ARTICLE OPEN

Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression

Alaa Abd-Alrazaq[1]✉, Rawan AlSaad[1,2], Farag Shuweihdi[3], Arfan Ahmed[1], Sarah Aziz[1] and Javaid Sheikh[1]

Given the limitations of traditional approaches, wearable artificial intelligence (AI) is one of the technologies that have been exploited to detect or predict depression. The current review aimed at examining the performance of wearable AI in detecting and predicting depression. The search sources in this systematic review were 8 electronic databases. Study selection, data extraction, and risk of bias assessment were carried out by two reviewers independently. The extracted results were synthesized narratively and statistically. Of the 1314 citations retrieved from the databases, 54 studies were included in this review. The pooled mean of the highest accuracy, sensitivity, specificity, and root mean square error (RMSE) was 0.89, 0.87, 0.93, and 4.55, respectively. The pooled mean of lowest accuracy, sensitivity, specificity, and RMSE was 0.70, 0.61, 0.73, and 3.76, respectively. Subgroup analyses revealed that there is a statistically significant difference in the highest accuracy, lowest accuracy, highest sensitivity, highest specificity, and lowest specificity between algorithms, and there is a statistically significant difference in the lowest sensitivity and lowest specificity between wearable devices. Wearable AI is a promising tool for depression detection and prediction although it is in its infancy and not ready for use in clinical practice. Until further research improve its performance, wearable AI should be used in conjunction with other methods for diagnosing and predicting depression. Further studies are needed to examine the performance of wearable AI based on a combination of wearable device data and neuroimaging data and to distinguish patients with depression from those with other diseases.

npj Digital Medicine (2023)6:84 ; https://doi.org/10.1038/s41746-023-00828-5

INTRODUCTION

Depression is a serious illness that affects ~3.8% of the population worldwide (i.e., 280 million people)[1]. Depression “causes feelings of sadness and/or a loss of interest in activities that were once enjoyed” and can lead to a variety of emotional and physical problems for those affected. Individuals with depression may have a decreased ability to interact and function at home and/or at work[2]. They may experience feelings of sadness, changes in appetite, altered sleep patterns, and/or feelings of fatigue. Depressed individuals may also experience feelings of worthlessness and guilt, poor concentration, and impaired decision-making, as well as being at increased risk of suicide and/or death[2]. If left untreated, it can become disabling and can lead to poor quality of life[2]. One study found that depressed adults had 28 more years of quality-adjusted life expectancy (QALE) than non-depressed adults, resulting in a 28.9-year QALE loss due to depression[3]. Therefore, it is very crucial to detect depression as soon as possible.

Current approaches for the assessment of depression disorders are primarily based on clinical observations of patients’ mental states, clinical history, and self-reported questionnaires (e.g., Patient Health Questionnaire-9 (PHQ-9)) for depression. These methods are subjective, time-consuming, and challenging to repeat. As a result, contemporary psychiatric assessments can be inaccurate and ineffective at assessing depression symptoms in a reliable and personalized manner. Furthermore, shortage of mental health professionals worldwide is one of the largest barriers to detecting depression in its early stages[2,4]. For example, there are 9 psychiatrists per 100,000 people in developed countries[5]. The situation is more concerning in middle to low-income countries, where there are 0.1 psychiatrists for every 1,000,000 in low-income countries[6]. Additionally, traditional methods of capacity building (i.e., increasing the number of trained mental health professionals) may take years to achieve[3]. Another factor that can prevent the early detection of depression is the stigma of being labelled as an individual living with a mental health disorder.

While technology has been implemented in healthcare settings with promising results, there is a need to utilize technologies to overcome the challenges of current approaches in depression assessment. Wearable devices have been one of the technologies used for detecting and predicting depression. Wearable devices are usually sensors worn by individuals to collect and analyze biomarkers or biosignals such as heart rates, physical activities, sleep patterns and quality, blood oxygen, and repository rate. Wearable devices are present in various forms such as watches, bands, jewellery, shoes, and clothing. Wearables can be classified into four categories: on-body devices (fixed directly on the body/skin), near-body devices (fixed close to the body with no direct contact with the body/skin), in-body devices (implantable electronics), and electronic textile (textiles with integrated electronics)[7]. The use of wearables has rapidly increased over the past few years; in 2020, 21% of Americans reported using a smartwatch or fitness tracker, a number which continues to grow[8]. Some countries report as high as 45% of their population using wearables[9].

Symptoms of depression can be assessed by many parameters collected by wearable devices. Due to the desire for automatic, objective, efficient, and real-time approaches to

[1]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar. [2]College of Computing and Information Technology, University of Doha for Science and Technology, Doha, Qatar. [3]School of Medicine, Leeds Institute of Health Sciences, University of Leeds, Leeds, UK. ✉email: aaa4027@qatar-med.cornell.edu

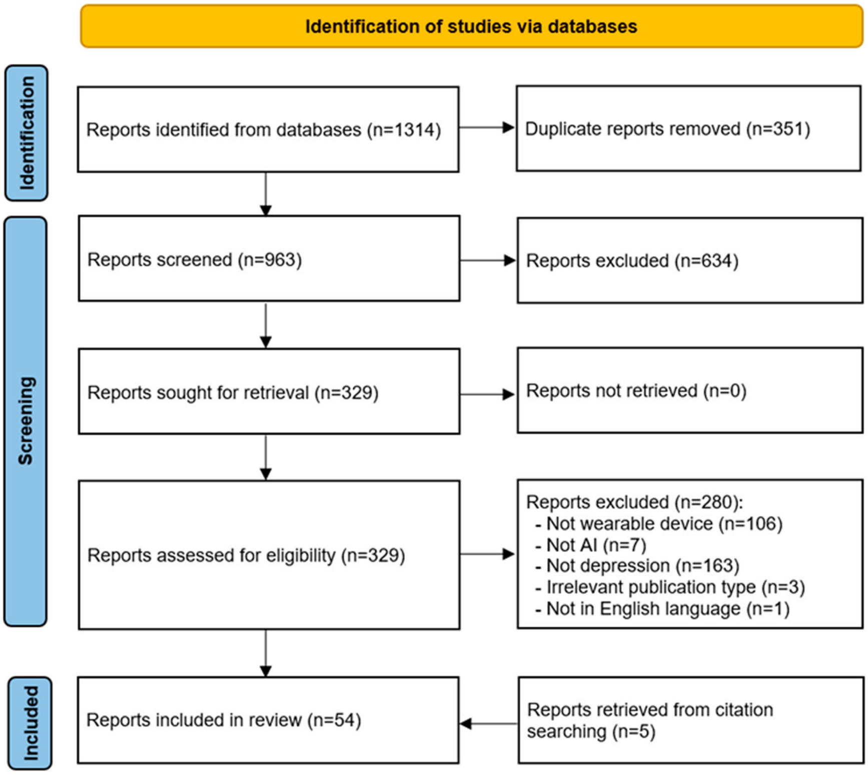

Identification of studies via databases
Identification
Screening
Included
Reports identified from databases (n=1314)
Duplicate reports removed (n=351)
Reports screened (n=963)
Reports excluded (n=634)
Reports sought for retrieval (n=329)
Reports not retrieved (n=0)
Reports assessed for eligibility (n=329)
Reports excluded (n=280):
- Not wearable device (n=106)
- Not AI (n=7)
- Not depression (n=163)
- Irrelevant publication type (n=3)
- Not in English language (n=1)
Reports included in review (n=54)
Reports retrieved from citation searching (n=5)

Fig. 1 Flow diagram of the study selection process. A total of 1314 publications were retrieved by searching all databases. Of these, 351 duplicates were removed. Screening titles and abstracts of the remaining publications led to excluding 634 citations. By reading the full text of the remaining 329 publications, we excluded 280 publications. Five additional publications were identified by checking the list of the included reviews. In total, 54 publications were included in the current review.

detect or predict depression, Artificial Intelligence (AI) has been utilized with wearable devices, introducing what we call "Wearable AI". Wearable AI refers to wearable devices that are paired with AI to analyze a large amount of wearable data and provide personalized feedback. Wearable AI has the potential to provide an early and accurate diagnosis and prediction of depression.

Numerous studies have been published on the performance of wearable devices and AI for detecting depression. Several reviews were conducted to summarize these studies; however, they had the following limitations. Firstly, they focused on wearable devices rather than wearable devices paired with AI[10–13]. Secondly, they only targeted certain age groups such as children and adolescents[11,12]. Thirdly, they did not search relevant databases such as PsychInfo[10,12,13], IEEE Xplore[10–12], and ACM Digital Library[10–13]. Fourthly, they focused on a specific type of algorithms (neural networks) and data (e.g., electroencephalogram (EEG) data[13], self-reported data[14], and neuroimaging data[15]). Lastly, and most importantly, they were not systematic reviews and did not assess the performance of the wearable AI in detecting depression via either a narrative approach or statistical approach (e.g., meta-analysis)[7,10–13]. Therefore, the need for a systematic review that focuses on the performance of wearable AI in detecting and predicting depression has never been higher. To address the above-mentioned limitations, the current review aimed at examining the performance of wearable AI in detecting and predicting depression.

RESULTS

Search results

As depicted in Fig. 1, we identified 1314 publications through searching all pre-identified databases. EndNote X9 found and removed 351 duplicates from those publications. Further 634 publications were excluded after screening titles and abstracts of the remaining 963 publications. We retrieved and read the full text of all the remaining 329 publications, and this process led to removing 280 records for several reasons shown in Fig. 1. We identified 5 additional publications relevant to this review by backward and forward reference list checking. Overall, 54 publications were included in the current review[16–69], and 38 of them were included in the meta-analyses[16–21,24–29,31,35,36,38,41,42,45–50,52,53,56–66,69].

Characteristics of included studies

The included studies were published between 2015 and 2022 (Table 1). The year in which the largest number of included studies was published was 2022 (15/54, 27.8%). Studies were carried out in 17 different countries (Table 1), and the country that published the largest number of the included studies was the United States (13/54, 24.1%). The included studies were peer-reviewed journal articles (40/54, 74.1%), conference papers (12/54, 22.2%), and theses (2/54, 3.7%).

Number of participants in the included studies ranged from 8 to 4036, with an average of 315.7 (standard deviation (SD) = 826.1) (Table 1). The mean age of participants was reported in 44 studies and ranged between 15.5 and 78 years, with an average of 39.9 (SD 13). Only 1 of the included studies targeted children (<18 years), and 3 studies focused on only older adults (≥65 years). The percentage of female participants was reported in 46 studies and varied between 2.4% and 100%, with an average of 60.2 (SD 15.2). Half of the studies (27/54, 50%) recruited both patients with depression and healthy individuals. Supplementary Table 1 shows characteristics of each included study.

Table 1. Characteristics of the included studies.

Feature	Number of studies (%)	References
Year of publication		
2022	15 (27.8)	17,22,25,26,33,34,38,41,42,47,50,53,54,57,62
2021	11 (20.4)	16,18–20,23,43,46,52,59,63,64
2020	11 (20.4)	24,28,36,37,48,49,51,56,60,61,65
2019	11 (20.4)	21,27,29,35,39,40,45,55,58,68,69
2018	4 (7.4)	30,31,44,67
2017	1 (1.9)	32
2015	1 (1.9)	66
Country of publication		
USA	13 (24.1)	20,25,32,34,35,44,50,56,57,64,67,68
Mexico	7 (13)	26,29,55,60–62,69
South Korea	7 (13)	21–23,38,39,42,51
Norway	6 (11.1)	18,27,30,31,36,40
Japan	4 (7.4)	28,53,58,65
United Kingdom	3 (5.6)	24,33,41
China	2 (3.7)	19,37
India	2 (3.7)	45,48
Switzerland	2 (3.7)	46,47
Others (Bangladesh, Finland, Italy, Netherlands, Poland, Singapore, Spain, Taiwan)	1 (each) (1.9)	16,43,49,52,54,59,63,66
Type of publication		
Journal article	40 (74.1)	16,17,19–26,29,33–37,39,41–45,47–51,54–57,60–66,68,69
Conference Paper	12 (22.2)	18,27,28,30–32,38,52,53,58,59,67
Thesis	2 (3.7)	40,46
Number of participants		
Mean (Standard Deviation)	315 (826)	16–69
Range	8-4036	16–69
Age of participants		
Mean (Standard Deviation)	39.9 (13)	16,18,21–27,29–36,38–42,45,47,49,50,52–65,67,69
Range	15.5-78	16,18,21–27,29–36,38–42,45,47,49,50,52–65,67,69
Gender (Female %)		
Mean (Standard Deviation)	60.2 (15.2)	16,18,21–27,29–42,44,47–50,52–65,67,69
Range	2.4-100	16,18,21–27,29–42,44,47–50,52–65,67,69
Health conditions[1]		
Depression	38 (70.4)	16–21,25–27,29–33,35,36,38–42,44,49–52,54–62,64,65,69
Healthy	27 (50)	16,18,20,26,27,29–31,35,36,38,40,41,44,49,51,52,54,55,57–62,65,69
Any health condition	13 (24.1)	22–24,28,34,37,46–48,53,63,67,68
Bipolar	4 (7.4)	21,42,43,66
Schizophrenia	1 (1.9)	62
Mood swings	1 (1.9)	45

Features of wearable AI

The included studies used 30 different wearable devices, but the most common wearable devices used were Actiwatch AW4 (19/54, 35.2%) and Fitbit series (e.g., Fitbit Charge, Fitbit Flex, Fitbit Altra) (14/54, 25.9%) (Table 2). The wearable devices in the included studies were worn on 8 different parts of the body, but the wrist-worn devices were most common in the included studies (50/54, 92.6%).

AI in this review was used to detect the current depression status in 48 studies or predict the occurrence or level of depression in the future based on previous and current biosignals in 6 studies (Table 2). Studies used algorithms to solve classification problems (44/54, 81.5%), regression problems (5/54, 9.3%), and both classification and regression problems (5/54, 9.3%). There were 36 different algorithms used in the included studies, but the most commonly used algorithms were Random Forest (RF) (32/54, 59.3%), Logistic Regression (LogR) (13/54, 24.1%), and Support Vector Machine (SVM) (11/54, 20.4%). The included studies used datasets from either closed sources (i.e., collected by authors of the study or obtained from previous studies) (34/54, 63%) or open sources (i.e., public databases) (20/54, 37%). Depresjon was the most common dataset obtained from open sources and used in the included studies (17/20, 85%).

The included studies used >30 types of data to develop the model (Table 2). The most common data used to develop the models were physical activity data (e.g., step counts, calories,

Table 2. Features of wearable AI.

Feature	Number of studies (%)	References
Wearable device		
Actiwatch AW4	19 (35.2)	16,18,26,27,29–31,35,36,40,41,52,55,57,59–62,69
Fitbit series	14 (25.9)	20,21,25,28,33,34,42,44,46,47,50,53,63,68
Empatica series	3 (5.6)	22,32,56
Mi Band	2 (3.7)	19,45
GENEActiv	2 (3.7)	43,49
Others	1 each (1.9)	17,23,24,38,39,48,49,51,54,58,64–67
Not reported	1 (1.9)	37
Placement		
Wrist	50 (92.6)	16–37,39–47,49–53,55–65,67–69
Head	1 (1.9)	48
Lower back	1 (1.9)	38
Fingers	1 (1.9)	54
Chest	1 (1.9)	66
Waist	1 (1.9)	23
Thigh	1 (1.9)	23
Ankle	1 (1.9)	23
Aim of AI algorithms		
Detection	48 (88.9)	16–20,22–24,26–41,43–46,48–67,69
Prediction	6 (11.1)	21,25,34,42,47,68
Problem-solving approaches		
Classification	44 (81.5)	16–26,28–31,33,34,36–43,45–49,51–55,57–62,66,68,69
Regression	5 (9.3)	32,50,56,64,67
Classification and regression	5 (9.3)	27,35,44,63,65
AI Algorithms		
Random Forest	32 (59.3)	16–19,21,22,25,28–34,36,39,42,43,45,46,48,50,51,54,55,57,59–62,64,69
Logistic Regression	13 (24.1)	16,17,19,20,23,25,39,43,45,46,49,54,67
Support Vector Machine	11 (20.4)	16,18,19,22,25,30,43,54,55,58,64
Extreme Gradient Boosting	10 (18.5)	17,18,22,23,35,46,50,54,63,65
Decision Tree	8 (14.8)	18,19,22,30,39,43,48,55
AdaBoost	8 (14.8)	20,25,30,32,50,59,64,68
Convolutional Neural Network	6 (11.1)	26,27,36,40,41,52
Ensemble model	6 (11.1)	32,45–47,52,56
K-Nearest Neighbours	6 (11.1)	17,20,22,30,54,55
Long Short-Term Memory	5 (9.3)	24,37,38,40,41
Gradient Boosting	4 (7.4)	17,20,22,64
Multilayer Perceptron	3 (5.6)	22,23,66
Artificial Neural Network	3 (5.6)	25,30,59
Naive Bayes	3 (5.6)	30,48,55
Gradient-Boosted Decision Trees	2 (3.7)	25,46
Ridge Regression	2 (3.8)	32,44
Gaussian Process	2 (3.7)	30,32
Linear regression	2 (3.7)	32,67
Deep Neural Network	2 (3.7)	31,36
elasticNet	2 (3.7)	34,64
Support Vector Classifier	2 (3.7)	17,23
least Absolute Shrinkage and Selection Operator	2 (3.7)	20,44
Others	1 each (1.9)	18,25,30,32,39,41,48,53,64
Dataset source		
Closed	34 (63)	19–22,24,25,28,30,32–34,37–39,42–51,53,54,56,58,63–68
Open	20 (37)	16–18,23,26,27,29,31,35,36,40,41,52,55,57,59–62,69
Data input to AI algorithm		
Physical activity data	47 (87)	16–22,25–27,29–45,49–65,67–69

Table 2 continued

Feature	Number of studies (%)	References
Sleep data	26 (48.1)	19–21,25,28,32–34,39,42–47,49–51,53,54,56,63–65,67,68
Heart rate data	17 (31.5)	17,19,21,22,24,25,42,44,45,50,51,53,54,56,64,65,67
Mental health measures	12 (22.2)	21,25,32,34,39,44,46,47,49,51,54,64
Smartphone usage data	9 (16.7)	19,20,32,50,51,54,56,67,68
Location data	9 (16.7)	19,20,32,44,50,54,56,67,68
Social interaction data	8 (14.8)	19,20,32,50,51,56,67,68
Light exposure	5 (9.3)	21,39,42,51,65
Demographic data	5 (9.3)	25,46,47,49,59
Electrodermal activity data	4 (7.4)	17,22,32,56
Circadian rhythms	3 (5.6)	23,49,63
Skin temperature data	2 (3.7)	22,65
Weather data	2 (3.7)	53,56
Others	1 each (1.9)	32,37,43,46–48,51,53,64,66
Ground truth assessment		
MADRS	19 (35.2)	16,18,26,27,29–31,35,36,40,41,52,55,57,59–62,69
PHQ-4, -8, and -9	14 (25.9)	19,22,23,25,33,34,46–48,50,51,63,64,67
DSM-IV and -5	5 (9.3)	21,38,44,51,66
HDRS	5 (9.3)	32,39,45,56,65
BDI-II	5 (9.3)	17,20,37,51,68
Clinical assessment	2 (3.7)	42,43
STAI	2 (3.7)	24,37
DASS	2 (3.7)	24,54
DAMS	2 (3.7)	28,53
QIDS	2 (3.7)	44,66
GDS	2 (3.7)	22,39
Others	1 each (1.9)	49,62,66
Not reported	1 (1.9)	58
Validation approach		
K-fold cross-validation	25 (46.3)	17,19,22,25–27,29,30,32,33,38,40,45–47,51–53,56,59,62,63,65–67
Hold-out cross-validation	22 (40.7)	21,23,24,26,27,29,32,37,39,41–43,45,51,52,56,58,60,61,66,68,69
Leave-one-out cross-validation	12 (22.2)	18,20,27,28,31,32,35,36,44,48,50,68
Nested cross-validation	5 (9.3)	16,34,54,57,64
External validation	2 (3.7)	49,68
Time-series cross-validation	1 (1.9)	54
Not reported	1 (1.9)	55

BDI-II Beck Depression Inventory-II, *DASS* Depression, Anxiety and Stress Scale, *DSM* Diagnostic and Statistical Manual of Mental Health, *HDRS* Hamilton Depression Rating Scale, *MADRS* Montgomery-Asberg Depression Rating Scale, *PHQ-9* Patient Health Questionnaire-9, *QIDS* Quick Inventory of Depressive Symptomatology, *STAI* State-Trait-Anxiety-Inventory.

metabolic rate) (47/54, 87%), sleep data (e.g., duration and patterns) (26/54, 48.1%), heart rate data (e.g., heart rate, heart rate variability, interbeat interval) (17/54, 31.5%), mental health measures (e.g., depression level, anxiety level, stress level, mood status) (12/54, 22.2%), smartphone usage data (e.g., display on/off, charging activity, number of apps used) (9/54, 16.7%), location data (e.g., latitude, longitude, % of time at home) (9/54, 16.7%), and social interaction (e.g., call and message logs) (8/54, 14.8%).

The included studies identified the ground truth based on 13 different tools, but the most common tool was Montgomery-Asberg Depression Rating Scale (MADRS) (19/54, 35.2%). The included studies used 6 different validation methods for the models (Table 2). The most commonly used validation methods were K-fold cross-validation (25/54, 46.3%) and hold-out cross-validation (22/54, 40.7%). Supplementary Table 2 shows features of wearable AI in each included study.

Results of risk of bias appraisal

More than two-thirds of the studies (37/54, 69%) did not provide sufficient information to verify if an appropriate consecutive or random sample of eligible patients was used. Majority of the studies (50/54, 93%) avoided inappropriate exclusions. An adequate balance in the number of patients between the subgroups was used in 73% (34/54) of the studies. Researchers have used an insufficient sample size in 44% (24/54) of the included studies. Thus, the risk of bias owing to the "selection of participants" was rated as low in only 33% (18/54) of the studies (Fig. 2). Concerns regarding the matching between the spectrum of participants and the pre-stated requirements in the review question were rated as low in 87% (47/54) of the studies (Fig. 3).

The AI models were described in detail in 72% (39/54) of the studies. The features (predictors) used in the models were clearly described in almost all studies (53/54, 98%) and were assessed in

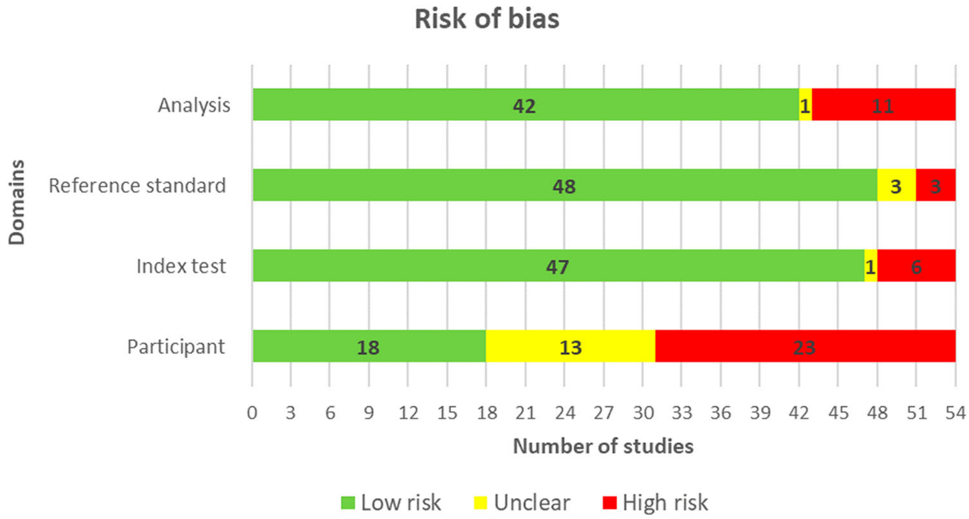


Fig. 2 Results of the assessment of risk of bias in the included studies. A modified version of QUADAS-2 was used to assess the risk of bias in the included studies in terms of 4 domains (participants, index test, reference standard, and analysis). Low risk (green) refers to the number of studies that have a low risk of bias in the respective domain. Unclear (yellow) refers to the number of studies that have an unclear risk of bias in the respective domain due to lack of information reported by the study. High risk (Red) refers to the number of studies that have a high risk of bias in the respective domain.

the same way for all participants in 94% (51/54) of the studies. Features were collected without the knowledge of outcome data in 93% (50/54) of the studies. Therefore, there was a low risk of bias because of the "index test" in 87% (47/54) of the studies (Fig. 2). All the included studies (54/54, 100%) were judged to have low concerns that the definition, assessment, or timing of predictors in the model do not match the review question (Fig. 3).

Researchers in 98% (53/54) of the studies assessed the outcome of interest (i.e., depression level) using appropriate tools. In 94% (51/54) of the studies, the outcome was defined in a similar way for all participants and was determined without knowledge of predictor information. However, only 10 studies (19%) used an appropriate interval between the index test and the reference standard. According to these judgments, the risk of bias because of the "reference standard" was low in 89% (48/54) of the studies (Fig. 2). Nearly all studies (53/54, 98%) were judged to have low concerns that the outcome definition, timing, or determination do not match the review question (Fig. 3).

All participants enroled in the study were included in the data analysis in 65% (35/54) of the studies. In 94% (51/54) of the studies, the data preprocessing was carried out appropriately and the breakdown of the training, validation, and test sets was adequate. The performance of the model was evaluated using suitable measures in 85% (46/54) of the studies. Accordingly, 78% (42/54) of the studies had a low risk of bias in the analysis domain (Fig. 2). Supplementary Table 3 shows reviewers' judgments about each domain in "risk of bias" and "applicability concerns" for each included study.

Results of the studies

Meta-analyses were carried out for the highest and lowest results of 4 measures: accuracy, sensitivity, specificity, and RMSE. Further, when applicable, subgroups meta-analyses were performed to assess the performance of wearable AI based on different AI algorithms, aims of AI, used wearable devices, data sources, types of data, and reference standards. The following sections show the above-mentioned results.

Accuracy. Wearable AI accuracy, which is the ability of the AI to correctly classify patients with and without depression, was reported in 35 studies. We identified 75 estimates of accuracy from these studies as many of them reported accuracy for more than one algorithm. The highest accuracy in these studies ranged from 0.56 to 1.00. As presented in Fig. 4, a meta-analysis of the 75 estimates from 249,203 participants in the 35 studies showed a pooled mean accuracy of 0.89 (95% confidence interval (CI) 0.83 to 0.93). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 99.5\%$). As shown in Supplementary Table 4, subgroup analyses revealed that there is no statistically significant difference in the highest accuracy between subgroups in all groups except for the "algorithms" group (Cochran's $p < 0.001$).

We extracted 39 estimates of the lowest accuracy from 24 studies. The lowest accuracy estimates ranged between 0.20 and 1.00. As demonstrated in Fig. 5, a meta-analysis of the 39 estimates of the lowest accuracy from 44,846 participants in the 24 studies showed a pooled mean of 0.70 (95% CI 0.62 to 0.78). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 98.9\%$). As shown in Supplementary Table 5, subgroup analyses revealed that there is no statistically significant difference in the lowest accuracy between subgroups in all groups except for the "algorithms" group (Cochran's $p < 0.001$).

Sensitivity. The wearable AI sensitivity, which is the ability of the AI to correctly detect patients with depression, was reported in 29 studies. We identified 58 estimates of sensitivity from these studies because many of them reported sensitivity for more than one algorithm. The highest sensitivity in these studies ranged from 0.53 to 1.00. As presented in Fig. 6, a meta-analysis of the 58 estimates from 54,169 participants in the 29 studies showed a pooled mean sensitivity of 0.87 (95% CI 0.79 to 0.92). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 98.1\%$). As exhibited in Supplementary Table 6, subgroup analyses revealed that there is no statistically significant difference in the highest sensitivity between subgroups in all groups except for the "algorithms" group (Cochran's $p = 0.002$).

We extracted 30 estimates of the lowest sensitivity from 21 studies. The lowest sensitivity estimates ranged between 0.00 and 0.98. As demonstrated in Fig. 7, a meta-analysis of the 30 estimates of the lowest sensitivity from 13,015 participants in the 21 studies showed a pooled mean of 0.61 (95% CI 0.49 to 0.72).

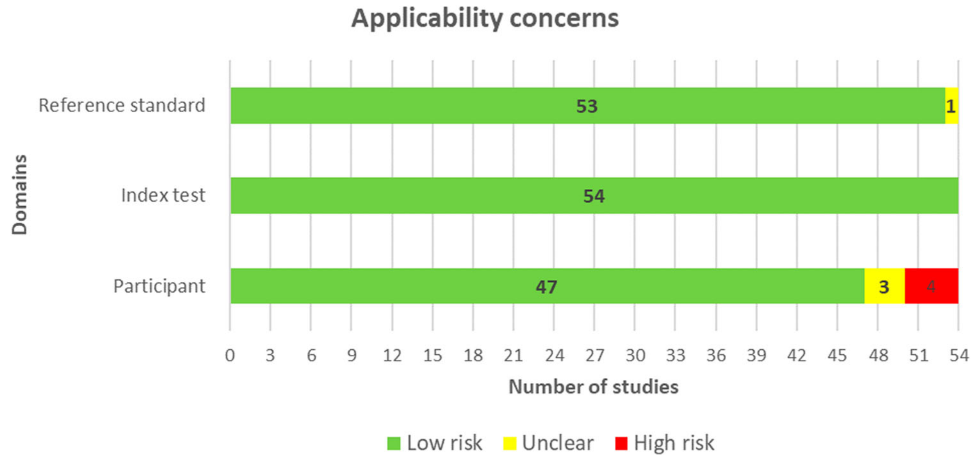


Fig. 3 Results of the assessment of applicability concerns in the included studies. A modified version of QUADAS-2 was used to assess the applicability concerns in the included studies in terms of 3 domains (participants, index test, and reference standard). Low risk (green) refers to the number of studies that have a low risk of applicability concerns in the respective domain. Unclear (yellow) refers to the number of studies that have an unclear risk of applicability concerns in the respective domain due to lack of information reported by the study. High risk (Red) refers to the number of studies that have a high risk of applicability concerns in the respective domain.

The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 98.6\%$). As shown in Supplementary Table 7, subgroup analyses revealed that there is no statistically significant difference in the lowest sensitivity between subgroups in all groups except for the "wearable devices" group (Cochran's $p = 0.038$).

Specificity. The wearable AI specificity, which is the ability of the AI to correctly detect patients without depression, was reported in 28 studies. We identified 54 estimates of specificity from these studies given that many of them reported specificity for more than one algorithm. The highest specificity in these studies ranged from 0.51 to 1.00. As presented in Fig. 8, a meta-analysis of the 54 estimates from 157,576 participants in the 28 studies showed a pooled mean specificity of 0.93 (95% CI 0.87 to 0.97). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 99.6\%$). As shown in Supplementary Table 8, subgroup analyses revealed that there is no statistically significant difference in the highest specificity between subgroups in all groups except for the "algorithms" group (Cochran's $p = 0.042$).

We extracted 27 estimates of the lowest specificity from 20 studies. The lowest specificity estimates ranged between 0.25 and 0.99. As demonstrated in Fig. 9, a meta-analysis of the 27 estimates of the lowest specificity from 26,654 participants in the 20 studies showed a pooled mean of 0.73 (95% CI 0.62 to 0.82). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 98.6\%$). As shown in Supplementary Table 9, subgroup analyses revealed that there is no statistically significant difference in the lowest specificity between subgroups in all groups except for the "algorithms" group (Cochran's $p < 0.001$) and the "wearable devices" group (Cochran's $p = 0.038$).

Root Mean Suare Error (RMSE). The wearable AI RMSE, which estimates the average difference between depression scores predicted by wearable AI and the actual depression scores as assessed by depression assessment tools (e.g., PHQ-9 and HDRS), was reported in 3 studies. We identified 5 estimates of the RMSE from these studies given that one study reported RMSE for 3 algorithms. The highest RMSE in these studies ranged from 3.2 to 6.00. As presented in Fig. 10, a meta-analysis of the 5 estimates from 1,705 participants in the 3 studies showed a pooled mean RMSE of 4.55 (95% CI 3.05 to 6.05). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 100\%$).

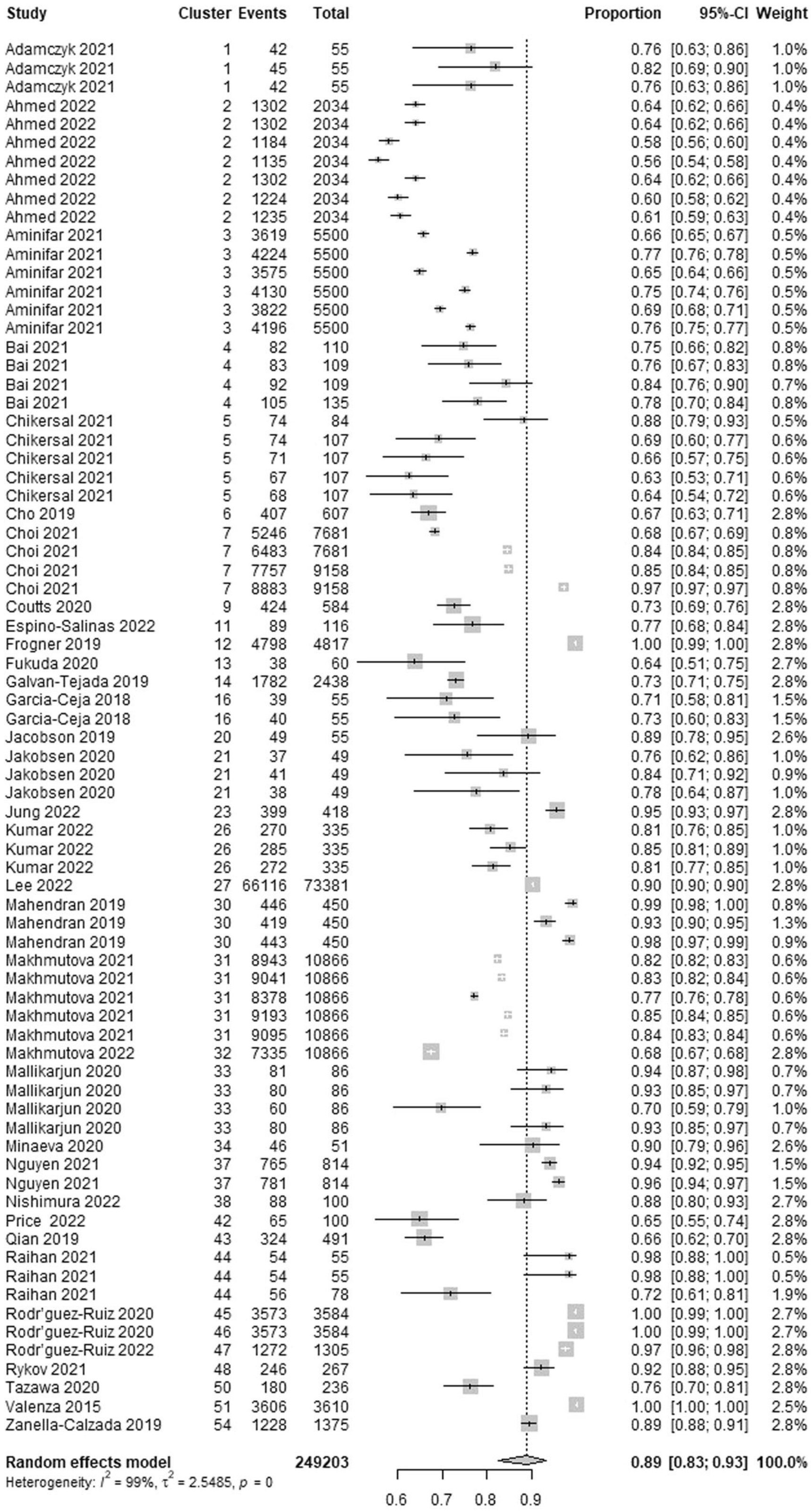
We extracted 5 estimates of the lowest RMSE from 3 studies. The lowest RMSE estimates ranged between 0.11 and 1.16. As shown in Fig. 11, a meta-analysis of the 5 estimates of the lowest RMSE from 1,705 participants in the 3 studies showed a pooled mean RMSE of 3.76 (95% CI 2.45 to 5.07). The statistical heterogeneity of the evidence was considerable (Cochran's $p < 0.001$; $I^2 = 99.9\%$).

DISCUSSION

This review examined the performance of wearable AI in detecting and predicting depression. Meta-analyses of estimates from 38 studies revealed AI has a good performance in diagnosing depression using wearable device data, but it is not optimal. Specifically, this review showed that AI could correctly classify patients with and without depression in between 70% from 89% of cases. The review demonstrated that AI has a slightly higher performance in detecting patients without depression (73–93%) than patients with depression (61%-87%). Similarly, this review found that AI has good performance in predicting depression scores using wearable device data, but it is not optimal (RMSE 3.76-4.55).

Subgroup analyses in this review showed that the performance of wearable AI is statistically different between algorithms. To be more precise, AdaBoost outperformed all other algorithms in most analyses. In contrast, logistic regression and decision trees were the worst in most analyses. These results should be interpreted with caution as most of the pooled estimates of the above-mentioned algorithms were based on a few studies (i.e., ≥4) and small sample sizes. Some subgroup analyses found that the performance of wearable AI is affected by the wearable device used to collect data. Specifically, wearable AI has better performance when data is collected by Actiwatch in comparison with Fitbit. This finding should also be interpreted carefully because all studies that used Actiwatch are based on the same dataset (i.e., Depresjon[30]). None of the subgroup analyses showed a statistically significant difference between subgroups in the remaining groups (i.e., aims of AI, data sources, data types, and reference standards).

Similar to the current review, two previous systematic reviews showed that AI has a slightly higher performance in detecting patients without depression (specificity) than patients with depression (sensitivity)[14,15]. However, the two reviews showed pooled sensitivity (80%[14] and 77%[15]) and specificity (85%[14] and 78%[15]) that are slightly lower than those in the current review



Study	Cluster	Events	Total	Proportion	95%-CI	Weight
Adamczyk 2021	1	42	55	0.76	[0.63; 0.86]	1.0%
Adamczyk 2021	1	45	55	0.82	[0.69; 0.90]	1.0%
Adamczyk 2021	1	42	55	0.76	[0.63; 0.86]	1.0%
Ahmed 2022	2	1302	2034	0.64	[0.62; 0.66]	0.4%
Ahmed 2022	2	1302	2034	0.64	[0.62; 0.66]	0.4%
Ahmed 2022	2	1184	2034	0.58	[0.56; 0.60]	0.4%
Ahmed 2022	2	1135	2034	0.56	[0.54; 0.58]	0.4%
Ahmed 2022	2	1302	2034	0.64	[0.62; 0.66]	0.4%
Ahmed 2022	2	1224	2034	0.60	[0.58; 0.62]	0.4%
Ahmed 2022	2	1235	2034	0.61	[0.59; 0.63]	0.4%
Aminifar 2021	3	3619	5500	0.66	[0.65; 0.67]	0.5%
Aminifar 2021	3	4224	5500	0.77	[0.76; 0.78]	0.5%
Aminifar 2021	3	3575	5500	0.65	[0.64; 0.66]	0.5%
Aminifar 2021	3	4130	5500	0.75	[0.74; 0.76]	0.5%
Aminifar 2021	3	3822	5500	0.69	[0.68; 0.71]	0.5%
Aminifar 2021	3	4196	5500	0.76	[0.75; 0.77]	0.5%
Bai 2021	4	82	110	0.75	[0.66; 0.82]	0.8%
Bai 2021	4	83	109	0.76	[0.67; 0.83]	0.8%
Bai 2021	4	92	109	0.84	[0.76; 0.90]	0.7%
Bai 2021	4	105	135	0.78	[0.70; 0.84]	0.8%
Chikersal 2021	5	74	84	0.88	[0.79; 0.93]	0.5%
Chikersal 2021	5	74	107	0.69	[0.60; 0.77]	0.6%
Chikersal 2021	5	71	107	0.66	[0.57; 0.75]	0.6%
Chikersal 2021	5	67	107	0.63	[0.53; 0.71]	0.6%
Chikersal 2021	5	68	107	0.64	[0.54; 0.72]	0.6%
Cho 2019	6	407	607	0.67	[0.63; 0.71]	2.8%
Choi 2021	7	5246	7681	0.68	[0.67; 0.69]	0.8%
Choi 2021	7	6483	7681	0.84	[0.84; 0.85]	0.8%
Choi 2021	7	7757	9158	0.85	[0.84; 0.85]	0.8%
Choi 2021	7	8883	9158	0.97	[0.97; 0.97]	0.8%
Coutts 2020	9	424	584	0.73	[0.69; 0.76]	2.8%
Espino-Salinas 2022	11	89	116	0.77	[0.68; 0.84]	2.8%
Frogner 2019	12	4798	4817	1.00	[0.99; 1.00]	2.8%
Fukuda 2020	13	38	60	0.64	[0.51; 0.75]	2.7%
Galvan-Tejada 2019	14	1782	2438	0.73	[0.71; 0.75]	2.8%
Garcia-Ceja 2018	16	39	55	0.71	[0.58; 0.81]	1.5%
Garcia-Ceja 2018	16	40	55	0.73	[0.60; 0.83]	1.5%
Jacobson 2019	20	49	55	0.89	[0.78; 0.95]	2.6%
Jakobsen 2020	21	37	49	0.76	[0.62; 0.86]	1.0%
Jakobsen 2020	21	41	49	0.84	[0.71; 0.92]	0.9%
Jakobsen 2020	21	38	49	0.78	[0.64; 0.87]	1.0%
Jung 2022	23	399	418	0.95	[0.93; 0.97]	2.8%
Kumar 2022	26	270	335	0.81	[0.76; 0.85]	1.0%
Kumar 2022	26	285	335	0.85	[0.81; 0.89]	1.0%
Kumar 2022	26	272	335	0.81	[0.77; 0.85]	1.0%
Lee 2022	27	66116	73381	0.90	[0.90; 0.90]	2.8%
Mahendran 2019	30	446	450	0.99	[0.98; 1.00]	0.8%
Mahendran 2019	30	419	450	0.93	[0.90; 0.95]	1.3%
Mahendran 2019	30	443	450	0.98	[0.97; 0.99]	0.9%
Makhmutova 2021	31	8943	10866	0.82	[0.82; 0.83]	0.6%
Makhmutova 2021	31	9041	10866	0.83	[0.82; 0.84]	0.6%
Makhmutova 2021	31	8378	10866	0.77	[0.76; 0.78]	0.6%
Makhmutova 2021	31	9193	10866	0.85	[0.84; 0.85]	0.6%
Makhmutova 2021	31	9095	10866	0.84	[0.83; 0.84]	0.6%
Makhmutova 2022	32	7335	10866	0.68	[0.67; 0.68]	2.8%
Mallikarjun 2020	33	81	86	0.94	[0.87; 0.98]	0.7%
Mallikarjun 2020	33	80	86	0.93	[0.85; 0.97]	0.7%
Mallikarjun 2020	33	60	86	0.70	[0.59; 0.79]	1.0%
Mallikarjun 2020	33	80	86	0.93	[0.85; 0.97]	0.7%
Minaeva 2020	34	46	51	0.90	[0.79; 0.96]	2.6%
Nguyen 2021	37	765	814	0.94	[0.92; 0.95]	1.5%
Nguyen 2021	37	781	814	0.96	[0.94; 0.97]	1.5%
Nishimura 2022	38	88	100	0.88	[0.80; 0.93]	2.7%
Price 2022	42	65	100	0.65	[0.55; 0.74]	2.8%
Qian 2019	43	324	491	0.66	[0.62; 0.70]	2.8%
Raihan 2021	44	54	55	0.98	[0.88; 1.00]	0.5%
Raihan 2021	44	54	55	0.98	[0.88; 1.00]	0.5%
Raihan 2021	44	56	78	0.72	[0.61; 0.81]	1.9%
Rodr'guez-Ruiz 2020	45	3573	3584	1.00	[0.99; 1.00]	2.7%
Rodr'guez-Ruiz 2020	46	3573	3584	1.00	[0.99; 1.00]	2.7%
Rodr'guez-Ruiz 2022	47	1272	1305	0.97	[0.96; 0.98]	2.8%
Rykov 2021	48	246	267	0.92	[0.88; 0.95]	2.8%
Tazawa 2020	50	180	236	0.76	[0.70; 0.81]	2.8%
Valenza 2015	51	3606	3610	1.00	[1.00; 1.00]	2.5%
Zanella-Calzada 2019	54	1228	1375	0.89	[0.88; 0.91]	2.8%
Random effects model			**249203**	**0.89**	**[0.83; 0.93]**	**100.0%**

Heterogeneity: $I^2 = 99\%$, $\tau^2 = 2.5485$, $p = 0$

Fig. 4 Meta-analysis of the highest accuracy estimates. A total of 75 estimates of the highest accuracy from 35 studies were used in this meta-analysis. The square shape represents the highest accuracy in each study. The rhombus shape represents the pooled estimates of the highest accuracy in all studies. CI Confidence interval. p *p*-value.

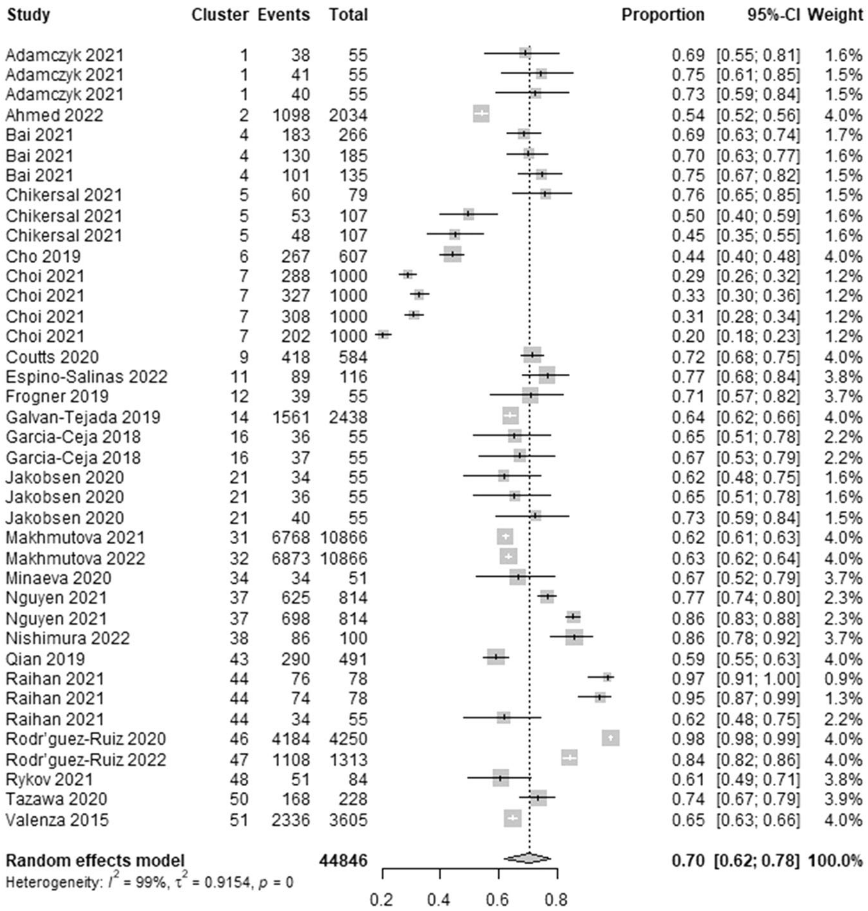


Fig. 5 **Meta-analysis of the lowest accuracy estimates.** A total of 39 estimates of the lowest accuracy from 24 studies were used in this meta-analysis. The square shape represents the lowest accuracy in each study. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. CI Confidence interval. p *p*-value.

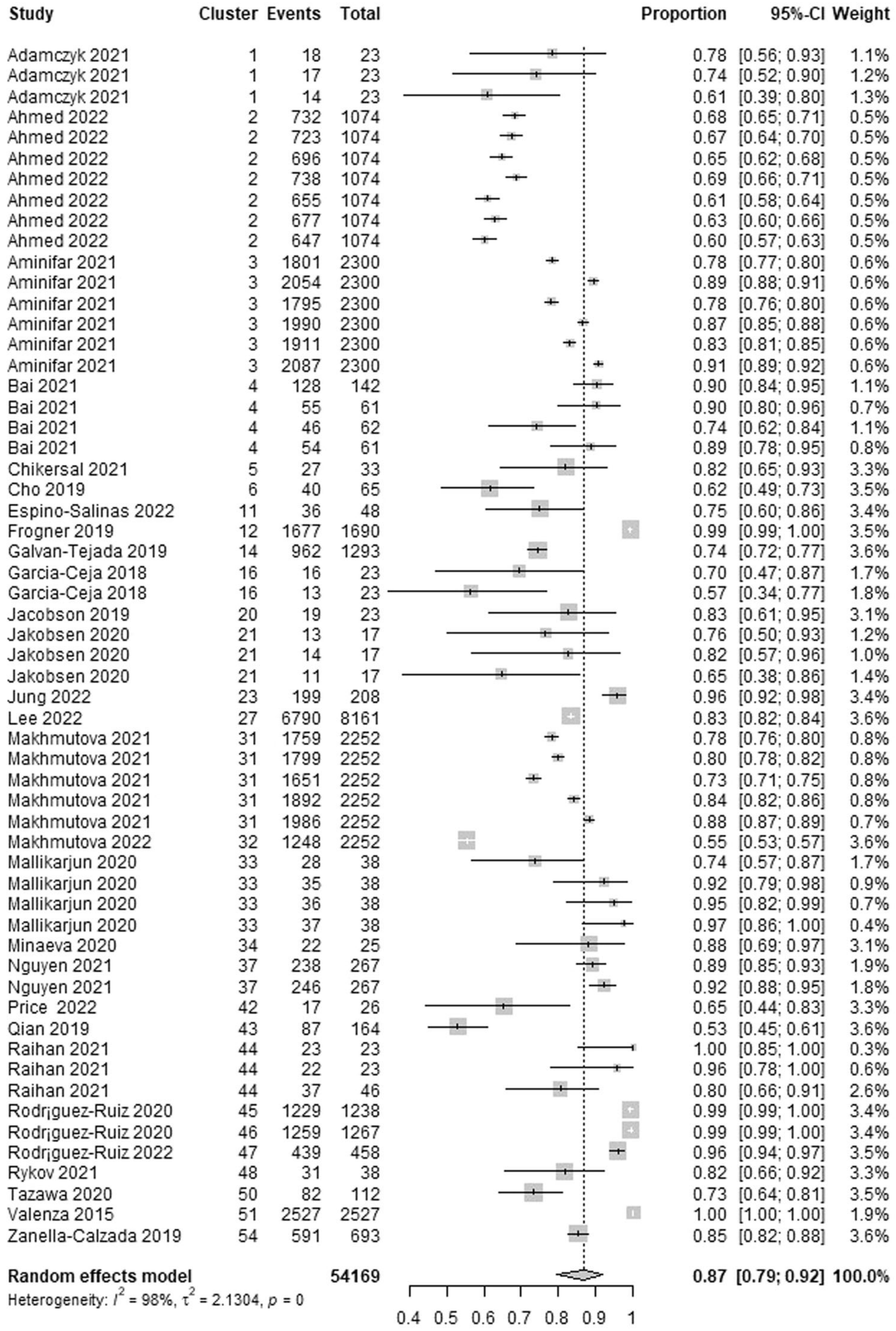
although they were within the range reported in our review. This may be attributed to the fact that the previous reviews focused on the performance of AI based on only self-reported data collected using mobile-based PHQ-9[14] or neuroimaging data[15].

From the findings of this review and previous reviews, it seems that AI has a better performance in detecting and predicting depression than in predicting treatment responses in depression. Specifically, a systematic review conducted by Cohen et al. [70] found an overall area under the curve of 84%, sensitivity of 77%, and specificity of 79% for AI in predicting response to antidepressant treatment using magnetic resonance imaging (MRI). Another review reported a pooled accuracy of 82% for AI in predicting the outcome of different therapeutic interventions (pharmacological, neuromodulatory, or manual-based psychotherapeutic interventions) using different data types (neuroimaging data, genetic data, and phenomenological data)[71]. A review carried out by Watts et al. [72] found a pooled accuracy of 84% for AI in predicting response to pharmacological and nonpharmacological interventions using EEG data. One rationale for AI's higher performance in detecting and predicting depression rather than predicting treatment responses might be the present research emphasis on diagnostic and predictive analysis of depression rather than prescriptive analysis of depression treatment in this area. More focus should be placed on prescriptive analytic research using wearable devices since these are the gadgets that patients can quickly examine and can cure or reduce the severity of depression on the spot without causing serious effects.

The current review showed that wearable AI is a promising tool for detecting and predicting depression. However, we cannot advocate that wearable AI is ready to be implemented in clinical practices for the following reasons: (1) its performance is not optimal at the present, thus, there is still room for improvement, (2) the sample size was small (≤55) in more than half of the studies (57.4%), (3) about 37% of the studies used publicly available datasets; especially Depresjon, and (4) few studies were judged to have a low risk of bias in all domains. Therefore, wearable AI should be used in conjunction with other methods for diagnosing and predicting depression, such as self-report questionnaires or interviews, to provide a more comprehensive understanding of a patient's condition.

In this review, AI was not embedded in any of the commercial wearable devices; instead, AI was embedded in a host device (e.g., computers) where the data collected by wearable devices is stored. Thus, we encourage tech companies to develop wearable devices that can detect and predict depression immediately as those that can detect stress (e.g., Fitbit Charge 5, Garmin Instinct Solar 2, Apple Watch Series 7, and Samsung Galaxy Watch 4). We envisage that this could happen in the near future especially as the computing power of wearables increases as new chips are developed and the tech improves. This may encourage researchers to conduct more studies in this area.

None of the included studies used neuroimaging data in addition to wearable device data to detect or predict depression. Several studies showed that AI has a high diagnostic performance (ranging from 92% to 98%) when using neuroimaging data (e.g., diffusion tensor imaging and functional and structural magnetic resonance imaging)[73–77]. Accordingly, future research vistas are to assess the performance of wearable AI in the detection and



Study	Cluster	Events	Total	Proportion	95%-CI	Weight
Adamczyk 2021	1	18	23	0.78	[0.56; 0.93]	1.1%
Adamczyk 2021	1	17	23	0.74	[0.52; 0.90]	1.2%
Adamczyk 2021	1	14	23	0.61	[0.39; 0.80]	1.3%
Ahmed 2022	2	732	1074	0.68	[0.65; 0.71]	0.5%
Ahmed 2022	2	723	1074	0.67	[0.64; 0.70]	0.5%
Ahmed 2022	2	696	1074	0.65	[0.62; 0.68]	0.5%
Ahmed 2022	2	738	1074	0.69	[0.66; 0.71]	0.5%
Ahmed 2022	2	655	1074	0.61	[0.58; 0.64]	0.5%
Ahmed 2022	2	677	1074	0.63	[0.60; 0.66]	0.5%
Ahmed 2022	2	647	1074	0.60	[0.57; 0.63]	0.5%
Aminifar 2021	3	1801	2300	0.78	[0.77; 0.80]	0.6%
Aminifar 2021	3	2054	2300	0.89	[0.88; 0.91]	0.6%
Aminifar 2021	3	1795	2300	0.78	[0.76; 0.80]	0.6%
Aminifar 2021	3	1990	2300	0.87	[0.85; 0.88]	0.6%
Aminifar 2021	3	1911	2300	0.83	[0.81; 0.85]	0.6%
Aminifar 2021	3	2087	2300	0.91	[0.89; 0.92]	0.6%
Bai 2021	4	128	142	0.90	[0.84; 0.95]	1.1%
Bai 2021	4	55	61	0.90	[0.80; 0.96]	0.7%
Bai 2021	4	46	62	0.74	[0.62; 0.84]	1.1%
Bai 2021	4	54	61	0.89	[0.78; 0.95]	0.8%
Chikersal 2021	5	27	33	0.82	[0.65; 0.93]	3.3%
Cho 2019	6	40	65	0.62	[0.49; 0.73]	3.5%
Espino-Salinas 2022	11	36	48	0.75	[0.60; 0.86]	3.4%
Frogner 2019	12	1677	1690	0.99	[0.99; 1.00]	3.5%
Galvan-Tejada 2019	14	962	1293	0.74	[0.72; 0.77]	3.6%
Garcia-Ceja 2018	16	16	23	0.70	[0.47; 0.87]	1.7%
Garcia-Ceja 2018	16	13	23	0.57	[0.34; 0.77]	1.8%
Jacobson 2019	20	19	23	0.83	[0.61; 0.95]	3.1%
Jakobsen 2020	21	13	17	0.76	[0.50; 0.93]	1.2%
Jakobsen 2020	21	14	17	0.82	[0.57; 0.96]	1.0%
Jakobsen 2020	21	11	17	0.65	[0.38; 0.86]	1.4%
Jung 2022	23	199	208	0.96	[0.92; 0.98]	3.4%
Lee 2022	27	6790	8161	0.83	[0.82; 0.84]	3.6%
Makhmutova 2021	31	1759	2252	0.78	[0.76; 0.80]	0.8%
Makhmutova 2021	31	1799	2252	0.80	[0.78; 0.82]	0.8%
Makhmutova 2021	31	1651	2252	0.73	[0.71; 0.75]	0.8%
Makhmutova 2021	31	1892	2252	0.84	[0.82; 0.86]	0.8%
Makhmutova 2021	31	1986	2252	0.88	[0.87; 0.89]	0.7%
Makhmutova 2022	32	1248	2252	0.55	[0.53; 0.57]	3.6%
Mallikarjun 2020	33	28	38	0.74	[0.57; 0.87]	1.7%
Mallikarjun 2020	33	35	38	0.92	[0.79; 0.98]	0.9%
Mallikarjun 2020	33	36	38	0.95	[0.82; 0.99]	0.7%
Mallikarjun 2020	33	37	38	0.97	[0.86; 1.00]	0.4%
Minaeva 2020	34	22	25	0.88	[0.69; 0.97]	3.1%
Nguyen 2021	37	238	267	0.89	[0.85; 0.93]	1.9%
Nguyen 2021	37	246	267	0.92	[0.88; 0.95]	1.8%
Price 2022	42	17	26	0.65	[0.44; 0.83]	3.3%
Qian 2019	43	87	164	0.53	[0.45; 0.61]	3.6%
Raihan 2021	44	23	23	1.00	[0.85; 1.00]	0.3%
Raihan 2021	44	22	23	0.96	[0.78; 1.00]	0.6%
Raihan 2021	44	37	46	0.80	[0.66; 0.91]	2.6%
Rodrıguez-Ruiz 2020	45	1229	1238	0.99	[0.99; 1.00]	3.4%
Rodrıguez-Ruiz 2020	46	1259	1267	0.99	[0.99; 1.00]	3.4%
Rodrıguez-Ruiz 2022	47	439	458	0.96	[0.94; 0.97]	3.5%
Rykov 2021	48	31	38	0.82	[0.66; 0.92]	3.3%
Tazawa 2020	50	82	112	0.73	[0.64; 0.81]	3.5%
Valenza 2015	51	2527	2527	1.00	[1.00; 1.00]	1.9%
Zanella-Calzada 2019	54	591	693	0.85	[0.82; 0.88]	3.6%
Random effects model			**54169**	**0.87**	**[0.79; 0.92]**	**100.0%**

Heterogeneity: $I^2 = 98\%$, $\tau^2 = 2.1304$, $p = 0$

Fig. 6 Meta-analysis of the highest sensitivity estimates. A total of 58 estimates of the highest sensitivity from 29 studies were used in this meta-analysis. The square shape represents the highest sensitivity in each study. The rhombus shape represents the pooled estimates of the highest sensitivity in all studies. CI Confidence interval. p *p*-value.

prediction of depression based on a combination of wearable device data and neuroimaging data.

Most studies (89%) in this review used AI for detecting the current depression status rather than predicting the occurrence or level of depression in the future. Prediction of depression is as important as, or even more important than, detection of depression as this will enable the development of early mental health warning systems and more effective, timely interventions targeted to the individual. Therefore, we urge researchers to conduct further studies on wearable AI for predicting depression.

We noticed that only a few studies in this review used wearable AI to differentiate depression from other disorders (e.g., bipolar, schizophrenia, anxiety, and stress). In clinical practice, complex and error-prone diagnostic processes are usually used to differentiate between various patient groups rather than solely distinguishing them from healthy individuals. Further studies should be conducted to distinguish patients with depression from

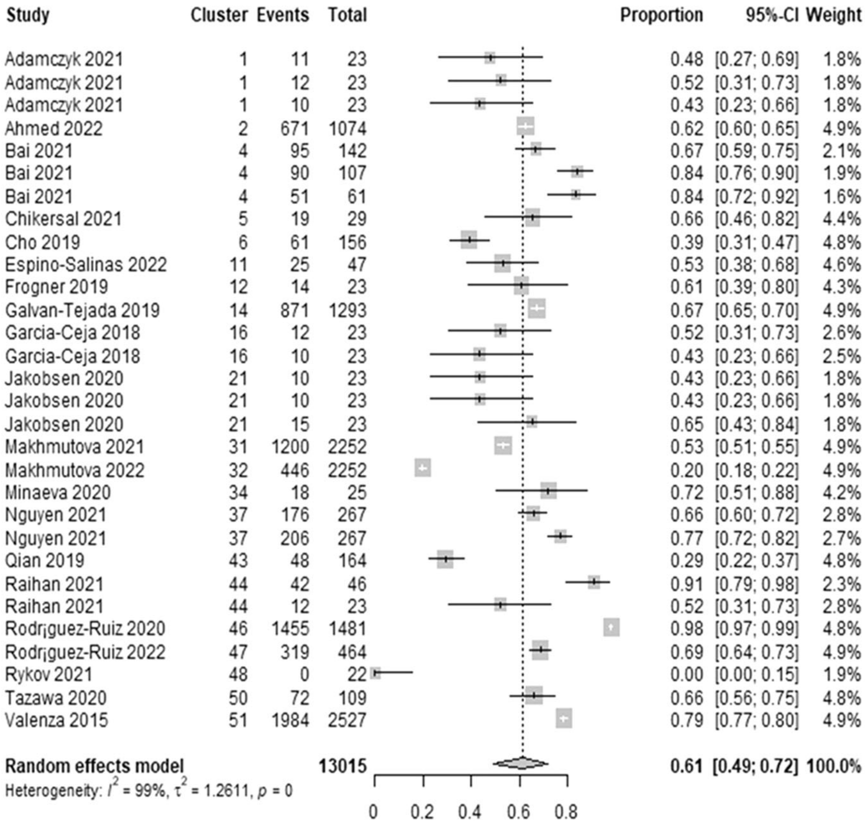


Fig. 7 Meta-analysis of the lowest sensitivity estimates. A total of 30 estimates of the lowest sensitivity from 21 studies were used in this meta-analysis. The square shape represents the lowest sensitivity in each study. The rhombus shape represents the pooled estimates of the lowest sensitivity in all studies. CI Confidence interval. p *p*-value.

those with other diseases that have similar signs and symptoms of depression.

As mentioned earlier, the sample size was small (≤55) in more than half of the studies (57.4%). For this reason, potential differences in the performance of wearable AI in subgroup analyses might not have manifested. This might also have prevented researchers to use some algorithms that need a very large sample size to be trained and tested. We urge researchers to conduct further studies with larger samples and over longer periods of time to ensure adequate statistical power as well as to enable the utilization of more complex and efficient algorithms requiring a larger amount of data.

About 61% of the included studies used Fitbit or Actiwatch AW4 to collect biomarkers although there are many other wearable devices in the market. For this reason, most subgroup analyses included only Fitbit or Actiwatch AW4, thereby, differences in the performance of different wearable devices in subgroup analyses might not have manifested. Further, none of the included studies compared the performance of different wearable devices. We recommend researchers use other wearable devices and compare the performance of different wearable devices.

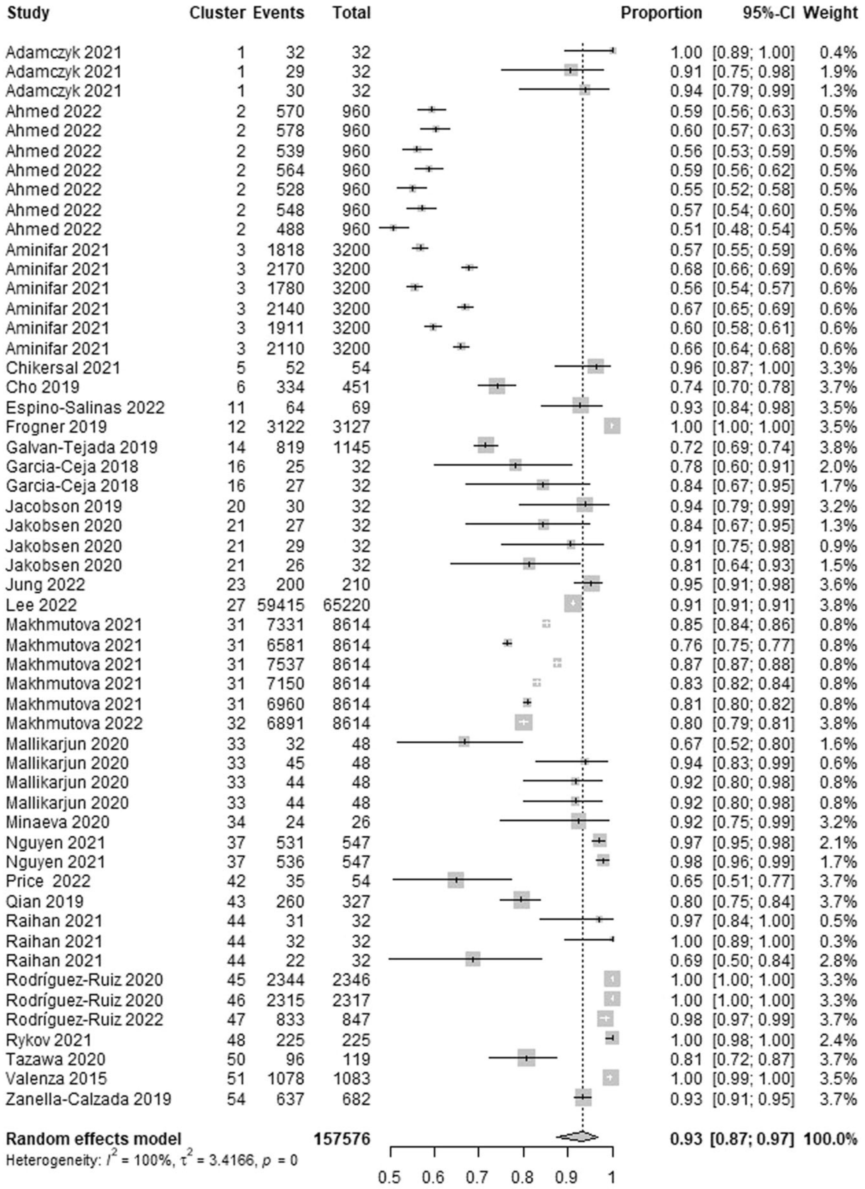
This review cannot comment on (1) the performance of wearable AI in detecting other mental disorders, (2) the performance of wearable AI in predicting outcomes of treatment for depression, and (3) the performance of non-wearable devices, hand-held devices, near-body wearable devices, in-body wearable devices, wearable devices connected with non-wearable devices using wires, and wearable devices that need an expert to apply on users. This is because such disorders, outcomes, and wearable devices were beyond the scope of this review, thereby, our findings may not be generalizable to such contexts. Further, we likely missed some studies given that we restricted our search, for practical constraints, to studies published in the English language from 2015 onwards. Results of our meta-analyses are likely to be overestimated or underestimated given that several included studies were included in the meta-analyses because they did not report results appropriate for the meta-analyses.

Wearable AI is a promising tool for detecting and predicting depression, but it is still in its infancy; meaning, it is not quite ready to be implemented in clinical practice. Until further research improve its performance, wearable AI should be used in conjunction with other methods for diagnosing and predicting depression (e.g., self-report questionnaires or interviews) to provide a more comprehensive understanding of a patient's condition. Tech companies should embrace the use of AI for the purpose of detecting and predicting depression using wearables. Researchers should examine the performance of wearable AI in the detection and prediction of depression based on a combination of wearable device data and neuroimaging data. Further studies should be conducted to distinguish patients with depression from those with other diseases that have similar signs and symptoms of depression. Wearables utilizing AI for detecting and predicting depression are getting better over time and we will likely see further development in this field via more accurate sensors and improved AI algorithms, we envisage this eventually leading to the possibility of use in clinical practice.

METHODS

Overview

We adhered to Preferred Reporting Items for Systematic Reviews and Meta-Analyses- Extension for Diagnostic Test Accuracy (PRISMA-DTA)[78] in reporting this review. Supplementary Table 10 outlines the PRISMA-DTA Checklist for this review. The protocol



Study	Cluster	Events	Total	Proportion	95%-CI	Weight
Adamczyk 2021	1	32	32	1.00	[0.89; 1.00]	0.4%
Adamczyk 2021	1	29	32	0.91	[0.75; 0.98]	1.9%
Adamczyk 2021	1	30	32	0.94	[0.79; 0.99]	1.3%
Ahmed 2022	2	570	960	0.59	[0.56; 0.63]	0.5%
Ahmed 2022	2	578	960	0.60	[0.57; 0.63]	0.5%
Ahmed 2022	2	539	960	0.56	[0.53; 0.59]	0.5%
Ahmed 2022	2	564	960	0.59	[0.56; 0.62]	0.5%
Ahmed 2022	2	528	960	0.55	[0.52; 0.58]	0.5%
Ahmed 2022	2	548	960	0.57	[0.54; 0.60]	0.5%
Ahmed 2022	2	488	960	0.51	[0.48; 0.54]	0.5%
Aminifar 2021	3	1818	3200	0.57	[0.55; 0.59]	0.6%
Aminifar 2021	3	2170	3200	0.68	[0.66; 0.69]	0.6%
Aminifar 2021	3	1780	3200	0.56	[0.54; 0.57]	0.6%
Aminifar 2021	3	2140	3200	0.67	[0.65; 0.69]	0.6%
Aminifar 2021	3	1911	3200	0.60	[0.58; 0.61]	0.6%
Aminifar 2021	3	2110	3200	0.66	[0.64; 0.68]	0.6%
Chikersal 2021	5	52	54	0.96	[0.87; 1.00]	3.3%
Cho 2019	6	334	451	0.74	[0.70; 0.78]	3.7%
Espino-Salinas 2022	11	64	69	0.93	[0.84; 0.98]	3.5%
Frogner 2019	12	3122	3127	1.00	[1.00; 1.00]	3.5%
Galvan-Tejada 2019	14	819	1145	0.72	[0.69; 0.74]	3.8%
Garcia-Ceja 2018	16	25	32	0.78	[0.60; 0.91]	2.0%
Garcia-Ceja 2018	16	27	32	0.84	[0.67; 0.95]	1.7%
Jacobson 2019	20	30	32	0.94	[0.79; 0.99]	3.2%
Jakobsen 2020	21	27	32	0.84	[0.67; 0.95]	1.3%
Jakobsen 2020	21	29	32	0.91	[0.75; 0.98]	0.9%
Jakobsen 2020	21	26	32	0.81	[0.64; 0.93]	1.5%
Jung 2022	23	200	210	0.95	[0.91; 0.98]	3.6%
Lee 2022	27	59415	65220	0.91	[0.91; 0.91]	3.8%
Makhmutova 2021	31	7331	8614	0.85	[0.84; 0.86]	0.8%
Makhmutova 2021	31	6581	8614	0.76	[0.75; 0.77]	0.8%
Makhmutova 2021	31	7537	8614	0.87	[0.87; 0.88]	0.8%
Makhmutova 2021	31	7150	8614	0.83	[0.82; 0.84]	0.8%
Makhmutova 2021	31	6960	8614	0.81	[0.80; 0.82]	0.8%
Makhmutova 2022	32	6891	8614	0.80	[0.79; 0.81]	3.8%
Mallikarjun 2020	33	32	48	0.67	[0.52; 0.80]	1.6%
Mallikarjun 2020	33	45	48	0.94	[0.83; 0.99]	0.6%
Mallikarjun 2020	33	44	48	0.92	[0.80; 0.98]	0.8%
Mallikarjun 2020	33	44	48	0.92	[0.80; 0.98]	0.8%
Minaeva 2020	34	24	26	0.92	[0.75; 0.99]	3.2%
Nguyen 2021	37	531	547	0.97	[0.95; 0.98]	2.1%
Nguyen 2021	37	536	547	0.98	[0.96; 0.99]	1.7%
Price 2022	42	35	54	0.65	[0.51; 0.77]	3.7%
Qian 2019	43	260	327	0.80	[0.75; 0.84]	3.7%
Raihan 2021	44	31	32	0.97	[0.84; 1.00]	0.5%
Raihan 2021	44	32	32	1.00	[0.89; 1.00]	0.3%
Raihan 2021	44	22	32	0.69	[0.50; 0.84]	2.8%
Rodríguez-Ruiz 2020	45	2344	2346	1.00	[1.00; 1.00]	3.3%
Rodríguez-Ruiz 2020	46	2315	2317	1.00	[1.00; 1.00]	3.3%
Rodríguez-Ruiz 2022	47	833	847	0.98	[0.97; 0.99]	3.7%
Rykov 2021	48	225	225	1.00	[0.98; 1.00]	2.4%
Tazawa 2020	50	96	119	0.81	[0.72; 0.87]	3.7%
Valenza 2015	51	1078	1083	1.00	[0.99; 1.00]	3.5%
Zanella-Calzada 2019	54	637	682	0.93	[0.91; 0.95]	3.7%
Random effects model			**157576**	**0.93**	**[0.87; 0.97]**	**100.0%**

Heterogeneity: $I^2 = 100\%$, $\tau^2 = 3.4166$, $p = 0$

Fig. 8 Meta-analysis of the highest specificity estimates. A total of 54 estimates of the highest specificity from 28 studies were used in this meta-analysis. The square shape represents the highest specificity in each study. The rhombus shape represents the pooled estimates of the highest specificity in all studies. CI Confidence interval. p *p*-value.

has been registered in with the International Prospective Register of Systematic Reviews (PROSPERO) (ID: CRD42022367856). The methods used in this review are detailed in the following subsections.

Search strategy

We identified the relevant studies having searched 8 electronic databases on October 3, 2022: MEDLINE (via Ovid), PsycInfo (via Ovid), EMBASE (via Ovid), CINAHL (via EBSCO), IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. An automatic search was set up with biweekly alerts for 3 months (ending on January 2, 2023). Only the first 100 hits (i.e.,10 pages) were checked for studies retrieved using Google Scholar in this review, due to the large number of results returned. Reference lists of included studies were checked (i.e., backward reference list checking), and studies that cited the included studies were screened (i.e., forward reference list checking) in order to identify additional studies.

Three experts in digital mental health were consulted whilst developing the search query, furthermore, previous reviews of relevance to the review were checked. Three groups of search terms were used: terms related to AI (e.g., artificial intelligence, machine learning, and deep learning), terms related to wearable devices (e.g., wearable, smartwatch, and smartwatch), and terms related to depression (e.g., depression and mood disorder). The search queries used in this review are highlighted in Supplementary Table 11.

Study eligibility criteria

This review examined papers that focused on building AI algorithms for depression utilizing wearable device data. We concentrated specifically on all AI algorithms utilized for detecting or predicting depression. We excluded studies that used AI for predicting the outcome of an intervention or treatment for depression. The data acquisition had to be non-invasive on-body wearables such as smartwatches, smart glasses, smart clothes,

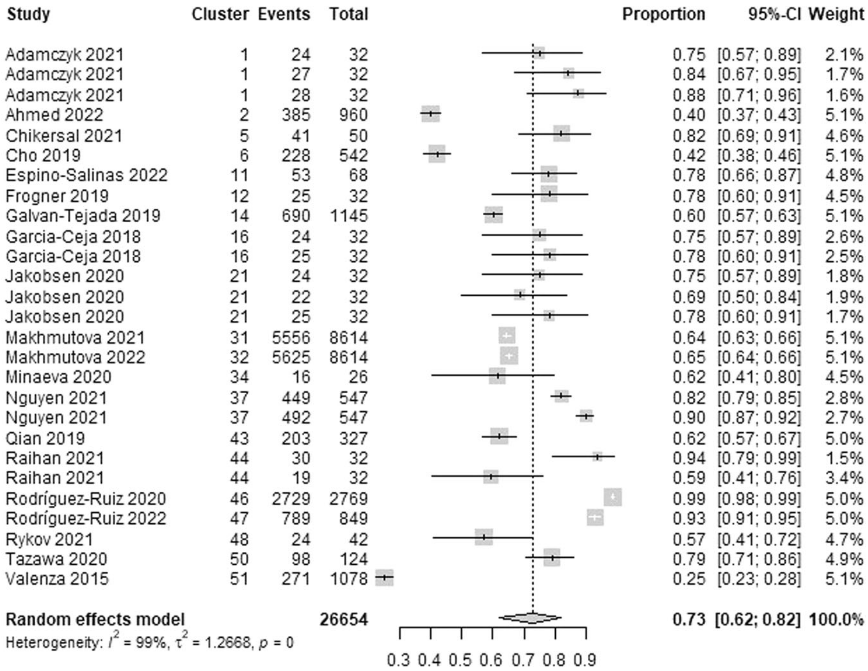


Fig. 9 Meta-analysis of the lowest specificity estimates. A total of 27 estimates of the lowest specificity from 20 studies were used in this meta-analysis. The square shape represents the lowest specificity in each study. The rhombus shape represents the pooled estimates of the lowest specificity in all studies. CI Confidence interval. p *p*-value.

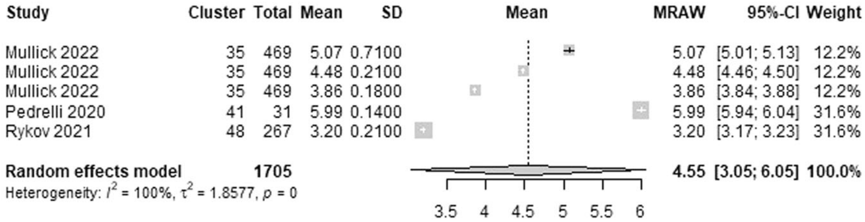


Fig. 10 Meta-analysis of the highest RMSE estimates. A total of 5 estimates of the highest RMSE from 3 studies were used in this meta-analysis. The square shape represents the highest RMSE in each study. The rhombus shape represents the pooled estimates of the highest RMSE in all studies. CI Confidence interval. p *p*-value.

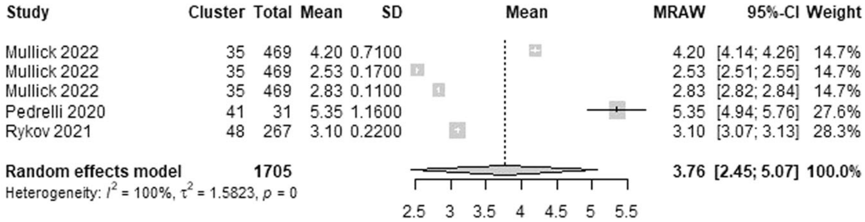


Fig. 11 Meta-analysis of the lowest RMSE estimates. A total of 5 estimates of the lowest RMSE from 3 studies were used in this meta-analysis. The square shape represents the lowest RMSE in each study. The rhombus shape represents the pooled estimates of the lowest RMSE in all studies. CI Confidence interval. p *p*-value.

smart wristbands, and smart tattoos. We excluded studies that used data collected by the following devices: non-wearable devices, hand-held devices (e.g., mobile phones), near-body wearable devices, in-body wearable devices (e.g., implants), wearable devices wired to non-wearable devices, and wearable devices that necessitate expert supervision (e.g., wearable devices composed of many electrodes that need to be placed in very specific points of the body). Studies that used data collected via other methods (e.g., non-wearable devices, questionnaires, and interviews) in addition to wearable devices were considered in this review. To be included in the current review, studies had to assess the performance of the AI algorithms in detecting or predicting depression and report the confusion matrix and/or performance measures (e.g., accuracy, sensitivity, specificity, etc.). We disregarded articles that typically demonstrated a theoretical foundation of AI-powered wearable devices for depression. We accepted journal articles, conference papers, and dissertations published in English since 2015. Reviews, preprints, conference abstracts, posters, protocols, editorials, and comments were not included. There were no constraints on the setting, reference standard, or country of publication.

Study selection

In the study selection process, we followed three procedures. EndNote X9 was used in the first stage to eliminate duplicates from all retrieved studies. The titles and abstracts of the remaining articles were examined in the second stage. Finally, we read over the whole texts of the studies that were included in the previous stage. The research selection procedure was carried out separately by two reviewers. Disagreements in the second and third phases were settled through dialogue. Cohen's kappa was used to calculate inter-rater agreement, which was 0.85 for "title and abstract" screening and 0.92 for full-text reading.

Data extraction

Two reviewers independently extracted data on study meta-data, wearable devices, AI algorithms, and results of studies using Microsoft Excel. Disagreements among the reviewers were overcome through discussion. When the raw data or confusion matrix is reported in the included studies, we calculated all possible performance measures such as accuracy, sensitivity, specificity, and precision. We did not extract results related to the performance of AI algorithms that are based on only non-wearable-device data (e.g., data collected by smartphones or questionnaires). Given that many studies conducted several experiments to test, for example, different numbers of features, data types, validation approaches, and AI techniques, they reported several results for the same performance measure. Therefore, we extracted the lowest and highest results for each performance measure for each algorithm. The data extraction form utilized in this review was trialled with five studies (Supplementary Table 12).

Risk of bias and applicability appraisal

We modified a well-known risk of bias assessment tool (Quality Assessment of Studies of Diagnostic Accuracy-Revised (QUADAS-2))[79] by removing some irrelevant criteria and adding other criteria from another relevant tool (the Prediction model Risk Of Bias ASsessment Tool (PROBAST))[80]. Similar to the original QUADAS-2, the modified version evaluates the risk of bias of the included studies in terms of four domains (participants, index test (AI algorithms), reference standard (ground truth), and analysis) whereas it evaluates their applicability to the review question in terms of three domains (participants, index test (AI algorithms), reference standard (ground truth)). Each domain consists of four signalling questions that were tailored to the goal of this review. Based on the answers to these questions, the risk of bias and applicability in the corresponding domain was assessed. Supplementary Table 13 shows the modified version of QUADAS-2, which was pilot tested using four included studies. Two reviewers independently used the modified version to assess the risk of bias and the applicability of the included studies. Disagreements between the two reviewers were settled by discussion.

Data synthesis

The extracted data were synthesized using narrative and statistical approaches. Knowing that studies or groups of studies reported multiple effect sizes will have a larger influence on the results of the meta-analysis than studies reporting only one effect size. Hence, the risk of biased estimates is high, meaning that the potential dependency between effect sizes, for studies that reported more than one effect size, is needed to be considered in our meta-analysis. Multi-level meta-analysis is a statistical technique used to combine the results of multiple studies while taking into account that data is nested (i.e., the observations are not independent) thereby reducing the likelihood of Type I errors. We, therefore, used a three-level model to analyze the data, where we anticipated a set of studies (level 3), repeated analysis nested within studies (level 2), and a sample of subjects for each analysis (level 1). Using three-level meta-analysis uses three sources of variance: population differences between study population effects, population differences between effects of experiments from the same study, and, finally, sampling variance. We used a random-effects model, assuming a priori significant heterogeneity resulting from diverse study populations and different models. The extracted data was used to compute pooled mean accuracy, sensitivity, specificity, and root mean square error (RMSE). Stratification (subgroup) analysis was conducted based on AI algorithms, aims of AI, wearable devices, data sources, types of data, and reference standards. To assess the degree of heterogeneity and the statistical significance of heterogeneity in the meta-analyzed studies, we computed I^2 and Cochran's *Q*-test. The presence of heterogeneity in the meta-analyzed studies is indicated by a Cochran's *p*-value ≤ 0.05[81]. The degree of heterogeneity was considered insignificant when I^2 ranged from 0% to 40%, moderate when it ranged from 30% to 60%, substantial when it ranged from 50% to 90%, or considerable when it ranged from 75% to 100%[81]. The R version 4.2.2 was used to perform meta-analyses.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 23 January 2023; Accepted: 19 April 2023;
Published online: 05 May 2023

REFERENCES

1. Institute of Health Metrics and Evaluation. *Global Health Data Exchange (GHDx)*. http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b (2019).
2. American Psychiatric Association. *What Is Depression*? https://www.psychiatry.org/patients-families/depression/what-is-depression#section_2 (2022).
3. Jia, H., Zack, M. M., Thompson, W. W., Crosby, A. E. & Gottesman, I. I. Impact of depression on quality-adjusted life expectancy (QALE) directly as well as indirectly through suicide. *Soc. Psychiatry Psychiatr. Epidemiol.* **50**, 939–949 (2015).
4. Wainberg, M. L. et al. Challenges and opportunities in global mental health: a research-to-practice perspective. *Curr. Psychiatry Rep.* **19**, 28 (2017).
5. Murray, C. J. 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* **380**, 2197–2223 (2012).
6. Oladeji, B. D. & Gureje, O. Brain drain: a challenge to global mental health. *BJPsych Int* **13**, 61–63 (2016).
7. Abd-alrazaq, A. et al. Wearable artificial intelligence for anxiety and depression: Scoping review. *J. Med. Internet Res.* **25**, e42672 (2023).
8. Statista. *Wearables—Statistics and Facts.* https://www.statista.com/topics/1556/wearable-technology/#topicHeader__wrapper (2022).
9. Statista. *Does You Household Own Wearables (e.g. Smart Watch, Health/Fitness Tracker)?* https://www.statista.com/forecasts/1101101/wearable-devices-ownership-in-selected-countries (2022).
10. Lee, S., Kim, H., Park, M. J. & Jeon, H. J. Current advances in wearable devices and their sensors in patients with depression. *Front. Psychiatry* **12**, 672347 (2021).
11. Sequeira, L. et al. Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: a scoping review. *J. Affect Disord.* **265**, 314–324 (2020).

12. Welch, V. et al. Use of mobile and wearable artificial intelligence in child and adolescent psychiatry: scoping review. *J. Med Internet Res.* **24**, e33560 (2022).
13. Yasin, S. et al. EEG based major depressive disorder and bipolar disorder detection using neural networks: a review. *Comput. Methods Programs Biomed.* **202**, 106007 (2021).
14. Kim, S. & Lee, K. Screening for depression in mobile devices using Patient Health Questionnaire-9 (PHQ-9) Data: a diagnostic meta-analysis via machine learning methods. *Neuropsychiatr. Dis. Treat.* **17**, 3415–3430 (2021).
15. Kambeitz, J. et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol. Psychiatry* **82**, 330–338 (2017).
16. Adamczyk, J. & Malawski, F. Comparison of manual and automated feature engineering for daily activity classification in mental disorder diagnosis. *Comput. Inform.* **40**, 850–879 (2021).
17. Ahmed, A. et al. Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning. *Information* **13**, 406 (2022).
18. Aminifar, A., Rabbi, F., Pun, V. K. I. & Lamo, Y. Monitoring motor activity data for detecting patients' depression using data augmentation and privacy-preserving distributed learning. *Annu Int Conf. IEEE Eng. Med. Biol. Soc.* **2021**, 2163–2169 (2021).
19. Bai, R. et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. *JMIR Mhealth Uhealth* **9**, e24365 (2021).
20. Chikersal, P. et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Trans. Comput-Hum. Interact.* **28**, Article 3 (2021).
21. Cho, C. H. et al. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian Rhythm: prospective observational cohort study. *J. Med. Internet Res.* **21**, e11029 (2019).
22. Choi, J., Lee, S., Kim, S., Kim, D. & Kim, H. Depressed mood prediction of elderly people with a wearable band. *Sensors (Basel)*. https://doi.org/10.3390/s22114174 (2022).
23. Choi, J. G., Ko, I. & Han, S. Depression level classification using machine learning classifiers based on actigraphy data. *IEEE Access* **9**, 116622–116646 (2021).
24. Coutts, L. V., Plans, D., Brown, A. W. & Collomosse, J. Deep learning with wearable based heart rate variability for prediction of mental and general health. *J. Biomed. Inf.* **112**, 103610 (2020).
25. Dai, R. et al. Multi-task learning for randomized controlled trials: a case study on predicting depression with wearable data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **6**, Article 50 (2022).
26. Espino-Salinas, C. H. et al. Two-dimensional convolutional neural network for depression episodes detection in real time using motor activity time series of depresjon dataset. *Bioeng. (Basel)* **9**, 458 (2022).
27. Frogner, J. I. et al. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In *Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care* 9–15 (Association for Computing Machinery, 2019).
28. Fukuda, S. et al. editors. *Predicting Depression and Anxiety Mood by Wrist-Worn Sleep Sensor. 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* 23–27 (IEEE, 2020).
29. Galván-Tejada, C. E. et al. Depression episodes detection in unipolar and bipolar patients: a methodology with feature extraction and feature selection with genetic algorithms using activity motion signal as information source. *Mob. Inf. Syst.* **2019**, 8269695 (2019).
30. Garcia-Ceja E., et al. Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM Multimedia Systems Conference* 472–477 (Association for Computing Machinery, 2018).
31. Garcia-Ceja E., et al. Motor activity based classification of depression in unipolar and bipolar patients. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)* 18–21 (IEEE, 2018).
32. Ghandeharioun A., et al., editors. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* 23-26 (IEEE, 2017).
33. Griffiths, C. et al. Investigation of physical activity, sleep, and mental health recovery in treatment resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rTMS) treatment. *J. Affect Disord. Rep.* **8**, 100337 (2022).
34. Horwitz A. G., et al. Using machine learning with intensive longitudinal data to predict depression and suicidal ideation among medical interns over time. *Psychol. Med.* https://doi.org/10.1017/s0033291722003014 (2022).
35. Jacobson, N. C., Weingarden, H. & Wilhelm, S. Digital biomarkers of mood disorders and symptom change. *NPJ Digital Med.* **2**, 3 (2019).
36. Jakobsen, P. et al. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. *PLoS One* **15**, e0231995 (2020).
37. Jin, J. et al. Attention-block deep learning based features fusion in wearable social sensor for mental wellbeing evaluations. *IEEE Access* **8**, 89258–89268 (2020).
38. Jung, D., Kim, J. & Mun, K. R. Identifying depression in the elderly using gait accelerometry. *Annu Int Conf. IEEE Eng. Med Biol. Soc.* **2022**, 4946–4949 (2022).
39. Kim, H. et al. Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: observational study on older adults living alone. *JMIR Mhealth Uhealth* **7**, e14149 (2019).
40. Kulam, S. *Time-Series Classification With Uni-Dimensional Convolutional Neural Networks: An Experimental Comparison With Long Short-Term Memory Networks*. https://www.duo.uio.no/handle/10852/73250?locale-attribute=en (2019).
41. Kumar, A., Sangwan, S. R., Arora, A. & Menon, V. G. Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT. *Appl. Soft Comput.* **122**, 108863 (2022).
42. Lee H. J., et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. *Psychol. Med.* https://doi.org/10.1017/s0033291722002847 (2022).
43. Llamocca P., López V., Santos M., Čukić M. Personalized characterization of emotional states in patients with bipolar disorder. *Mathematics* https://doi.org/10.3390/math9111174 (2021).
44. Lu, J. et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, Article 21 (2018).
45. Mahendran N., et al. Sensor-assisted weighted average ensemble model for detecting major depressive disorder. *Sensors (Basel)* https://doi.org/10.3390/s19224822 (2019).
46. Makhmutova, M. *Predicting Changes in Depression Using Person-Generated Health Data* (Ecole Polytechnique Federale de Lausanne, 2021).
47. Makhmutova, M. et al. Predicting changes in depression severity using the PSYCHE-D (prediction of severity change-depression) model involving person-generated health data: longitudinal case-control observational study. *JMIR Mhealth Uhealth* **10**, e34148 (2022).
48. Mallikarjun, H. M. & Manimegalai, P. Manoglanistara—emotional wellness phases prediction of adolescent female students by using brain waves. *Curr. Signal Transduct. Ther.* **15**, 315–323 (2020).
49. Minaeva, O. et al. Screening for depression in daily life: development and external validation of a prediction model based on actigraphy and experience sampling method. *J. Med Internet Res.* **22**, e22634 (2020).
50. Mullick, T., Radovic, A., Shaaban, S. & Doryab, A. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. *JMIR Form. Res.* **6**, e35807 (2022).
51. Narziev, N. et al. STDD: Short-term depression detection with passive sensing. *Sensors (Basel)*. https://doi.org/10.3390/s20051396 (2020).
52. Nguyen D-K, Chan C-L, Li A-HA, Phan D-V. Deep stacked generalization ensemble learning models in early diagnosis of Depression illness from wearable devices data. In *2021 5th International Conference on Medical and Health Informatics* 7–12 (Association for Computing Machinery, Japan, 2021).
53. Nishimura, Y. et al. *Toward the Analysis of Office Workers' Mental Indicators Based on Wearable, Work Activity, and Weather Data. Sensor-and Video-Based Activity and Behavior Computing* 1–26 (Springer; 2022).
54. Opoku Asare, K. et al. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. *Pervasive Mob. Comput.* **83**, 101621 (2022).
55. Pacheco-González, S. L. et al. Evaluation of five classifiers for depression episodes detection. *Res. Comput Sci.* **148**, 129–138 (2019).
56. Pedrelli, P. et al. Monitoring changes in depression severity using wearable and mobile sensors. *Front Psychiatry* **11**, 584711 (2020).
57. Price, G. D. et al. An unsupervised machine learning approach using passive movement data to understand depression and schizophrenia. *J. Affect Disord.* **316**, 132–139 (2022).
58. Qian K., et al. Automatic detection of major depressive disorder via a bag-of-behaviour-words approach. In *Proceedings of the Third International Symposium on Image Computing and Digital Medicine* 71–75 (ACM, 2019).
59. Raihan M., Bairagi A. K., and Rahman S. A machine learning based study to predict depression with monitoring actigraph watch data. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* 6–8 (IEEE, 2021).
60. Rodríguez-Ruiz, J. G., Galván-Tejada, C. E., Vázquez-Reyes, S., Galván-Tejada, J. I. & Gamboa-Rosales, H. Classification of depressive episodes using nighttime data; a multivariate and univariate analysis. *Program Comput. Softw.* **46**, 689–698 (2020).

61. Rodríguez-Ruiz, J. G. et al. Comparison of night, day and 24 h motor activity data for the classification of depressive episodes. *Diagnostics (Basel)*. https://doi.org/10.3390/diagnostics10030162 (2020).

62. Rodríguez-Ruiz, J. G. et al. Classification of depressive and Schizophrenic episodes using night-time motor activity signal. *Healthcare (Basel)*. https://doi.org/10.3390/healthcare10071256 (2022).

63. Rykov, Y., Thach, T. Q., Bojic, I., Christopoulos, G. & Car, J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR Mhealth Uhealth* **9**, e24872 (2021).

64. Shah, R. V. et al. Personalized machine learning of depressed mood using wearables. *Transl. Psychiatry* **11**, 338 (2021).

65. Tazawa, Y. et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon* **6**, e03274 (2020).

66. Valenza, G. et al. Characterization of depressive states in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment. *IEEE J. Biomed. Health Inf.* **19**, 263–274 (2015).

67. Wang, R. et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, Article 43 (2018).

68. Xu, X. et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **3**, 116 (2019).

69. Zanella-Calzada L. A., et al. Feature extraction in motor activity signal: towards a depression episodes detection in unipolar and bipolar patients. *Diagnostics (Basel)*. https://doi.org/10.3390/diagnostics9010008 (2019).

70. Cohen, S. E., Zantvoord, J. B., Wezenberg, B. N., Bockting, C. L. H. & van Wingen, G. A. Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. *Transl. Psychiatry* **11**, 168 (2021).

71. Lee, Y. et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J. Affect. Disord.* **241**, 519–532 (2018).

72. Watts, D. et al. Predicting treatment response using EEG in major depressive disorder: a machine-learning meta-analysis. *Transl. Psychiatry* **12**, 332 (2022).

73. Korgaonkar, M. S., Williams, L. M., Song, Y. J., Usherwood, T. & Grieve, S. M. Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder. *Br. J. Psychiatry* **205**, 321–328 (2014).

74. Lord, A., Horn, D., Breakspear, M. & Walter, M. Changes in community structure of resting state functional connectivity in unipolar depression. *PLoS One* **7**, e41282 (2012).

75. Qin, J. et al. Predicting clinical responses in major depression using intrinsic functional connectivity. *NeuroReport* **26**, 675–680 (2015).

76. Qin, J. et al. Altered anatomical patterns of depression in relation to antidepressant treatment: evidence from a pattern recognition analysis on the topological organization of brain networks. *J. Affect. Disord.* **180**, 129–137 (2015).

77. Wei, M. et al. Identifying major depressive disorder using Hurst exponent of resting-state brain networks. *Psychiatry Res. Neuroimaging* **214**, 306–312 (2013).

78. McInnes, M. D. F. et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* **319**, 388–396 (2018).

79. Whiting, P. F. et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).

80. Wolff, R. F. et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern Med.* **170**, 51–58 (2019). PMID: 30596875.

81. Deeks J. J., Higgins J. P., Altman D. G. and Cochrane Statistical Methods Group. *Analysing Data and Undertaking Meta-Analyses*. https://training.cochrane.org/handbook/current/chapter-10 (2019).

AUTHOR CONTRIBUTIONS

A.Ab., A.Ah., and S.A.developed the protocol with guidance from and under the supervision of Javaid Sheikh. A.Ab. searched the electronic databases and conducted backward and forward reference list checking. The study selection process, data extraction, and risk of bias assessment were carried out by R.A. and A.Ab. Data synthesis was conducted by F.S. and A.Ab. The introduction section was written by R.A. A.Ab. wrote the methods, results, and discussion sections. The article was revised critically for important intellectual content by all authors. All authors approved the manuscript for publication and agree to be accountable for all aspects of the work.

FUNDING

Open Access funding provided by the Qatar National Library.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00828-5.

Correspondence and requests for materials should be addressed to Alaa Abd-Alrazaq.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.