing speech into fixed frames (10 to 20 ms), potentially including frames with relatively less information. Two studies utilized large data from Sonde Health to count sudden changes in articulation as speech landmark events, classifying depressed speakers.[20,22]

Some studies have conducted semantic analysis of speech using automatic speech recognition software (Google Speech-to-Text). One study analyzed the participants' words based on linguistic and psychological dimensions using Linguistic Inquiry and Word Count (Version 2015; Text Analysis Portal for Research, University of Alberta) software to assess the relevance of the detected theme words to depressive symptoms.[7] In more advanced research, both acoustic analysis of prosodic features and semantic analysis have been integrated. However, this study did not provide statistical results, including the predictive power for depressive symptoms.[26]

There were also studies estimating speaking time in everyday life[6] or calculating the proportion of silence in spoken sentences to elucidate the relationship with depressive symptoms.[16,23]

Clinical outcome measurement and key findings

Clinical outcome measurement

Most studies used self-report measures such as the Patient Health Questionnaire (PHQ),[6,7,20–22,26] Montgomery–

Table 1. Results of included studies.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
Studies presenting statistical results						
Abbas et al.	2021	18, MDD	Investigating the quantification of response to antidepressant therapy using visual and auditory digital markers.	Data collected through automated smartphone tasks assessing facial, vocal, and head movement characteristics over a four-week period (including baseline, two-week, and four-week assessments) of treatment; four weeks; the Montgomery–Åsberg Depression Rating Scale.	Speaking rate (Calculating the ratio of speech to silence within spoken utterances).	During the initial four weeks following treatment initiation, speech activity significantly increased, indicating a reduction in symptom severity assessed by clinicians using the MADRS.
Ben-Zeev et al.	2015	47, university students	Stress monitoring utilizing smartphone sensor data.	Equipped with smartphones featuring diverse sensors and software for continuous tracking of geographical activity (via GPS and WiFi usage), physical activity (using a triaxial accelerometer), sleep duration (modeled using device usage data, accelerometer, ambient sound, and ambient light levels), and time spent in proximity to human voices (i.e. utilizing microphone and voice detection algorithms); 10weeks; the Perceived Stress Scale, Patient Health Questionnaire (PHQ), the Revised UCLA Loneliness Scale.	The speaking time is calculated as the total time (in minutes) the participant spends in proximity to speech.	In panelized functional regression analysis, speech duration was significantly associated with changes in PHQ-9 scores over the study period ($p = .048$).
Braun et al.	2017	36; stressed participants for pilot study	The application of voice technology was implemented on a sample of 36 subjects in stressful environments, based on	The voice app's recording process involves three steps: (1) counting from 1 to 40, (2) reading out loud the standard reading probe, and (3) counting again from 1 to 40;	Detailed analysis was conducted using five parameters for speech behavior assessment (pause duration, speech duration, energy per speech (volume), dynamics (volume	In the analyzed data, variations in "F0 amplitude" frequently indicate the richness of voice. A narrow distribution suggests a lack of emotion and empathy, while a wider

(continued)

Table 1. Continued.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
			normative research across five major languages.	two weeks; Hamilton Depression Rating Scale (HAMD-17).	variation), energy per second), and another five parameters for evaluating voice characteristics (average voice pitch F0, F0 amplitude, F0 modulation, F0 6 dB bandwidth, and F0 contour).	distribution implies liveliness and thoughtfulness. Changes in "55–440 Hz Power" demonstrated a strong correlation ($r \geq 0.8$) with HAMD-17 scores.
De la Fuente et al.	2021	109; adults	A model for predicting impact scores based on voices collected during the COVID winter lockdown period, analyzing 109 voices.	Extracted from voice recordings collected via home and mobile devices (e.g. smartphones, tablets); during the winter COVID-19 lockdown period (no specific duration provided); emotional slider evaluations based on two dimensions of emotion: valence and arousal.	Segment-level acoustic information, encompassing extracted features like F0 pitch, loudness, spectrum flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, and slope V0, as well as commonly used statistical features, is utilized. An Active Data Representation Method (ADR) is then employed to generate data representations for each audio recording.	Using its own configuration map, ADR clusters the original acoustic features and then calculates 39 features for these clusters to extract new characteristics. It predicts the impact with the best CCC of 0.4230 (using Random Forest) for Arousal and 0.3354 (using Decision Trees) for Valence.
Di Matteo et al.	2021	86; adults	To determine a correlation between the severity of symptoms of social anxiety disorder, generalized anxiety disorder, depression, and somatic disorders, using the linguistic characteristics of manually detected words from environment audio recorded with participants' smartphones.	The app collects audio recordings for 15 s every 5 min. Transcripts of consistently captured. Audio-recorded throughout the study using automatic speech recognition software (Google Speech-to-Text) are generated consistently throughout the day over the entire duration of the study; two weeks; PHQ-8.	Language analysis, including an individual's word choice, can be related to symptoms of depression and anxiety.	The rates of words detected in the negative emotion category were most strongly correlated with the PHQ-8 ($r = 0.15$, $p = .17$). The rates of words detected in the positive emotion category were also most strongly correlated with the PHQ-8 ($r = -0.18$, $p = .09$).

(continued)

Table 1. Continued.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
Higuchi et al.	2020	1814; adults	To evaluate the potential of the Mind Monitoring System (MIMOSYS), which utilizes voice data from daily phone calls to monitor mental health, in detecting depressive states and monitoring mental changes due to stress.	Whenever a user makes a phone call with their smartphone, MIMOSYS automatically executes and records the analysis results continuously on the same server. The voice data of the calls is temporarily recorded on the smartphone; two years; Beck Depression Inventory.	Estimating energy levels based on the ratio of speech in conversation.	Consistently observed is that energy levels in patients with depression tend to be lower, while those of healthy individuals are more widely distributed. Women tend to have lower energy and mental activity levels compared to men.
Huang et al.	2022	4584 files (695 speakers) for training and 1279 files (192 speakers) for testing; N/A	To determine landmarks, sudden changes associated with speech articulation, are useful in predicting depressive symptoms.	SH2 is a subset of a large dataset collected by Sonde Health, where participants completed various speech tasks using their personal smartphones in uncontrolled natural environments. Speech tasks included speech tasks such as reading rainbow passages and Harvard sentences, as well as induced tasks like sustained vowel “ah” and diadochokinetic repetitions. For example, participants were instructed to repeat sentences from the Harvard Sentence database or freely respond to common topics such as “How is the weather outside?” for up to 30 s; N/A; PHQ-9.	The landmark method is utilized to characterize the articulatory elements of speech and detect timestamp boundaries indicating abrupt changes in speech articulation.	Landmark-based performance demonstrated performance with F1 scores of 0.77 in SH2 data’s free speech and 0.64 in SH2 data, presenting a new set of speech features for depression prediction, in addition to existing prosodic features.

(continued)

Table 1. Continued.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
Huang et al.	2020	444 files (438 speakers) for training and 130 files (128 speakers) for testing; N/A	Detecting depression through information related to vocal tract coordination (VTC) features in the voice data.	A large dataset collected by Sonde Health; N/A; PHQ-9.	Vocal Tract Coordination (VTC) features.	A new approach is proposed, suggesting the extraction of Full Vocal Tract Coordination (FVTC) features using Convolutional Neural Networks (CNNs). Assessment of the proposed FVTC-CNN architecture on depressive speech data shows an average F1 score enhancement of at least 16.4% compared to the existing VTC baseline system under clean conditions, with comparable results under noisy conditions.
Huang et al.	2020	444 files (436speakers) for training and 130files (128speakers) for testing; N/A	Assessing the prediction of depression based on landmark words.	A large dataset collected by Sonde Health; N/A; PHQ-9.	The landmark method is utilized to characterize the articulatory elements of speech and detect timestamp boundaries indicating abrupt changes in speech articulation.	Tokenizing acoustic space region into "words" representing speech events and combines these with speech landmarks for analysis, significantly improving depression detection accuracy with a 15% increase in F1 scores for the dataset.
Place et al.	2017	73; depression or PTSD patients	Exploring the relationship between passive data features and psychiatric symptoms.	Sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality; participants are instructed to leave an audio diary entry on the app at least once a week, designed as short "voice-mail" style entries about their mood or how their day is going; 12 weeks; "depressed mood" items.	The digital recordings of the audio diary are processed to extract measurements related to speech, speaking rate, intonation, prosody, and voice quality.	Depressed mood most of the day: MeanPitchVar + MeanVocalEffort + MeanVocalEffort:MeanPitchVar AUC .74.

(continued)

Table 1. Continued.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
Nonstatistical research (e.g. application development studies)						
Bilal et al.	2022	N/A; perinatal depression	To create a predictive model for identifying high-risk women with mental and physical comorbidities using digital phenotypic data from the Mom2B smartphone application.	The Mom2B app collects GPS data, timestamped smartphone connections, social media activity on smartphones, survey metadata, and voice data. For voice data collection, the app sends participants voice recording tasks every 2–4 weeks, requesting them to record brief text, number sequences, or verbal readings; N/A; Edinburgh Postnatal Depression Scale.	Using features of voice acoustics such as pitch, speed, timing, and timbre to assess voice quality.	Statistical results are not provided for the application development study.
Emden et al.	2021	Total of 997 (including major depressive disorder ($n = 409$), bipolar disorder ($n = 48$), anxiety disorder ($n = 58$), and psychotic disorder ($n = 21$)	To assess emotional symptoms by combining manual and active data formats, we developed the smartphone app Remote Monitoring Application in Psychiatry. Its validity and compliance were evaluated based on the frequency of transmitted data and the duration of participation.	Step count and distance, GPS data, accelerometer, and voice information; participation for one year is recommended; BDI.	The specific details regarding the app development study are not provided.	Statistical results are not provided for the application development study.
Lind et al.	2018	24; university students	The development of the EARS application captures various indicators of an individual's social and emotional behavior through natural usage of smartphones.	EARS application includes facial expressions, acoustic voice quality, natural language use, physical activity, music selection, and geographic location. Voice data is recorded through the device's microphone (excluding earpieces) during phone calls, encrypted, and then uploaded	Not applicable.	There is no exploration of the association between depression indicators and the validity of the EARS application's usage in the research.

(continued)

Table 1. Continued.

Study reference	Year published	Sample size and characteristics	Study aim	Method of data capture, duration, clinical evaluations	Audio data captured	Key findings
				for acoustic voice quality analysis; 1 semester; perceived stress and self-reported symptoms of mental health.		
Zhang et al.	2022	60; adolescents diagnosed with cancer	Recruitment of 60 adolescents diagnosed with cancer with the aim of monitoring psychological distress utilizing an AI-supported voice-based monitoring tool for a duration of six months.	After logging in, the patient engages in a conversation with the app on various topics such as health, work, or life for 90s. Over six months, the patient must use the EH voice tool once a month, totaling six sessions (T1–T6); 24 weeks; PHQ-8, GAD-7.	The Ellipsis Health Voice Tool integrates both semantic and acoustic information through AI-supported algorithms.	Statistical results are not provided for the application development study.

Åsberg Depression Rating Scale (SIGMA-MADRS),[23] Beck Depression Inventory,[16,24] Hospital Anxiety and Depression Scale (HADS),[15] and Hamilton Depression Rating Scale (HAMD)[19] to measure depressive symptoms. Some studies have categorized emotions into two dimensions, valence and arousal, and presented correlations using the Depression Scale (HADS).[17] Studies have measured only some aspects of depressive symptoms[25] and those with no specific information on measuring depression.[18]

Relationship between voice data and depression

Studies investigating the relationship between prosodic features of speech and depression as well as the predictive capacity of these features for depressive symptoms have yielded intriguing results. In Braun et al.'s study,[19] fluctuations in F0 amplitude were construed as indicators of speech "richness." A narrow F0 distribution suggests a lack of emotion and empathy, whereas a wider distribution indicates liveliness and thoughtfulness. They reported a close correlation, observed in 65–75% of cases, between HAMD-17 scores and speech sound characteristics, with a coefficient of $r \geq 0.8$. Place et al.[25] found that pitch information in speech data demonstrated significant predictive power for one depressive symptom, depressed mood, with an AUC value of .74. De La Fuente Garcia et al.[17] selected acoustic segments containing various features, including the F0 semi-tone, intensity, spectrum flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, and V0 slope functions, based on their theoretical significance and potential to detect physiological changes in speech production. Using an active data representation method to generate data representations for each audio recording, they calculated and extracted 39 features. These newly derived features were found to predict the influence of emotional arousal and induction with concordance correlation coefficients of 0.4230 (using random forests) and 0.3354 (using decision trees), respectively.

Furthermore, it has been suggested that the predictive power for depression in deep learning analysis improves when a sufficient length of speech data is provided, particularly when utilizing vocal tract coordination.[21] Additionally, in studies using Sonde Health (SH2) data collected via smartphones, landmark-based performance using abrupt changes in articulation as speech landmark events demonstrated performance with F1 scores of 0.77 in SH2 data's free speech and 0.64 in SH2 data, presenting a new set of speech features for depression prediction, in addition to existing prosodic features.[20,22]

Research investigating the relationship between depression and speech characteristics such as speech rate or the proportion of silence during speech has observed reduced depressive symptoms associated with increased speech activity, particularly in relation to antidepressant medication intake.[23] Moreover, the findings suggested a significant association ($p = .048$) between changes in speech duration and PHQ-9 scores during the early phase of the study (approximately 10–30 days).[6]

Studies assessing participants' vitality levels based on speech rate have consistently observed lower vitality values in patients with depression than in healthy individuals, with vitality values showing a wider dispersion among healthy individuals. Additionally, females exhibited lower vitality and mental activity than males, which is consistent with the tendency of females to have higher rates of depression.[16]

Although there are studies that provide protocols for extracting voice data to predict and monitor depressive symptoms, these studies primarily focus on introducing extraction methods without presenting statistical results regarding the relationship with depression.[15,26] Similarly, some studies have evaluated the feasibility of smartphone apps and participant compliance but lacked statistical results regarding their relationship with depression.[18,24]

Discussion

Principal findings

This review synthesizes the existing research on the use of smartphone-based voice data for predicting and monitoring depression. Collecting voice data through smartphones, which are widely used by the population, has high accessibility and the potential for real-time monitoring and prediction of individual depression.

Previous studies that reviewed the scope of smartphone voice data use for mood disorders primarily included research on bipolar disorder, resulting in limited findings on the utility of voice data for depression.[10] This scoping review included 14 studies selected from searches of major databases, such as Medline, Scopus, and Web of Science, focusing on relevant studies that utilized smartphone voice data to explore its association with depression.

Each study varied in duration (e.g. two weeks to 24 months) and method of capturing voice data (e.g. manual capture, predefined text speech, and spontaneous speech). Manual capture methods involve users activating smartphone microphones to record everyday speech or phone calls, which poses potential privacy concerns. While some studies have suggested that reading fixed texts may be more advantageous for depression prediction,[27] recent research indicates the potential applicability of spontaneous speech, such as free conversations or interviews, for more autonomous and ecological monitoring of mood disorders.[10]

Several prosodic features extracted from voice data are associated with depression. Studies on prosodic features revealed that fundamental frequency (F0), representing the vocal fold vibration rate, was significantly correlated

with depressive state,[19] and combinations of pitch information demonstrated predictive power for depression.[25] Moreover, studies also explored combining multiple prosodic features[17] or utilizing sudden changes in articulation as speech landmarks to predict depression.[20,22] Combining prosodic features to create new data representations appears more promising and valid than monitoring depression using a single dimension of vocal features because individuals naturally exhibit various changes and characteristics during speech. Therefore, it is difficult to claim that a single-level unimodal prosodic feature provides sufficient information as a clinical indicator of the broad clinical state of depression; a multilevel approach using combinations of voice data is necessary.

Some studies have examined the association between depressive symptoms and fluctuations in speech volume[6] and the ratio of silence during speech,[16,23] as well as studies extracting semantic information from speech to explore the relationship between the frequency of specific words and depressive symptoms.[7]

However, there are currently insufficient accumulated research results to conclude promising voice features for predicting depression using voice data. While some studies included in our review developed applications and proposed methods for collecting voice data to monitor depression, they did not provide statistical metrics regarding the relationship between depression and the predictive power of voice data. These studies introduced methods for collecting voice data for depression monitoring and explicitly evaluated user engagement and compliance with applications.[15,18,24,26] This suggests that depression monitoring and detection using voice data are still in the early stages. Once these applications are implemented, data are collected, and results reporting the predictive power for depression accumulate, allowing more integrated conclusions to be drawn about which voice data features are superior for predicting depression.

Furthermore, several studies included in this research not only collected voice data but also handled various sensors (e.g. GPS and accelerometers) together.[6,15,18,23–25] However, most studies have primarily examined the predictive capabilities of individual sensor data for depression in isolation, without integrating diverse sensor inputs.[6,23,25] In contrast, Hong's[28] study adopted a multimodal approach, combining activity data from location tracking, sleep patterns, and facial expression markers to predict depression. The findings revealed a significant enhancement in predictive accuracy through multimodal data analysis compared to single data analysis. While Osmani's[29] study wasn't specifically focused on depression, it investigated behavior changes in patients with bipolar disorder to see if the onset of bipolar episodes could be predicted through sensor data. The research found that combining different sensor modalities, including call sound analysis, accelerometer data, and location data, resulted in the highest accuracy (97.4%) in predicting the start of bipolar episodes.[29] Consequently, there is a pressing need for further research to explore multimodal methodologies that integrate voice data with other sensor inputs to improve depression prediction.

Implications for practice

The results of this review suggest that smartphone-based voice monitoring has a significant potential for evaluating and managing depression. When data generated by such monitoring systems are provided to users and clinicians in a timely manner, they allow immediate assessment and intervention, moving away from retrospective self-reporting and clinical interviews. Additionally, because it does not require face-to-face interaction, it offers advantages in terms of cost reduction and accessibility owing to the widespread use of smartphones.

However, although smartphone-based voice monitoring provides a certain level of statistical validity, it remains challenging to conclude whether voice data alone provide sufficient predictive power for mood disorders. Therefore, rather than using smartphone-based voice data as the sole method for clinical management, it can serve as an additional tool for clinicians to detect relapse and remission in patients.[10]

The emerging body of research underscores the potential of smartphones as tools for detecting and monitoring mental health conditions. By integrating a variety of smartphone-based sensor data, including but not limited to voice data, a more comprehensive approach to mental health monitoring may be achieved.[3,8]

Advanced machine learning techniques play a pivotal role in this integration, offering not only the means to predict mood disorders but also the ability to identify which sensor data are most valuable for the assessment and management of such disorders.[30] Particularly, the application of machine learning in conjunction with digital sensor data has shown promise in predicting mood states among individuals with mood disorders, like major depressive disorder and bipolar disorder. This integration extends to wearable, environmental, and smartphone-based passive sensors, highlighting the significance of digital markers such as location data, conversation frequency, and physical activity in understanding mood disorders.[31]

The convergence of digital technology's rapid advancement with the capabilities of machine learning algorithms heralds a new era in psychiatry and mood disorder management. By harnessing these technologies, there's a substantial opportunity for groundbreaking improvements in how mental health conditions are detected, monitored, and treated.

To further understand the link between voice data and depression, it's crucial to develop technologies that can more accurately classify voice features related to depressive

symptoms.[32] Additionally, longitudinal studies that track the effectiveness of voice-based monitoring over time are necessary.[33] This approach can lead to more consistent outcomes in identifying depression's relevance and predictability through voice data.

Limitations

This study has some limitations. First, while conducting a review of studies examining the relationship between voice data and depression using a systematic paper identification method, it is important to note that relevant research is still in its early stages, with ongoing development and pilot smartphone applications, including voice data.[15,26] Therefore, new publications might have been overlooked. Additionally, owing to the evolving nature of this field, varying terminology and a lack of conceptual consensus among researchers may have resulted in the omission of relevant papers owing to limitations in the search methods.

Second, the ethical and acceptability aspects of accessing and extracting voice data from smartphones were not reviewed. User resistance to the extraction and use of voice data can affect the practical implementation of such technologies. While some studies[19] have mentioned privacy protection measures such as avoiding collection by timestamps for privacy concerns, discussions on handling voice data are limited in most studies.

Bauer et al.[32] noted that while leveraging technology to analyze behavioral data obtained from tracking people's digital activities in mental health treatment presents a new direction for healthcare, it also poses challenges to patient privacy and security. There is a need for discussion on data handling and privacy protection within studies, and future reviews should incorporate agreed-upon methods of data collection when utilizing voice data for research purposes.

In particular, the practical implementation of voice data collection carries several privacy and ethical concerns that must be seriously considered to maximize the potential benefits and minimize risks. Slavich et al.[34]outlined requirements for minimizing risks, including: (a) clearly informing users about which devices transmit and evaluate voice data and the potential risks involved; (b) enabling users to easily activate or deactivate the listening function of devices as they wish; (c) allowing users to physically control the listening function of devices; (d) ensuring users can control who accesses their data and how it is used; (e) permitting users to utilize devices within their personal environments; and (f) providing users with the option to opt-out of voice recording or analysis.

Future research exploring the use of voice data and applications within smartphones would benefit from incorporating data handling protocols aligned with the recommendations outlined by Slavich et al.[34] This will ensure that the development and deployment of such technologies are in alignment with ethical standards and privacy considerations, thereby safeguarding user data while maximizing the utility and safety of voice data applications. It's important for researchers and developers to consider these guidelines carefully when designing their studies or applications to ensure they adhere to best practices for data privacy and security.

Conclusions

This study aims to comprehensively review the current status of research utilizing smartphone-based voice data to monitor and diagnose depression. This comprehensive review aimed to activate research that utilizes voice data as a behavioral marker for depression and influences mental health service providers to identify and respond to patients at risk of depression early. Globally, several research teams are developing smartphone-based tools for diagnosing and monitoring a wide range of mental disorders,[8] and the utilization of voice data among them could be a promising extension of existing digital phenotypic data for mental disorders.

Author contributions: Shin designed the study, analyzed the data, wrote the first draft of the manuscript, and edited the final version of the paper. Bae contributed to the design of the study, supervised the study, and edited the final version of the paper. All authors have read and agreed to the published version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (Grant No. NRF-2022S1A5A2A03050428).

Institutional Review Board Statement: The study was approved by the Institutional Review Board of Dankook University (IRB No. 2024-001).

ORCID iD: Sung Man Bae https://orcid.org/0000-0001-5762-4306

References

1. Organization. WH. Depression. [Fact sheet]. 2023. Available at: https://www.who.int/news-room/fact-sheets/detail/depression.
2. Association AP. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, DC: American psychiatric association, 2013.
3. Kamath J, Leon Barriera R, Jain N, et al. Digital phenotyping in depression diagnostics: integrating psychiatric and

engineering perspectives. *World J Psychiatry* 2022; 12: 393–409.
4. Petrizzo D and Popolo PS. Smartphone use in clinical voice recording and acoustic analysis: a literature review. *J Voice* 2021; 35: 499 e23–e28.
5. Gibbons M, Rothbard A, Farris K, et al. Changes in psychotherapy utilization among consumers of services for major depressive disorder in the community mental health system. *Adm Policy Ment Hlth* 2011; 38: 495–503.
6. Ben-Zeev D, Scherer EA, Wang R, et al. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J* 2015; 38: 218–226.
7. Di Matteo D, Wang W, Fotinos K, et al. Smartphone-detected ambient speech and self-reported measures of anxiety and depression: exploratory observational study. *JMIR Form Res* 2021; 5: e22723.
8. Dogan E, Sander C, Wagner X, et al. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. *J Med Internet Res* 2017; 19: e262.
9. Almaghrabi SA, Clark SR and Baumert M. Bio-acoustic features of depression: a review. *Biomed Signal Process Control* 2023; 85: 105020.
10. Flanagan O, Chan A, Roop P, et al. Using acoustic speech patterns from smartphones to investigate mood disorders: scoping review. *JMIR mHealth UHealth* 2021; 9: e24352.
11. Or F, Torous J and Onnela JP. High potential but limited evidence: using voice data from smartphones to monitor and diagnose mood disorders. *Psychiatr Rehabil J* 2017; 40: 320–324.
12. Arksey H and O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005; 8: 19–32.
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Br Med J* 2021; 372: n71.
14. Kappen M, Vanderhasselt M-A and Slavich GM. Speech as a promising biosignal in precision psychiatry. *Neurosci Biobehav Rev* 2023; 148: 105121.
15. Bilal AM, Fransson E, Bränn E, et al. Predicting perinatal health outcomes using smartphone-based digital phenotyping and machine learning in a prospective Swedish cohort (Mom2B): study protocol. *BMJ Open* 2022; 12: e059033.
16. Higuchi M, Nakamura M, Shinohara S, et al. Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study. *JMIR Form Res* 2020; 4: e16455.
17. De La Fuente Garcia S, Haider F and Luz S. COVID-19: Affect recognition through voice analysis during the winter lockdown in Scotland. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Mexico, 2021, pp.2326–2329. IEEE.
18. Lind MN, Byrne ML, Wicks G, et al. The effortless assessment of risk states (EARS) tool: an interpersonal approach to mobile sensing. *JMIR Ment Health* 2018; 5: e10334.
19. Braun S, Annovazzi C, Botella C, et al. Assessing chronic stress, coping skills, and mood disorders through speech analysis: a self-assessment 'voice app' for laptops, tablets, and smartphones. *Psychopathology* 2017; 49: 406–419.
20. Huang ZC, Epps J and Joachim D. Investigation of speech landmark patterns for depression detection. *IEEE Trans Affect Comput* 2022; 13: 666–679.
21. Huang ZC, Epps J and Joachim D and IEEE. Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020, pp.6549–6553 New York: IEEE.
22. Huang ZC, Epps J, Joachim D, et al. Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE J Sel Top Signal Process* 2020; 14: 435–448.
23. Abbas A, Sauder C, Yadav V, et al. Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study. *Front Digit Health* 2021; 3: 610006.
24. Emden D, Goltermann J, Dannlowski U, et al. Technical feasibility and adherence of the remote monitoring application in psychiatry (ReMAP) for the assessment of affective symptoms. *J Affect Disord* 2021; 294: 652–660.
25. Place S, Blanch-Hartigan D, Rubin C, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res* 2017; 19: e75.
26. Zhang AA, Kamat A, Acquati C, et al. Evaluating the feasibility and acceptability of an artificial-intelligence-enabled and speech-based distress screening mobile app for adolescents and young adults diagnosed with cancer: a study protocol. *Cancers (Basel)* 2022; 14: 914.
27. Kim AY, Jang EH, Lee SH, et al. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. *J Med Internet Res* 2023; 25: e34474.
28. Hong J, Kim J, Kim S, et al. Depressive symptoms feature-based machine learning approach to predicting depression using smartphone. *Healthcare* 2022; 10: 1189.
29. Osmani V. Smartphones in mental health: detecting depressive and manic episodes. *IEEE Pervasive Comput* 2015; 14: 10–13.
30. Rutledge RB, Chekroud AM and Huys QJM. Machine learning and big data in psychiatry: toward clinical applications. *Curr Opin Neurobiol* 2019; 55: 152–159.
31. Sheikh M, Qassem M and Kyriacou PA. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health* 2021; 3: 662811.
32. Pan W, Flint J, Shenhav L, et al. Re-examining the robustness of voice features in predicting depression: compared with baseline of confounders. *PLOS ONE* 2019; 14: e0218172.
33. Wang Y, Liang L, Zhang Z, et al. Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study. *Front Psychiatry* 2023; 14: 1195276.
34. Slavich GM, Taylor S and Picard RW. Stress measurement using speech: recent advancements, validation issues, and ethical and privacy considerations. *Stress* 2019; 22: 408–413.