Utilizing Digital Phenotyping to Discriminate Unipolar Depression and Bipolar Disorder: A Systematic Review

Rongrong Zhong, XiaoHui Wu, Jun Chen, Yiru Fang

Submitted to: Journal of Medical Internet Research
on: February 18, 2025

Table of Contents

Utilizing Digital Phenotyping to Discriminate Unipolar Depression and Bipolar Disorder: A Systematic Review

Rongrong Zhong[1]; XiaoHui Wu[1]; Jun Chen[1]; Yiru Fang[2]

[1] Shanghai Jiao Tong University School of Medicine, Shanghai Mental Health Center, Clinical Research Center & Division of Mood Disorders Shanghai CN
[2] Shanghai Jiao Tong University School of Medicine, Ruijin Hospital, Department of Psychiatry & Affective Disorders Center Shanghai CN

Corresponding Author:
Yiru Fang

Shanghai Jiao Tong University School of Medicine, Ruijin Hospital, Department of Psychiatry & Affective Disorders Center
197 Rui Jin Er Road, Huangpu District
Shanghai
CN

Abstract

Background: Differentiating bipolar disorder (BD) from unipolar depression (UD) is essential, as these conditions differ greatly in their progression and treatment approaches. Digital phenotyping, which involves using data from smartphones or other digital devices to assess mental health, has emerged as a promising tool for distinguishing between these two disorders.

Objective: This systematic review aims to summarize the findings of existing studies that assess the use of digital devices in distinguishing between UD and BD. Additionally, it seeks to identify gaps in the current research and suggest directions for future studies.

Methods: A comprehensive search was conducted in Scopus, IEEE Xplore, PubMed, and EMBASE databases until 20/9/2024. The search was focused on studies evaluating the use of digital tools, such as smartphone apps, wearable devices, audio/video recordings, and multimodal technologies, for patients with UD and BD. The review protocol was registered in PROSPERO (CRD42024624202).

Results: We identified 18 studies involving 732 participants with UD and 590 participants with BD. Among the included studies, 4 utilized smartphone apps, 3 employed wearable devices, 10 analyzed audio/video recordings, and 1 used multimodal technologies. Importantly, 9 studies directly focused on differentiating between UD and BD, excluding healthy controls or other mental health conditions from their classifications.

Conclusions: Digital phenotyping offers significant promise in distinguishing between UD and BD, with ongoing advances in technology. However, challenges such as data privacy, security concerns, and equitable access must be addressed to fully harness the potential of digital tools for inclusive mental health care. Further research should focus on overcoming these challenges and refining digital phenotyping methodologies to ensure broader applicability in clinical settings.

(JMIR Preprints 18/02/2025:72229)
DOI: https://doi.org/10.2196/preprints.72229

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Utilizing Digital Phenotyping to Discriminate Unipolar Depression and Bipolar Disorder: A Systematic Review

Rongrong Zhong[1], XiaoHui Wu[1], Jun Chen[1], Yiru Fang[1,2,*]

1 Clinical Research Center & Division of Mood Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, 200030, China
2 Department of Psychiatry & Affective Disorders Center, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 200025, China

*Corresponding author at: Yiru Fang, 197 Rui Jin Er Road, Huangpu District, Shanghai, China.
E-mail addresses: yirufang@aliyun.com (Y. Fang).

Abstract

Background: Differentiating bipolar disorder (BD) from unipolar depression (UD) is essential, as these conditions differ greatly in their progression and treatment approaches. Digital phenotyping, which involves using data from smartphones or other digital devices to assess mental health, has emerged as a promising tool for distinguishing between these two disorders.

Objective: This systematic review aims to achieve two goals: (1) to summarize the existing literature on the use of digital phenotyping to directly distinguish between UD and BD; (2) to review studies that utilize digital phenotyping to classify UD, BD, and healthy controls (HC). Furthermore, the review seeks to identify gaps in the current research and propose directions for future studies.

Methods: We systematically searched Scopus, IEEE Xplore, PubMed, EMBASE, Web of Science, and PsycINFO databases up to March 20, 2025. Studies were included if they used portable or wearable digital tools to directly distinguish between UD and BD, or to classify UD, BD and HC. Original studies published in English, including both journal and conference papers, were included, while reviews, narrative reviews, systematic reviews, and meta-analyses were excluded. Articles were excluded if the diagnosis was not made through a professional medical evaluation, or if they relied on electronic health records or clinical data. For each included study, the following information was extracted: geographic region, sample size, epidemiological data, diagnostic criteria or psychiatric assessments, details of the technological tools and data types, duration of data collection, data preprocessing methods, selected variables/features, machine learning algorithms and/or statistical tests, validation, and main findings. The review protocol was registered in PROSPERO (CRD42024624202).

Results: We included 21 studies: 11 focused on directly distinguishing between UD and BD, while 10 classified UD, BD, and HC. The studies were categorized into four groups based on the type of digital tool used: 6 utilized smartphone apps, 3 employed wearable devices, 11 analyzed audiovisual recordings, and 1 used multimodal technologies. Features such as activity levels from smartphone apps or wearable devices, emerged as potential markers for directly distinguishing UD and BD. BD patients generally exhibited lower activity levels than UD patients. They also tended to show higher activity in the morning and lower in the evening, while UD patients showed the opposite pattern. Moreover, speech modalities or the integration of multiple modalities achieved better classification performance across UD, BD, and HC groups, although the specific contributing features remained unclear.

Conclusions: Digital phenotyping shows promising potential in distinguishing between UD and BD, with ongoing advances in technology. However, challenges such as data privacy, security concerns, and equitable access must be addressed to fully harness the potential of digital tools for inclusive mental health care. Further research should focus on overcoming these challenges and refining

digital phenotyping methodologies to ensure broader applicability in clinical settings.

Keywords: Digital phenotyping; Depression; Bipolar disorder; Smartphone; Wearable; Audiovisual; Multimodal

Introduction

Background

Mood disorders are a highly prevalent and recurrent group of mental disorders associated with a significant risk of suicide [1], primarily encompassing unipolar depression (UD) and bipolar disorder (BD) [2]. Approximately one in four individuals is estimated to experience an affective disorder at least once in their lifetime, often leading to significant and lasting disability for those affected [3]. UD is primarily characterized by significant and persistent low mood. In contrast, BD involves (hypo-)manic episodes (elevated mood, racing thoughts, increased activity) and depressive episodes (low mood, slowed thinking, reduced activity). Both disorders are marked by significant and lasting mood changes impacting emotional, cognitive, and behavioral domains [4]. However, the course of BD often begins with depressive episodes, leading to a substantial risk of misdiagnosis, as approximately 40% to 69% of patients with BD are initially diagnosed with UD [5,6]. This misdiagnosis can have serious consequences, including inappropriate medication prescriptions, triggering manic episodes, prolonged illness duration, increased risk of recurrence, heightened suicide risk, and an overall poorer response to treatment [5,7–10].

Currently, psychiatrists typically diagnose based on established criteria (such as DSM-5 and ICD-11), relying on one-time self-reports from patients and their families. This approach heavily depends on the clinician’s experience and often lacks continuity and objectivity. In contrast, digital phenotyping offers a promising solution, with this concept being a groundbreaking advancement in the field, first introduced in 2015 [11]. Digital phenotyping includes long-term active data (e.g., participants completing daily self-assessment questionnaires via smartphone apps) [12], providing clinicians with a more comprehensive and continuous flow of information. It also offers objective data based on digital devices, such as physiological measurements (e.g., skin temperature, heart rate, blood volume pulse, electrodermal activity) and behavioral indicators (e.g., acoustic features, gestures, facial expressions) [13].

Compared to traditional diagnostic methods, digital phenotyping has great potential to improve diagnostic accuracy and timeliness. However, it generates vast amounts of data that require more robust processing and analysis. To address this, leveraging artificial intelligence (AI), including machine learning, in mental health is essential. Currently, machine learning has gradually emerged as a powerful tool for exploring high-dimensional and real-time data associated with digital phenotyping. It provides an opportunity to "make sense" of these digital phenotypes and the realities they attempt to represent in the context of mental health [14–16]. Some believe this technology has the potential to offer deeper insights into the neurobiological mechanisms underlying psychiatric disorders [17]. It may also facilitate the development of new transdiagnostic models for understanding symptoms [12]. This aligns with the “Research Domain Criteria” (RDoC) perspective proposed in recent years [18]. In addition to classification tasks, machine learning algorithms may also have the capability to predict episodes or even suicidal risk, enabling clinicians to make more accurate and timely clinical decisions. Consequently, digital phenotyping, supported by machine learning, has carved out an important role in psychiatry [19], helping clinicians access individualized behavioral, emotional, and other data from patients with mental disorders. This approach not only enhances psychiatrists’ understanding of symptoms and the disorders themselves [20,21], but may also compensate for the current lack of reliable biomarkers.

Objectives

In this study, we conducted a systematic review of original articles from both journals and conference proceedings, exploring the use of portable or wearable digital tools for distinguishing UD

and BD, as well as classifying UD, BD, and healthy controls (HC). We examined the study populations by considering factors such as geographic regions, sample sizes, epidemiological data, and the diagnostic criteria or psychiatric assessments used. Additionally, we analyzed the technological tools and data types employed, the duration of data collection, data preprocessing methods, selected variables or features, computational techniques applied, validation approaches, and the results achieved.

This review is essential because many previous studies in this field have either focused on specific tools or were limited by small sample sizes, which restricts their applicability to broader populations. Furthermore, the lack of consistency in methodologies and the absence of comprehensive validation across different groups have hindered the generalizability of the findings. This review aims to address these limitations and provide valuable insights into the potential of digital phenotyping for more accurate and reliable differentiation between these groups.

Methods

Information sources and search strategy

This Systematic Review was conducted in agreement with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Multimedia Appendix 1) [22], and is listed in the PROSPERO register (registration number: CRD42024624202). While the review was generally conducted according to the registered protocol, some minor deviations occurred during the study. These adjustments primarily involved refining and clarifying the study objectives and title to better reflect the focus of the review. Additionally, we expanded the scope to include studies that classified UD, BD, and HC, which was not initially specified in the protocol. Finally, we excluded studies based on non-clinically diagnosed data, such as those relying on social media samples, to ensure that only studies adhering to professional diagnostic standards were included.

We conducted a comprehensive search across six major databases—Scopus, IEEE Xplore, PubMed, EMBASE, Web of Science, and PsycINFO—for articles published up to March 20, 2025. The search terms used in each database are provided in the supplementary material (Multimedia Appendix 2). We included original studies, both journal articles and conference papers, published in English, with no restrictions on publication date.

Eligibility criteria

Eligible studies involved participants diagnosed with UD, BD, or HC, and used portable or wearable digital devices such as smartphone apps, wearable sensors, or audio and/or visual recordings. The studies were required to either compare digital phenotyping results with diagnostic outcomes from professional medical evaluations, compare UD with BD, or perform a classification task involving UD, BD, and HC. The primary outcome of interest was classification performance metrics such as sensitivity, specificity, area under the curve (AUC), accuracy, recall, and precision, but studies reporting t-tests, ANOVA, non-parametric tests, or correlation analyses were also considered.

Following the PICOS criteria, we excluded studies that were reviews, meta-analyses, or written in languages other than English. Additionally, studies were excluded if they used technologies unsuitable for daily monitoring, were based on electronic health records or clinical data, or had diagnoses not made through professional medical evaluations (e.g., studies based on social media data).

Selection process

Title and abstract selection

The titles and abstracts of all articles that matched the search criteria were double-screened. After reviewing them, we excluded studies that did not address the research question.

Full-text selection

We included in the systematic review only papers that aimed to differentiate between UD and BD, either directly or by classifying UD, BD, and HC, using digital phenotyping, according to the previously defined PICOS criteria. The included papers were read thoroughly to extract the relevant data.

All articles that met the search criteria were independently screened by two reviewers (ZRR and WXH) during both the title and abstract selection and the full-text selection stages. In cases of disagreement, a third reviewer (FYR) was consulted to achieve consensus.

Data extraction

For each study, the following information was extracted: geographic region, sample size, epidemiological data of the sample (number and percentage of females, mean age), diagnostic criteria or psychiatric assessments, type of technology and data collected, duration of data collection, data preprocessing methods, specific variables/features selected, machine learning algorithms and/or statistical tests used, validation, and main findings. Data extraction was independently conducted by two authors (ZRR and WXH). Discrepancies were resolved through discussion or with the involvement of a third author (FYR).

Synthesis method

The included papers were grouped based on the type of digital tool used (smartphone apps, wearable devices, audiovisual recordings, or multimodal technologies) and the comparison type (UD vs. BD or UD vs. BD vs. HC). A narrative synthesis approach was employed to summarize and compare study characteristics, data modalities, analytical methods, and classification performance. Quantitative pooling was not performed due to the heterogeneity in study designs and feature selection.

Study risk of bias assessment

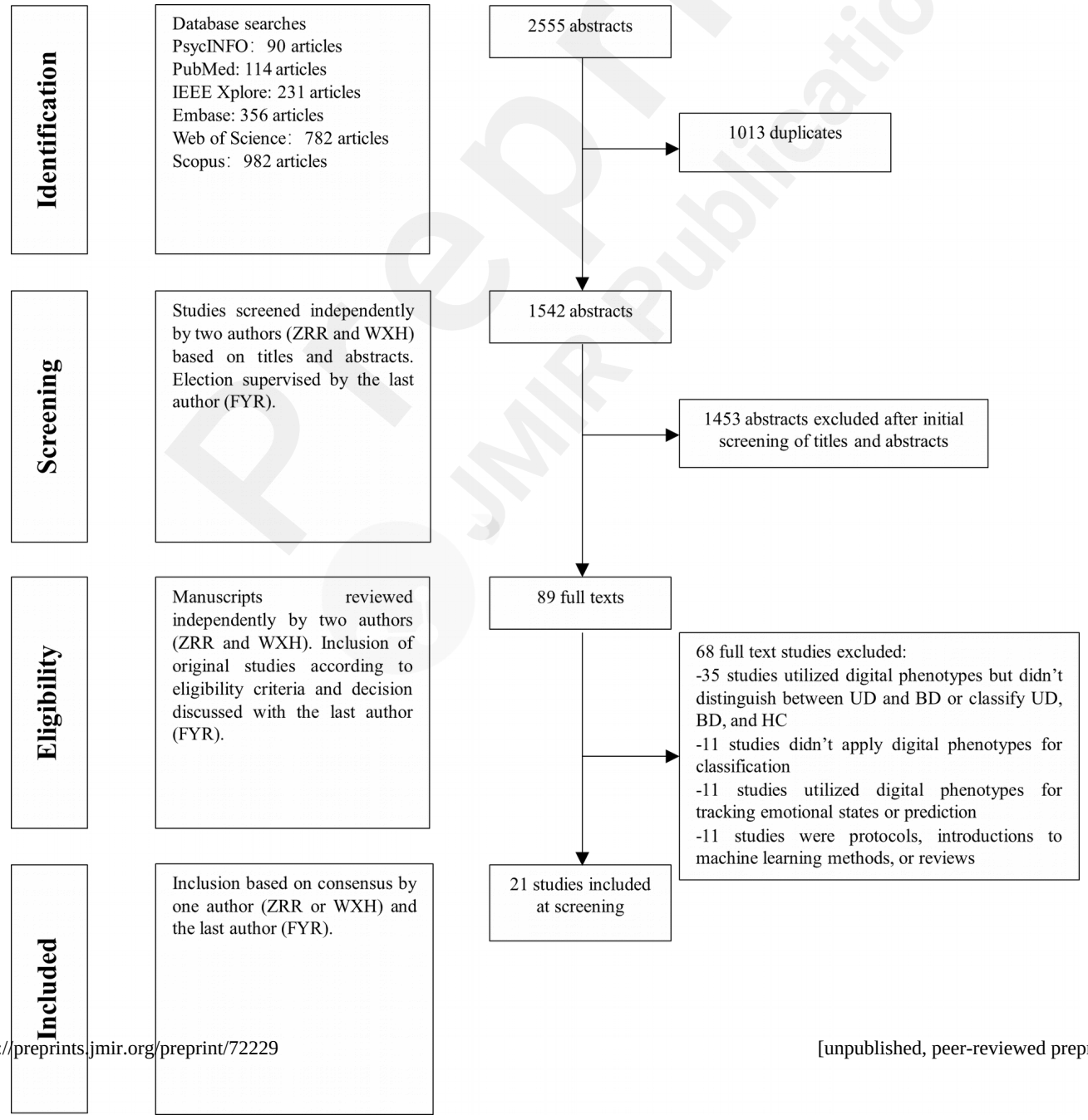
The quality of the selected studies was independently evaluated by two reviewers (ZRR and WXH) using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [23] tool. QUADAS-2 comprises four key domains: patient selection, index test, reference standard, and flow and timing. For each domain, the risk of bias and concerns regarding applicability were assessed and categorized as low, high, or unclear risk.

Results

As shown in Figure 1, a thorough search was conducted across six major databases, yielding a total of 2,555 articles. Specifically, PsycINFO identified 90 articles, PubMed provided 114 articles, IEEE returned 231 articles, Embase found 356 articles, Web of Science collected 782 articles, and Scopus gathered 982 articles. After removing 1,013 duplicate records, the titles and abstracts of the remaining 1,542 articles were screened for relevance to the topic, narrowing the selection to 89

articles for full-text eligibility assessment. During the eligibility screening, 68 full-text studies were excluded after a detailed evaluation. Among these, 35 studies applied digital phenotyping but did not distinguish between UD and BD or classify UD, BD, and HC. 11 studies did not utilize digital phenotyping at all. Another 11 studies used digital phenotyping to track emotional states or make predictions. The remaining 11 studies were either protocols, introductions to machine learning methods, or reviews. These exclusions ensured that the final selection aligned precisely with the study's specific objectives. The final sample comprised 21 articles, with the earliest study dating back to 2016, emphasizing the novelty of this research topic. Of these, 11 studies directly distinguished between UD and BD (Table 1-2), while 10 studies classified UD, BD, and HC (Table 3-4). It is important to note that studies based on the RADMIS trials or those using the CHI-MEI mood disorder database may have partly overlapping populations. However, they were not excluded due to their very different methods and results. Additionally, although studies using the CHI-MEI database did not explicitly report the diagnostic criteria or psychiatric assessments, data collection was carried out in collaboration with clinicians at Chi Mei Medical Center. As such, these studies were not excluded.

Figure 1 Flowchart of review process and study selection



BD, bipolar disorder; HC, healthy control; UD, unipolar depression.

Table 1 Characteristics of included studies that directly distinguish between UD and BD.

Study	Region	Population	Sample size	Female %	Mean age	Diagnostic criteria or psychiatric assessments
Smartphone apps						
(Faurholt-Jepsen et al, 2022 [24])	Denmark	UD (n = 75) BD (n = 65)	140	UD 52.7 BD 61.1	UD 44.3 BD 44.1	ICD-10
(Faurholt-Jepsen et al, 2022 [25])	Denmark	UD (n = 48) BD (n = 121)	169	UD 60.0 BD 60.0	UD 45.6 BD 35.7	ICD-10
(Langholm et al [26], 2023)	USA	UD (n = 53) BD (n = 28)	81	NG	NG	MINI
(Faurholt-Jepsen et al [27], 2023)	Denmark	UD (n = 58) BD (n = 316)	374	UD 48.20 BD 67.72	UD 43.34 BD 38.50	ICD-10
(Faurholt-Jepsen et al [28], 2024)	Denmark	UD (n = 74) BD (n = 64)	138	UD 52.7 BD 61.1	UD 44.3 BD 44.1	ICD-10
(Faurholt-Jepsen et al [29], 2025)	Denmark	UD (n = 59) BD (n = 47)	106	UD 47.5 BD 51.7	UD 44.5 BD 41.6	ICD-10
Wearable devices						
(Tanaka et al [30], 2018)	Japan	UD (n = 59) BD depression (n = 35)	94	UD 69.5 BD 34.3	UD 53.7 BD 46.9	DSM-IV
Audiovisual recordings						
(Yang et al [31], 2017)	Taiwan	UD (n = 13) BD (n = 13)	26	NG	NG	NG
(Horigome et al [32], 2020)	Japan	UD (n = 17) BD (n = 14)	31	UD 52.9 BD 50.0	UD 57.0 BD 61.0	DSM-5
(Yamamoto et al [33], 2020)	Japan	UD (n = 84) BD (n = 68)	152	UD 56.0 BD 55.9	UD 50.0 BD 55.1	DSM-5
(Pan et al [34], 2023)	China	UD actively symptomatic state (n = 22) BD euthymic state (n = 22)	44	UD 50.0 BD 50.0	UD 30.94 BD 29.58	DSM-IV

BD, bipolar disorder; DSM, Diagnostic and Statistical Manual of Mental Disorders; ICD, International Classification of Diseases; MINI, Mini International Neuropsychiatric Interview; NG, not given; UD, unipolar depression.

Table 2 Methodological details of included studies that directly distinguish between UD and BD.

Study	Data recording	Data utilized	Data preprocessing	Specific variable/ feature selection	Machine learning models/ statistic test	Validation	Findings
Smartphone apps							
(Faurholt-Jepsen et al, 2022 [24])	6 months	Active data (self-assessments) and passive data (location information) with smartphone app (Monsenso	Minimum of 50 location samples per day; exclusion of points with unrealistic acceleration	Stops, moves, places, routine index, radius of gyration, location entropy	BBC	10-fold stratified cross-validation	Discrimination between patients with UD and patients with BD (based on passive data): ● UD all vs. BD all: AUC = 0.75 ● UD euthymic state

		system)					vs. BD euthymic state: AUC = 0.79 ● UD depressive state vs. BD depressive state: AUC = 0.79
(Faurholt-Jepsen et al, 2022 [25])	1~972 days	Naturalistic phone calls voice collected from smartphone app (Monsenso system)	Removal of voice data without a corresponding patient-reported smartphone-based data entry of either mood, activity, or sleep	Acoustic features such as pitch, loudness and energy	RF	5-fold cross-validation	Discrimination between patients with UD and patients with BD: ● UD all vs. BD all: ACC = 0.73 AUC = 0.58 ● UD euthymic state vs. BD euthymic state: ACC = 0.76 AUC = 0.43 ● UD depressive state vs. BD depressive state: ACC = 0.66 AUC = 0.48
(Langholm et al [26], 2023)	12 weeks	Active data (self-reported survey) and passive data (user activity, geolocation, motion, exercise and device rotation) with smartphone app (mindLAMP)	Exclusion of bins with data quality below 0.8; imputation of missing values using mean feature values	Home time, entropy, sleep duration, and screen duration, survey scores	SVM LR DT	Repeated (n=3) Stratified K-fold (k=5)	Binary classification results of UD vs. BD: ● All features: ACC = 57.1 AUC = 0.61 ● Active features: ACC = 64.3 AUC = 0.62 ● Passive features: ACC = 57.1 AUC = 0.61
(Faurholt-Jepsen et al [27], 2023)	6 months	Active data (self-assessments) (Monsenso system)	Missing at random for missing data	Patient-reported daily evaluations about irritability, mood, activity, sleep, stress, and anxiety	Mixed effects regression models	-	Patients with UD spent a significantly higher proportion of time with presence of irritability as compared with patients with BD (depressive state).
(Faurholt-Jepsen et al [28], 2024)	6 months	Active data (self-assessments) and passive data (phone calls, text messages and screen) with smartphone app (Monsenso system)	Removal of entire day's data without a single reading for a day; imputation techniques for missing data	Number or duration of phone calls, text messages and screen usage	RF	Leave-one-patient-out cross validation	Discrimination between patients with UD and patients with BD (based on passive data): ● UD all vs. BD all: AUC = 0.48 ● UD euthymic state vs. BD euthymic state: AUC = 0.46 ● UD depressive state vs. BD depressive state: AUC = 0.42
(Faurholt-Jepsen et	6 months	Active data (self-	Missing at random for missing data	Patient-reported daily mood and	Linear mixed-	-	Patients with BD presented with lower level

al [29], 2025)		assessments) (Monsenso system)		activity scales	effects regression models		of activity as compared with patients with UD (overall, euthymic state and depressive state); there were no differences in mood and activity instability between two groups.
Wearable devices							
(Tanaka et al [30], 2018)	3 weeks	Daytime activity data with wearable activity trackers (Actiwatch)	Exclusion of participants with missing values (days with zero activity across all epochs)	5 principal component such as total amount of activity and activity ratio	PCA	-	Several temporal patterns of intraday activities were associated with the differences between UD and BD: ● BD showed high activity pattern in the morning and low activity pattern in evening ● UD showed low activity pattern in the morning and high activity pattern in evening
Audiovisual recordings							
(Yang et al [31], 2017)	1 day	Facial expressions and speech responses from interviews with a clinician after participants watched six videos	Application of a domain adaptation method called hierarchical spectral clustering based denoising autoencoder	384-dimensional acoustic features, 98-dimensional facial feature vector	HMM CHMM	13-fold cross validation	Classification of patients with UD vs. patients with BD: optimal ACC = 65.38%
(Horigome et al [32], 2020)	NG	Body motion with a Red-Green-Blue-Depth sensor	Nonlinear fitting with smoothing spline; splitting data into smaller segments for missing data > 5 seconds	7 types of features such as position, speed, acceleration, and jerk at four body joints	Chi-square test; One-way ANOVA; Tukey's honestly significant difference test	-	No significant difference in any head motion features between the UD and BD groups.
(Yamamoto et al [33], 2020)	10 minutes	Non-structured interviews (non-specific topics) with a research psychiatrist or psychologist	Exclusion of overlapping voice frames and outlier data using interquartile range	Speech rate, pause time, and response time	ANCOVA	-	Patients with UD showed longer response time than patients with BD, but there were no significant differences in speech rate and pause time.
(Pan et al [34], 2023)	NG	Vocal features collected from four vocal tasks: video watching, text reading, question answering, picture description	Exact gender matching using random sampling; case-control matching within classification tasks	Mel Frequency Cepstral Coefficients	LR	5-fold cross-validation	Classification of patients with UD vs. patients with BD: ● ACC = 0.50 ● AUC = 0.50

ACC, accuracy; ANCOVA, Analyses of Covariance; ANOVA, Analysis of Variance; AUC, area under the curve; AUROC, area under

the receiver operating characteristic; BBC, Balanced Bagging Classifier; BD, bipolar disorder; CHMM, Coupled Hidden Markov Model; DT, Decision Trees; HMM, Hidden Markov Model; LR, Logistic Regression; NG, not given; PCA, Principal Component Analysis; RF, Random Forest; SVM, Support Vector Machine; UD, unipolar depression.

Table 3 Characteristics of included studies that classify UD, BD, and HC.

Study	Region	Population	Sample size	Female %	Mean age	Diagnostic criteria or psychiatric assessments
Wearable devices						
(Anmella et al [35], 2023)	Spain	UD euthymic (n = 2) UD depressed episode (n = 2) BD euthymic (n = 2) BD depressed episode (n = 2) BD manic episode (n = 2) BD mixed episode (n = 2) HC (n = 7)	19	UD 25.0 BD 12.5 HC 57.1	UD 55.5 BD 39.4	DSM-5
(Zakariah and Alotaibi [36], 2023)	Norway	UD (n = 15) BD-1 depression (n = 1) BD-2 depression (n = 7) HC (n = 32)	55	UD 46.7 BD-1 0 BD-2 42.9 HC 62.5	UD 25-69 BD-1 40-44 BD-2 20-64 HC 20-69	DSM-IV
Audiovisual recordings						
(Yang et al [37], 2016)	Taiwan	UD (n = 13) BD (n = 13) HC (n = 13)	39	69.2	NG	NG
(Su et al [38], 2017)	Taiwan	UD (n = 12) BD (n = 12) HC (n = 12)	36	NG	NG	NG
(Hong et al [39], 2018)	Taiwan	UD (n = 12) BD (n = 12) HC (n = 12)	36	NG	NG	NG
(Huang et al [40], 2019)	Taiwan	UD (n = 15) BD (n = 15) HC (n = 15)	45	UD 73.3 BD 66.7 HC 66.7	UD 45.8 BD 46.4 HC 31.6	NG
(Su et al [41], 2020)	Taiwan	UD (n = 13) BD (n = 13) HC (n = 13)	39	69.2	NG	NG
(Hong et al [42], 2021)	Taiwan	UD (n = 12) BD (n = 12) HC (n = 12)	36	NG	NG	NG
(Luo et al [43], 2024)	China	UD (n = 50) BD (n = 50) HC (n = 50)	150	UD 64.0 BD 74.0 HC 58.0	UD 14.48 BD 14.24 HC 14.55	DSM-5
Multimodal technology						
(Wu et al [44], 2024)	Taiwan	UD mild (n = 31) UD moderate (n = 48) UD severe (n = 46) BD (n = 25) HC (n = 18)	168	UD mild 67.7 UD moderate 89.6 UD severe 82.6 BD 88.0 HC 16.7	UD mild 14.88 UD moderate 15.58 UD severe 14.96 BD 15.23 HC 23.67	ICD-10

BD, bipolar disorder; DSM, Diagnostic and Statistical Manual of Mental Disorders; HC, healthy control; ICD, International Classification of Diseases; NG, not given; UD, unipolar depression.

Table 4 Methodological details of included studies that that classify UD, BD, and HC.

Study	Data recording	Data utilized	Data preprocessing	Specific variable/ feature selection	Machine learning models/ statistic test	Validation	Findings
Wearable devices							
(Anmella et al [35], 2023)	2 days	Physiological data with wearable devices (Empatica E4)	Rules-based filter for invalid physiological data; time unit set to 1 second	X/Y/Z-axis acceleration, blood volume pulse, electrodermal activity, heart rate and skin temperature	BiLSTM	NG	7-class classification task: ● ACC = 0.7000 ● AUROC = 0.6900
(Zakariah and Alotaibi [36], 2023)	5~20 days	General levels of activity with a wearable Actiwatch	Imputation techniques (mean imputation/median imputation/regression-based imputation) for missing values; transformed categorical variables into numerical representations	Motor activity measurement from the Actiwatch	UMAP NN	Leave-One-Out Validation	4-class classification task: ● ACC = 0.991 ● F1-score: 0.9887
Audiovisual recordings							
(Yang et al [37], 2016)	1 day	Speech responses from 5 questions after participants watched six videos	Silence removal and speech segmentation based on energy and spectral centroid as features for threshold definition	39 dimensions of Mel-frequency cepstral coefficients; acoustic features of 384 dimensions	SVM MLP LSTM BiLSTM	13-fold cross validation	3-class classification task: optimal ACC = 76.92%
(Su et al [38], 2017)	1 day	Facial expressions elicited by six emotional video clips	Select time interval and segment each facial image into twelve mutually independent facial regions	8 basic orientations of motion vector in microscopic facial expression	HMM LSTM	12-fold cross validation	3-class classification task: optimal ACC = 67.7%
(Hong et al [39], 2018)	1 day	Facial expressions elicited by six emotional video clips	Select time interval and facial points were aligned to a new coordinate	12 action units	MLP SVM GMM LSTM	12-fold cross validation	3-class classification task: optimal ACC = 61.10%
(Huang et al [40], 2019)	1 day	Speech responses from interviews with a clinician after participants watched six videos	Utilization of hierarchical spectral clustering algorithm for database adaptation	32-dimensional acoustic features	SVM CNN LSTM	Leave-one cross-validation	3-class classification task: optimal ACC = 75.56%
(Su et al [41], 2020)	1 day	Facial expressions and speech responses from interviews with a	Hierarchical spectral clustering and denoising autoencoder method for databases adaptation	384 acoustic features; 49 facial expression feature points	SVM HMM MP GRU CNN RNN LSTM	13-fold cross validation	3-class classification task: optimal ACC = 76.9%

		clinician after participants watched six videos					
(Hong et al [42], 2021)	1 day	Facial expressions elicited by six emotional video clips	Selection of four 4-second intervals per elicitation video based on facial expression intensity of all subjects	Action units for macroscopic facial expressions; motion vectors for microscopic facial expressions	MP NN LSTM	12-fold cross-validation	3-class classification task: optimal ACC = 72.2%
(Luo et al [43], 2024)	1 day	Voice signals collected from seven pieces of reading materials	Power normalization; speech segmentation into seven parts corresponding to the seven reading materials	120 vocal features for classification such as the mean value of root mean square energy	DT NB SVM KNN EL CNN	NG	3-class classification task: optimal ACC = 95.6%
Multimodal technology							
(Wu et al [44], 2024)	1 day	Text, audio, facial attributes, heart rate, and eye movement with mobile devices while participants have a conversation with a virtual assistant	NG	Word embedding; 5 spectral features; facial attribute embedding; 23 heart rate variability indices; 7 eye movement features (fixation, saccade, etc.)	RF LSTM DT	5-fold cross validation	5-class classification task: optimal ACC = 0.90269

ACC, accuracy; AUROC, area under the receiver operating characteristic; BD, bipolar disorder; BiLSTM, Bidirectional Long Short-Term Memory; CNN, Convolutional Neural Network; DT, Decision Trees; EL, Ensemble Learning; GRU, Gated Recurrent Unit; HMM, Hidden Markov Model; KNN, K-Nearest Neighbor; LSTM, Long Short-Term Memory; MP, Multilayer Perceptron; NB, Naïve Bayes; NN, Neural Network; RF, Random Forest; RNN, Recurrent Neural Network; SVM, Support Vector Machine; UD, unipolar depression; UMAP, Uniform Manifold Approximation and Projection.

Synthesized Findings

In this subsection, we present the key findings relevant to this systematic review for each study. The studies are categorized according to the type of digital devices used, with some also detailing the clinical staging of UD or BD. Audio and/or visual recordings were the most commonly employed tools in this area, utilized in 11 studies. Additionally, 3 studies used wearable devices, while 6 utilized smartphone applications to collect digital phenotyping data for distinguishing between UD and BD. Only one study employed a multimodal approach, collecting diverse information such as text, audio, facial expressions, heart rate, and eye movement during participant interactions with a virtual assistant. This study was classified as “multimodal technology.” Among the 11 studies that directly distinguished between UD and BD, eight described the mood states of UD and BD patients (depressive, (hypo)manic, mixed, or euthymic state). Of these, 5 studies classified depressive and euthymic states based on participants’ self-reported scores [24,25,27–29], while the other 3 studies assessed mood states through genera clinical interviews [30], Hamilton Depression Scale (HAMD-17) and Young Mania Rating Scale (YMRS) [32], or confirmation of clinical staging by clinicians [34].

Smartphone apps

The six studies included, which are based on smartphone apps, directly distinguish between UD and BD. Among these, five studies [24,25,27–29] predominantly employed the Monsenso system—a smartphone-based monitoring platform installed on patients' personal devices (compatible with both iPhone and Android). The system gathered data through daily patient-reported entries, including subjective information such as mood, sleep, and activity levels. Additionally, it automatically collected objective data from smartphone sensors, such as phone usage patterns, mobility metrics, and voice features. These five studies were part of the RADMIS trials [45], a pragmatic, investigator-blinded, randomized controlled trial. In the intervention group, patients received the Monsenso system enabled patients to self-monitor their symptoms, access psychoeducational resources, and engage with cognitive modules. In contrast, the control group received standard treatment alone. The trial spanned six months, with outcome assessments conducted at baseline (0 months), 3 months, and 6 months. Another study utilized the mindLAMP app [26], which collected real-time data, including geolocation, accelerometer readings, and screen-state (on/off) information. Participants received notifications three times per week, prompting them to complete in-app surveys measuring self-reported depression (Patient Health Questionnaire-2, PHQ-2) and anxiety (Generalized Anxiety Disorder 2-item, GAD-2). In contrast, the Monsenso system used a different scale for patient-reported, smartphone-based mood evaluation. For patients experiencing depressive states, mood scores ranged from -3 to -1, with a neutral mood (euthymia) defined as a self-reported score between -0.5 and +0.5. Two studies, which used only participants' daily active data from self-assessment questionnaires via smartphone apps, revealed that patients with UD spent a significantly higher proportion of time with irritability compared to patients with BD (depressive state) [27]. In contrast, patients with BD exhibited a lower level of activity compared to those with UD (overall, euthymic state and depressive state) [29]. The remaining four studies employed various machine learning models. One study [24] demonstrated that using a Balanced Bagging Classifier with combined smartphone location data effectively distinguished BD from UD during both depressive and euthymic states, achieving a high AUC of 0.79. Another study [25] found that voice features could differentiate BD from UD with low sensitivity but relatively high specificity. A third study [26] employed Logistic Regression to classify UD and BD, reporting a best test accuracy of 64.3% and a test AUC of 0.62. The final study [28] applied a Random Forest model using combined sensor-based smartphone data, where leave-one-out cross-validation yielded a sensitivity of 0.54, specificity of 0.44, and an AUC of 0.48.

Wearable devices

Wearable devices/sensors are tools that monitor physiological and behavioral data, such as heart rate, movement, and temperature, enabling continuous, non-invasive tracking of health and activity patterns. Such tools were used in 3 studies, with two studies classifying UD, BD, and HC, and one study directly distinguishing between UD and BD. Two studies utilized the Actiwatch, a lightweight wrist-worn accelerometer, to differentiate between UD and BD [30] or classify UD, BD and HC [36]. The device recorded participants' activity in 2-minute epochs over several weeks, capturing data on sleep/activity patterns. Another study [35] utilized the research-grade wearable device Empatica E4 to collect physiological data across multiple channels, including acceleration, skin temperature, blood volume pulse, heart rate, and electrodermal activity. One study [30] used Principal Component Analysis to identify distinct temporal patterns of intraday activities differentiating BD from UD, including significant differences in activity patterns (e.g., morning hyperactivity in BD vs. morning hypoactivity in UD). Another small-sample study [36] achieved an accuracy of 0.991 and an F1-score of 0.9887 in a four-class classification task (UD/BD-1/BD-2/HC) using machine learning. The final study [35] utilized physiological data collected by the wearable device Empatica E4 to classify seven groups, including different episodes of UD and BD, remission phases of both disorders, and

HC, achieving an accuracy of 0.7000 and an F1-score of 0.6927.

Audiovisual recordings

A total of 11 studies utilized audiovisual recordings, with 4 studies directly distinguishing between UD and BD, while the remaining studies focused on the UD, BD, and HC classification task. Among these, five studies examined speech alone, three studies combined speech with facial expressions, two studies investigated only facial expressions, and one study focused exclusively on body motion. Seven of the studies [31,37–42] were based on the CHI-MEI mood disorder database, but they differed in the types of data used or the machine learning classification methods applied. Unlike studies utilizing smartphone apps or wearable devices/sensors, research involving audiovisual recordings collected participants' speech, facial expressions, and even body motion within a single day using a fixed paradigm. Studies based on speech analysis extracted a variety of voice features, including strictly acoustic features (e.g., speech rate, pause duration, response time) [33], prosodic features (e.g., pitch, energy, formants) [43,46], and spectral features (e.g., Gamma-tone Frequency Cepstral Coefficients [GFCC] and Mel-Frequency Cepstral Coefficients [MFCC]) [34,47]. Four studies [31,37,40,41] extracted emotion profiles from speech, which represent the local intensity of emotions for each speech response corresponding to a specific question. These emotion profiles were used for single-modality analysis or integrated with facial expressions for fusion analysis in machine learning classification tasks. For the facial expression modality, four studies [31,39,41,42] analyzed facial action units (AUs). Two of these studies additionally examined motion vectors (MVs) [38,42], which were used to capture subtle changes in facial expressions at a microscopic level. Among the four studies that directly distinguished between UD and BD, a single study explored the correlation between body motion and mood disorders [32], finding no significant differences in any head motion features between the UD and BD groups. The remaining three studies employed machine learning methods or analyses of covariance (ANCOVA) to differentiate between the two groups. In the binary classification task, one study reported specificity and accuracy of 0.60 and 0.50, respectively, after extracting i-vectors from MFCCs [34]. Another study, which combined facial expressions and speech for fusion analysis, achieved an optimal accuracy of 65.38% using the Coupled Hidden Markov Model [31]. Furthermore, one study found that patients with UD exhibited longer response times compared to those with BD, although no significant differences were observed in speech rate or pause time [33]. The remaining seven studies all employed machine learning methods (including Long Short-Term Memory, Random Forest, Support Vector Machines, etc.) to classify UD, BD and HC. The model accuracies ranged from 61.10 % to 95.6 %.

Multimodal technology

Only one study was categorized as "multimodal technology". Although this study also collected digital phenotyping data via smartphones, its approach differed from the previously mentioned "smartphone apps" section (as it did not involve long-term data collection over weeks or even months). Therefore, it was categorized separately. This study developed a virtual assistant for automatic depression-level stratification on mobile devices [44]. The assistant actively engages users through voice-based dialogues and dynamically adjusts conversation content based on emotion perception. During the interaction, multimodal features—including text, audio, facial expressions, heart rate, and eye movement—are extracted to facilitate precise depression-level stratification. The study employed a feature-level fusion framework to integrate five modalities with a deep neural network for classifying five groups, including UD (with mild, moderate, and severe levels), BD and HC. Using outcome data from 168 subjects, the experimental results demonstrated that the feature-level fusion of all five modalities achieved the highest overall accuracy of 90.26%.

Quality assessment

The results of the methodological quality assessment using the QUADAS-2 tool were presented in

Multimedia Appendix 3. The risk of bias in the included studies is considerable and cannot be overlooked. A detailed table summarizing the QUADAS-2 scores for each study is provided in Multimedia Appendix 4.

Discussion

Principal findings

We conducted a systematic review of original articles from both journals and conference proceedings, investigating the use of portable or wearable digital tools for distinguishing between UD and BD, as well as for classifying UD, BD, and HC. A total of 21 articles were included, categorized into four main groups based on the type of digital tool assessed: 1) smartphone apps for collecting active data such as mood self-assessments or passive data such as location, naturalistic phone calls, device rotation, and more; 2) wearable sensors, including the Actiwatch and the research-grade wearable device Empatica E4; 3) audiovisual recordings for analyzing speech characteristics, facial expressions, and upper body movements; and 4) multimodal technology, which combine text, audio, facial expressions, heart rate, and eye movement data. Overall, digital phenotyping data offer significant opportunities for advancing the differential diagnosis of mood disorders. Despite certain methodological limitations, our findings highlight the potential of these digital technologies to provide more precise and objective support in diagnosing mood disorders. Certain features, such as activity levels captured through smartphone apps or wearable devices, emerged as potential markers for directly distinguishing between UD and BD. Individuals with BD typically exhibited lower activity levels compared to those with UD. Moreover, BD patients tended to show higher activity levels in the morning and lower activity in the evening, whereas UD patients displayed the opposite pattern. Additionally, approaches leveraging speech modalities or integrating multiple modalities achieved better classification performance across the UD, BD, and HC groups, although the specific contributing features remain unclear.

Diagnostic confusion between BD and UD often occurs when patients are in a remitted or depressive state, as patients and their families may fail to recall previous (hypo-)manic episodes, making it challenging for clinicians to make an accurate diagnosis [24]. In the six included studies that used smartphone apps to directly distinguish between UD and BD, five studies using the Monsenso system considered the clinical staging of both groups. The five studies used different types of digital phenotyping data to differentiate between UD and BD. Among them, three studies employed machine learning algorithms, with the best AUC of 0.79 achieved in both euthymic and depressive states (based on location data) [24]. However, results from two other studies showed lower AUC values (0.42–0.58), which may indicate that variations in smartphone-based digital phenotyping are highly individualized [28]. Four studies utilized participants' active data, including daily self-assessments of irritability, mood, activity, sleep, stress, anxiety and naturalistic phone calls voice. Half of these studies used mixed-effects models, revealing differences in the level of activity [29] and presence of irritability [27] between UD and BD in a depressive state.

Interestingly, among the three studies that used wearable devices to collect digital phenotyping data, one study also revealed differences in activity patterns between UD and BD. In contrast to the study mentioned above that used daily self-report activity scales, this study collected participants' activity levels using activity trackers. This study provided a more detailed insight into the activity patterns of UD and BD: BD showed a high activity pattern in the morning and a low activity pattern in the evening, while UD exhibited the opposite [30]. Another study based on wearable watches also considered activity levels as a digital phenotype, but included HC as well. Using machine learning

for a four-class classification of UD, BD-1, BD-2, and HC, the study achieved an accuracy of 0.991 [36]. These three studies all demonstrate the potential of activity levels in the differential diagnosis of mood disorders.

Among the studies utilizing audiovisual recordings and multimodal technologies to classify UD, BD, and HC, all involved one-time data collection from participants in a standardized environment. Overall, approaches based on speech modalities or the integration of multiple modalities demonstrated better classification performance across these three groups, with accuracy ranging from 75.56% to 95.6% [37,40,41,43,44]. However, the specific features contributing to the classification performance remained unclear.

Accordingly, it can be concluded that smartphone apps and wearable devices, when combined with machine learning or other advanced analytical methods, demonstrate moderate diagnostic potential for differentiating UD and BD. However, their effectiveness is often limited by substantial individual variability in digital phenotyping data. Similarly, studies using audiovisual recordings and multimodal technology with machine learning have shown mixed outcomes—while they exhibit promising results in multi-class classification tasks, they have had limited success in directly distinguishing between UD and BD. These findings suggest that, although digital phenotyping holds potential for supporting the differential diagnosis of mood disorders, it should complement, rather than replace, comprehensive clinical evaluations conducted by trained professionals.

In particular, digital phenotyping data collected via smartphones or wearable devices/sensors may be more suitable for long-term dynamic monitoring and even intervention in the real world. This advantage lies in its ability to continuously gather active and/or passive data, offering a more comprehensive understanding of an individual’s mental and physical state. By capturing data in real-time, these approaches help mitigate recall biases commonly associated with other symptom-reporting methods [48]. Moreover, they address the issue of “back-filling”, a frequent problem with paper-based diaries [49], thereby improving the reliability and accuracy of symptom tracking. However, factors such as patients’ proficiency in using electronic devices, the need for continuous and uninterrupted use, and potential limitations in monitoring continuity due to lack of feedback must be considered [50], as these could impact its medical value. In contrast, multimodal data collected through fixed paradigms in controlled settings offer advantages such as capturing explicit behaviors (e.g., speech, facial expressions, gestures) and implicit physiological signals (e.g., eye movements, heart rate). Standardized protocols ensure reliable and reproducible data, making this structured approach particularly effective for mood disorder detection and classification. For instance, in AVEC 2013 [51] and AVEC 2018 [52], audiovisual signals have been proven to play a significant role in the detection and classification of UD and BD.

Our systematic review primarily focused on criterion validity and content validity. Criterion validity, defined as the extent to which the results of a specific test align with those of a reference standard [53], was demonstrated in the included studies by comparing classifications or distinctions based on digital phenotyping with diagnostic outcomes from professional medical evaluations. Content validity, on the other hand, refers to the degree to which an assessment tool is relevant to and representative of the targeted construct it aims to measure [54]. The included studies primarily focused on capturing one or a few dimensions of emotions (e.g., daily mood, social interactions, or activity levels), falling short of comprehensively reflecting all core features of mood disorders. Furthermore, the complexity and vastness of digital phenotyping data led most studies to employ machine learning algorithms for classification tasks, which posed significant challenges for the evaluation of discriminant validity and structural validity.

Future directions

The article that first introduced the concept of digital phenotyping stated that data collected through social media, forums, online communities, wearable technologies, and mobile devices offers substantial value that goes beyond traditional methods such as laboratory tests and clinical imaging [11]. This is particularly significant given the unclear pathogenesis, the unknown etiologies, and the lack of objective biomarkers of mood disorders. As a result, digital phenotyping stands out as a promising tool in supporting mood disorder diagnosis. We should also recognize that as medical technology advances, the types of digital phenotyping data are likely to expand. Beyond the digital devices discussed in our review, smart speakers may also play a role in future applications [55]. Besides, traditional laboratory tests and imaging methods are becoming more portable, including smartwatch-based 9-lead electrocardiograms [56], wearable electroencephalography [57], and even devices for long-term monitoring of metabolic indicators, such as continuous glucose monitor [58]. This evolution suggests that, in the future, more complex and extensive multimodal data may be used for the auxiliary diagnosis of mood disorders. However, the growing diversity of data types increases the demand for more sophisticated data analysis and processing techniques. To meet these challenges, integrating multimodal data from various sources could provide a more comprehensive understanding of mood disorders and expand the potential applications of digital phenotyping in mental health.

Despite the promise of digital phenotyping, we observed that this field is still relatively new, with a limited number of studies conducted thus far. Most of the research in this area involves small sample sizes and lacks consistency in evaluation standards and procedures. This makes it difficult to draw generalizable conclusions and underscores the need for further research to refine methodologies, improve data quality, and establish robust analytical frameworks. Therefore, it is crucial to develop standardized methodologies, guidelines, and protocols in the field of digital phenotyping to ensure consistency and reliability in data collection, analysis, and interpretation. Establishing these standards will enable better comparisons across studies, improve reproducibility, and ultimately enhance the clinical applicability of digital phenotyping for the diagnosis and monitoring of mood disorders.

Undoubtedly, digital phenotyping plays a significant role in the prevention, diagnosis, treatment, and management of mood disorders and even mental disorders in general. However, several key requirements must be met to transform the way mental health care is delivered [59]. First, given the unique nature of mental disorders, patient privacy and data security require greater attention than for the general population. Patients must be fully informed and consent to how their data is collected, managed, and utilized, as this is crucial to maintaining trust between patients and clinicians [16]. Moreover, the “black box” nature of machine learning models continues to raise long-term concerns among healthcare professionals [60], as these models obscure their decision-making processes [12]. Such opacity limits the clinical actionability of machine learning models, particularly for classification tasks, as clinicians often require clear explanations to trust and act upon this prediction. This underscores the importance of explainable AI [61], which aims to make the reasoning behind machine learning conclusions transparent, thus enhancing trust among healthcare professionals and ensuring that patients maintain greater control over their health data. However, the integration of digital tools into mental health care also faces broader social challenges. For example, these technological advances may exacerbate existing inequalities in access to psychiatric care, particularly disadvantaging underprivileged populations who lack access to the necessary technology or infrastructure [62]. Furthermore, the stigma surrounding mental illness [63], which already acts as a significant barrier to seeking help, may be compounded by the perceived invasiveness of digital health tools, discouraging their adoption by those most in need.

Strengths and limitations

To the best of our knowledge, this is the first systematic review to comprehensively synthesize existing evidence on distinguishing BD from UD based on digital phenotyping. By systematically evaluating studies employing smartphone apps, wearable devices, audiovisual recordings, and multimodal technologies, this review highlights the emerging potential of digital phenotyping as a supportive tool for the differential diagnosis of mood disorders. However, our systematic review has several limitations. First, many of the included studies had small sample sizes, which increases the risk of overfitting and limits the generalizability of the findings. Future research should involve larger and more diverse samples to improve model robustness and applicability. Second, none of the studies validated their findings using external datasets, raising concerns about the reliability and reproducibility of the results. Incorporating independent datasets for external validation is essential to ensure consistent performance across different populations. Third, there was substantial heterogeneity in study designs, including variations in data sources, analytical methods, and outcome measures. This heterogeneity complicates direct comparisons and weakens the strength of synthesized conclusions. To address this, future studies should adopt standardized methodologies and outcome definitions to enhance comparability.

Conclusions

In conclusion, our review shows that digital phenotyping for distinguishing UD and BD is progressing rapidly. However, challenges like privacy, data security, and equitable access must be addressed for digital healthcare to effectively transform mental health care. Only by overcoming these challenges can digital innovations fulfill their potential to improve mental health care inclusivity.

Acknowledgments

We would like to thank all participants who took part in this study. This work was financially supported by the Funding Project of Clinical Research Center of Shanghai Mental Health Center Key Project (CRC2021DX01).

Conflict of interest

We have no conflicts of interest concerning this paper.

References

1. Le GH, Wong S, Au H, Badulescu S, Gill H, Vasudeva S, Teopiz KM, Rhee TG, Ho R, Kwan ATH, Mansur RB, Rosenblat JD, McIntyre RS. Association between rumination, suicidal ideation and suicide attempts in persons with depressive and other mood disorders and healthy controls: A systematic review and meta-analysis. J Affect Disord Netherlands; 2025 Jan 1;368:513–527. PMID:39303880

2. Angst J, Ajdacic-Gross V, Rössler W. Classification of mood disorders. Psychiatr Pol Poland; 2015;49(4):663–671. PMID:26488343

3. Le-Niculescu H, Roseberry K, Gill SS, Levey DF, Phalen PL, Mullen J, Williams A, Bhairo S, Voegtline T, Davis H, Shekhar A, Kurian SM, Niculescu AB. Precision medicine for mood disorders: objective assessment, risk prediction, pharmacogenomics, and repurposed drugs. Mol Psychiatry England; 2021 Jul;26(7):2776–2804. PMID:33828235

4. Datta S, Suryadevara U, Cheong J. Mood Disorders. Contin Minneap Minn 2021 Dec;27(6):1712–1737. doi: 10.1212/con.0000000000001051

5. Hirschfeld RMA, Lewis L, Vornik LA. Perceptions and impact of bipolar disorder: How far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 Survey of individuals with bipolar disorder. J Clin Psychiatry 2003;64(2):161–174. doi: 10.4088/JCP.v64n0209

6. Judd LL, Akiskal HS, Schettler PJ, Endicott J, Maser J, Solomon DA, Leon AC, Rice JA, Keller MB. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry 2002;59(6):530–537. doi: 10.1001/archpsyc.59.6.530

7. Baldessarini RJ, Faedda GL, Offidani E, Vázquez GH, Marangoni C, Serra G, Tondo L. Antidepressant-associated mood-switching and transition from unipolar major depression to bipolar disorder: a review. J Affect Disord Netherlands; 2013 May 15;148(1):129–135. PMID:23219059

8. Drancourt N, Etain B, Lajnef M, Henry C, Raust A, Cochet B, Mathieu F, Gard S, Mbailara K, Zanouy L, Kahn JP, Cohen RF, Wajsbrot-Elgrabli O, Leboyer M, Scott J, Bellivier F. Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. Acta Psychiatr Scand United States; 2013 Feb;127(2):136–144. PMID:22901015

9. Post RM, Leverich GS, Kupka RW, Keck PEJ, McElroy SL, Altshuler LL, Frye MA, Luckenbaugh DA, Rowe M, Grunze H, Suppes T, Nolen WA. Early-onset bipolar disorder and treatment delay are risk factors for poor outcome in adulthood. J Clin Psychiatry United States; 2010 Jul;71(7):864–872. PMID:20667291

10. Stensland MD, Schultz JF, Frytak JR. Diagnosis of unipolar depression following initial identification of bipolar disorder: a common and costly misdiagnosis. J Clin Psychiatry United States; 2008 May;69(5):749–758. PMID:18363423

11. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol United States; 2015 May;33(5):462–463. PMID:25965751

12. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, Schoeller F, Mouchabac S, Adrien V. Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. J Med Internet Res Canada; 2023 Oct 4;25:e44502. PMID:37792430

13. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. J Med Internet Res 2023 Dec 13;25:e46778. doi: 10.2196/46778

14. Choudhary S, Thomas N, Ellenberger J, Srinivasan G, Cohen R. A Machine Learning Approach for Detecting Digital Behavioral Patterns of Depression Using Nonintrusive Smartphone Data (Complementary Path to Patient Health Questionnaire-9 Assessment): Prospective Observational Study. JMIR Form Res Canada; 2022 May 16;6(5):e37736. PMID:35420993

15. Galatzer-Levy IR, Onnela J-P. Machine Learning and the Digital Measurement of Psychological Health. Annu Rev Clin Psychol United States; 2023 May 9;19:133–154. PMID:37159287

16. Smith KA, Blease C, Faurholt-Jepsen M, Firth J, Van Daele T, Moreno C, Carlbring P, Ebner-Priemer UW, Koutsouleris N, Riper H, Mouchabac S, Torous J, Cipriani A. Digital mental health: challenges and next steps. BMJ Ment Health England; 2023 Feb;26(1). PMID:37197797

17. Orsolini L, Fiorani M, Volpe U. Digital phenotyping in bipolar disorder: Which integration with clinical endophenotypes and biomarkers? Int J Mol Sci 2020;21(20):1–21. doi: 10.3390/ijms21207684

18. Torous J, Onnela J-P, Keshavan M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Transl Psychiatry United States; 2017 Mar 7;7(3):e1053. PMID:28267146

19. Prakash J, Chaudhury S, Chatterjee K. Digital phenotyping in psychiatry: When mental health goes binary. Ind Psychiatry J India; 2021 Dec;30(2):191–192. PMID:35017799

20. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry Off J World Psychiatr Assoc WPA Italy; 2018 Oct;17(3):276–277. PMID:30192103

21. Insel TR. Digital Phenotyping: Technology for a New Science of Behavior. JAMA United States; 2017 Oct 3;318(13):1215–1216. PMID:28973224

22. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med Public Library of Science; 2009 Jul 21;6(7):e1000097. doi: 10.1371/journal.pmed.1000097

23. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med American College of Physicians; 2011 Oct 18;155(8):529–536. doi: 10.7326/0003-4819-155-8-201110180-00009

24. Faurholt-Jepsen M, Busk J, Rohani DA, Frost M, Tønning ML, Bardram JE, Kessing LV. Differences in mobility patterns according to machine learning models in patients with bipolar disorder and patients with unipolar disorder. J Affect Disord Elsevier B.V.; 2022 Jun 1;306((Faurholt-Jepsen M., maria@faurholt-jepsen.dk; Tønning M.L.; Kessing L.V.) Copenhagen Affective Disorder Research Center (CADIC), Psychiatric Center Copenhagen, Rigshospitalet, Copenhagen, Denmark):246–253. doi: 10.1016/j.jad.2022.03.054

25. Faurholt-Jepsen M, Rohani DA, Busk J, Tønning ML, Vinberg M, Bardram JE, Kessing LV. Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls. Acta Psychiatr Scand Wiley-Blackwell Publishing Ltd.; 2022 Mar 1;145(3):255–267. doi: 10.1111/acps.13391

26. Langholm C, Breitinger S, Gray L, Goes F, Walker A, Xiong A, Stopel C, Zandi P, Frye MA, Torous J. Classifying and clustering mood disorder patients using smartphone data from a feasibility study. Npj Digit Med 2023 Dec 21;6(1):238. doi: 10.1038/s41746-023-00977-7

27. Faurholt-Jepsen M, Busk J, Tønning ML, Bardram JE, Frost M, Vinberg M, Kessing LV. Irritability in bipolar disorder and unipolar disorder measured daily using smartphone-based

data: An exploratory post hoc study. Acta Psychiatr Scand Wiley-Blackwell Publishing Ltd.; 2023 Jun 1;147(6):593–602. doi: 10.1111/acps.13558

28. Faurholt-Jepsen M, Rohani DA, Busk J, Tønning ML, Frost M, Bardram JE, Kessing LV. Using digital phenotyping to classify bipolar disorder and unipolar disorder – exploratory findings using machine learning models. Eur Neuropsychopharmacol Elsevier B.V.; 2024 Apr;81((Faurholt-Jepsen M., maria@faurholt-jepsen.dk; Tønning M.L.; Kessing L.V.) Psychiatric Center Copenhagen, Copenhagen Affective Disorder Research Center (CADIC), Frederiksberg, Denmark):12–19. doi: 10.1016/j.euroneuro.2024.01.003

29. Faurholt-Jepsen M, Busk J, Tonning M, Rohani D, Bardram J, Kessing L. Mood, Activity, and Instability in Bipolar Disorder and Unipolar Disorder-An Exploratory Post Hoc Study Using Digital Data. ACTA Psychiatr Scand John Wiley and Sons Inc; 2025 Mar;151(3):426–433. doi: 10.1111/acps.13771

30. Tanaka T, Kokubo K, Iwasa K, Sawa K, Yamada N, Komori M. Intraday Activity Levels May Better Reflect the Differences Between Major Depressive Disorder and Bipolar Disorder Than Average Daily Activity Levels. Front Psychol Frontiers Media S.A.; 2018 Dec 7;9(DEC):2314. doi: 10.3389/fpsyg.2018.02314

31. Yang T-H, Wu C-H, Huang K-Y, Su M-H. Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. J Ambient Intell Humaniz Comput Springer Verlag; 2017 Nov;8(6):895–906. doi: 10.1007/s12652-016-0395-y

32. Horigome T, Sumali B, Kitazawa M, Yoshimura M, Liang K, Tazawa Y, Fujita T, Mimura M, Kishimoto T. Evaluating the severity of depressive symptoms using upper body motion captured by RGB-depth sensors and machine learning in a clinical interview setting: A preliminary study. Compr Psychiatry W.B. Saunders; 2020 Apr;98((Horigome T.; Kitazawa M.; Yoshimura M.; Liang K.-C.; Tazawa Y.; Mimura M.; Kishimoto T., tkishimoto@keio.jp) Department of Neuropsychiatry, Keio University School of Medicine, Tokyo, Japan):152169. doi: 10.1016/j.comppsych.2020.152169

33. Yamamoto M, Takamiya A, Sawada K, Yoshimura M, Kitazawa M, Liang K, Fujita T, Mimura M, Kishimoto T. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. Hashimoto K, editor. PLOS ONE United States; 2020 Sep 11;15(9 September):e0238726. PMID:32915846

34. Pan W, Deng F, Wang X, Hang B, Zhou W, Zhu T. Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls. Front Psychiatry Frontiers Media SA; 2023 Jul 20;14:1079448. doi: 10.3389/fpsyt.2023.1079448

35. Anmella G, Corponi F, Li BM, Mas A, Sanabra M, Pacchiarotti I, Valentí M, Grande I, Benabarre A, Giménez-Palomo A, Garriga M, Agasi I, Bastidas A, Cavero M, Fernández-Plaza T, Arbelo N, Bioque M, García-Rizo C, Verdolini N, Madero S, Murru A, Amoretti S, Martínez-Aran A, Ruiz V, Fico G, De Prisco M, Oliva V, Solanes A, Radua J, Samalin L, Young AH, Vieta E, Vergari A, Hidalgo-Mazzei D. Exploring Digital Biomarkers of Illness Activity in Mood Episodes: Hypotheses Generating and Model Development Study. JMIR MHealth UHealth Canada; 2023 May 4;11((Anmella G.; Mas A.; Sanabra M.; Pacchiarotti I.; Valentí M.; Grande I.; Benabarre A.; Giménez-Palomo A.; Garriga M.; Agasi I.; Bastidas A.; Cavero M.; Fernández-Plaza T.; Arbelo N.; Bioque M.; García-Rizo C.; Verdolini N.; Madero S.; Murru A.; Amoretti S.; Martínez-Aran A.; Ruiz V.; Fico G.; De Prisco M.; Oliva V.; Vieta E.; Hidalgo-

Mazzei D.) Department of Psychiatry and Psychology, Institute of Neuroscience, Hospital Clínic de Barcelona, Barcelona, Catalonia, Spain):e45405. PMID:36939345

36. Zakariah M, Alotaibi YA. Unipolar and Bipolar Depression Detection and Classification Based on Actigraphic Registration of Motor Activity Using Machine Learning and Uniform Manifold Approximation and Projection Methods. DIAGNOSTICS Multidisciplinary Digital Publishing Institute (MDPI); 2023 Jul;13(14). doi: 10.3390/diagnostics13142323

37. Yang T, Wu C, Huang K, Su M. Detection of Mood Disorder Using Speech Emotion Profiles and LSTM. In: Lee T, Xie L, Dang J, Wang H, Wei J, Feng H, Hou Q, Wei Y, editors. 2016 10th Int Symp Chin Spok Lang Process ISCSLP Institute of Electrical and Electronics Engineers Inc.; 2016. p. 1–5. doi: 10.1109/ISCSLP.2016.7918439

38. Su M, Wu C, Huang K, Hong Q, Wang H. Exploring Microscopic Fluctuation of Facial Expression for Mood Disorder Classification. In: Dong M, Wang L, Lu Y, Li H, editors. 2017 Int Conf Orange Technol ICOT Institute of Electrical and Electronics Engineers Inc.; 2017. p. 65–69. doi: 10.1109/ICOT.2017.8336090

39. Hong Q, Wu C, Su M, Huang K, IEEE. Exploring Macroscopic Fluctuation of Facial Expression for Mood Disorder Classification. 2018 First Asian Conf Affect Comput Intell Interact ACII Asia Institute of Electrical and Electronics Engineers Inc.; 2018. p. 1–6. doi: 10.1109/ACIIAsia.2018.8470337

40. Huang K-Y, Wu C-H, Su M-H. Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. Pattern Recognit Elsevier Ltd; 2019 Apr;88:668–678. doi: 10.1016/j.patcog.2018.12.016

41. M. -H. Su, C. -H. Wu, K. -Y. Huang, T. -H. Yang. Cell-Coupled Long Short-Term Memory With L -Skip Fusion Mechanism for Mood Disorder Detection Through Elicited Audiovisual Features. IEEE Trans Neural Netw Learn Syst Institute of Electrical and Electronics Engineers Inc.; 2020 Jan;31(1):124–135. doi: 10.1109/TNNLS.2019.2899884

42. Hong Q-B, Wu C-H, Su M-H, Chang C-C. Exploring Macroscopic and Microscopic Fluctuations of Elicited Facial Expressions for Mood Disorder Classification. IEEE Trans Affect Comput Institute of Electrical and Electronics Engineers Inc.; 2021 Dec 1;12(4):989–1001. doi: 10.1109/TAFFC.2019.2909873

43. Luo J, Wu Y, Liu M, Li Z, Wang Z, Zheng Y, Feng L, Lu J, He F. Differentiation between depression and bipolar disorder in child and adolescents by voice features. Child Adolesc Psychiatry Ment Health BioMed Central Ltd; 2024 Jan 29;18(1):19. doi: 10.1186/s13034-024-00708-0

44. Wu EH-K, Gao T-Y, Chung C-R, Chen C-C, Tsai C-F, Yeh S-C. Mobile Virtual Assistant for Multi-Modal Depression-Level Stratification. IEEE Trans Affect Comput Institute of Electrical and Electronics Engineers Inc.; 2024;1–14. doi: 10.1109/TAFFC.2024.3451114

45. Tønning ML. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: The RADMIS randomized controlled trial. J Affect Disord 2021;

46. Wanderley Espinola C, Gomes JC, Mônica Silva Pereira J, Dos Santos WP. Detection of major

depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study. Res Biomed Eng Springer Science and Business Media Deutschland GmbH; 2022 Jun 7;38(3):813–829. doi: 10.1007/s42600-022-00222-2

47. Assoc. Prof. Dr., Department of Psychiatry, Inonu University, Faculty of Medicine, Malatya, Turkey, Cansel N, Alcin ÖF, Assoc. Prof. Dr., Department of Software Engineering, Inonu University, Faculty of Engineering, Malatya, Turkey, Yılmaz ÖF, MD, Department of Psychiatry, Yeşilyurt Hasan Çalik Hospital, Malatya, Turkey, Ari A, Assist. Prof. Dr., Department of Computer Engineering, Inonu University, Faculty of Engineering, Malatya, Turkey, Akan M, Assist. Prof. Dr., Department of Psychiatry, Turgut Özal University, Faculty of Medicine, Malatya, Turkey, Ucuz İ, Assoc. Prof. Dr., Department of Child and Adolescent Psychiatry, Inonu University Faculty of Medicine Malatya, Turkey. A new artificial intelligence-based clinical decision support system for diagnosis of major psychiatric diseases based on voice analysis. Psychiatr Danub Medicinska Naklada Zagreb; 2023 Jan 1;35(4):489–499. doi: 10.24869/psyd.2023.489

48. Trull TJ, Ebner-Priemer UW. Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. J Abnorm Psychol United States; 2020 Jan;129(1):56–63. PMID:31868388

49. Piasecki TM, Hufford MR, Solhan M, Trull TJ. Assessing clients in their natural environments with electronic diaries: rationale, benefits, limitations, and barriers. Psychol Assess United States; 2007 Mar;19(1):25–43. PMID:17371121

50. C. -H. Wu, J. -H. Hsu, C. -R. Liou, H. -Y. Su, E. C. -L. Lin, P. -S. Chen. Automatic Bipolar Disorder Assessment Using Machine Learning With Smartphone-Based Digital Phenotyping. IEEE Access Institute of Electrical and Electronics Engineers Inc.; 2023;11:121845–121858. doi: 10.1109/ACCESS.2023.3328342

51. Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, Schnieder S, Cowie R, Pantic M. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proc 3rd ACM Int Workshop Audiov Emot Chall New York, NY, USA: Association for Computing Machinery; 2013. p. 3–10. doi: 10.1145/2512530.2512533

52. Ringeval F, Schuller B, Valstar M, Cowie R, Kaya H, Schmitt M, Amiriparian S, Cummins N, Lalanne D, Michaud A, Ciftçi E, Güleç H, Salah AA, Pantic M. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. Proc 2018 Audiov Emot Chall Workshop New York, NY, USA: Association for Computing Machinery; 2018. p. 3–13. doi: 10.1145/3266302.3266316

53. Piedmont RL. Criterion Validity. In: Michalos AC, editor. Encycl Qual Life Well- Res Dordrecht: Springer Netherlands; 2014. p. 1348–1348. doi: 10.1007/978-94-007-0753-5_618ISBN:978-94-007-0753-5

54. Rusticus S. Content Validity. In: Michalos AC, editor. Encycl Qual Life Well- Res Dordrecht: Springer Netherlands; 2014. p. 1261–1262. doi: 10.1007/978-94-007-0753-5_553ISBN:978-94-007-0753-5

55. Kim D. Can healthcare apps and smart speakers improve the health behavior and depression of older adults? A quasi-experimental study. Front Digit Health Switzerland; 2023;5:1117280.

PMID:36910571

56. Choi J, Kim J, Spaccarotella C, Esposito G, Oh I-Y, Cho Y, Indolfi C. Smartwatch ECG and artificial intelligence in detecting acute coronary syndrome compared to traditional 12-lead ECG. Int J Cardiol Heart Vasc Ireland; 2025 Feb;56:101573. PMID:39687687

57. Newton TJ, Frankel MA, Tosi Z, Kazen AB, Muvvala VK, Loddenkemper T, Spitz MC, Strom L, Friedman D, Lehmkuhle MJ. Validation of a discrete electrographic seizure detection algorithm for extended-duration, reduced-channel wearable EEG. Epilepsia United States; 2025 Mar 19; PMID:40108974

58. Wu Z, Lebbar M, Bonhoure A, Raffray M, Devaux M, Grou C, Messier V, Boudreau V, Vanasse A, Brazeau A-S, Rabasa-Lhoret R. Open-Source Versus Commercial Automated Insulin Delivery System for Type 1 Diabetes Management: A Prospective Observational Comparative Study from Canada. Diabetes Technol Ther United States; 2025 Mar 18; PMID:40100927

59. Bhugra D, Tasman A, Pathare S, Priebe S, Smith S, Torous J, Arbuckle MR, Langford A, Alarcón RD, Chiu HFK, First MB, Kay J, Sunkel C, Thapar A, Udomratn P, Baingana FK, Kestel D, Ng RMK, Patel A, Picker LD, McKenzie KJ, Moussaoui D, Muijen M, Bartlett P, Davison S, Exworthy T, Loza N, Rose D, Torales J, Brown M, Christensen H, Firth J, Keshavan M, Li A, Onnela J-P, Wykes T, Elkholy H, Kalra G, Lovett KF, Travis MJ, Ventriglio A. The WPA-Lancet Psychiatry Commission on the Future of Psychiatry. Lancet Psychiatry 2017;4(10):775–818. doi: https://doi.org/10.1016/S2215-0366(17)30333-4

60. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 2020 Dec 1;32(24):18069–18083. doi: 10.1007/s00521-019-04051-w

61. Balcombe L, De Leo D. Digital Mental Health Challenges and the Horizon Ahead for Solutions. JMIR Ment Health Canada; 2021 Mar 29;8(3):e26811. PMID:33779570

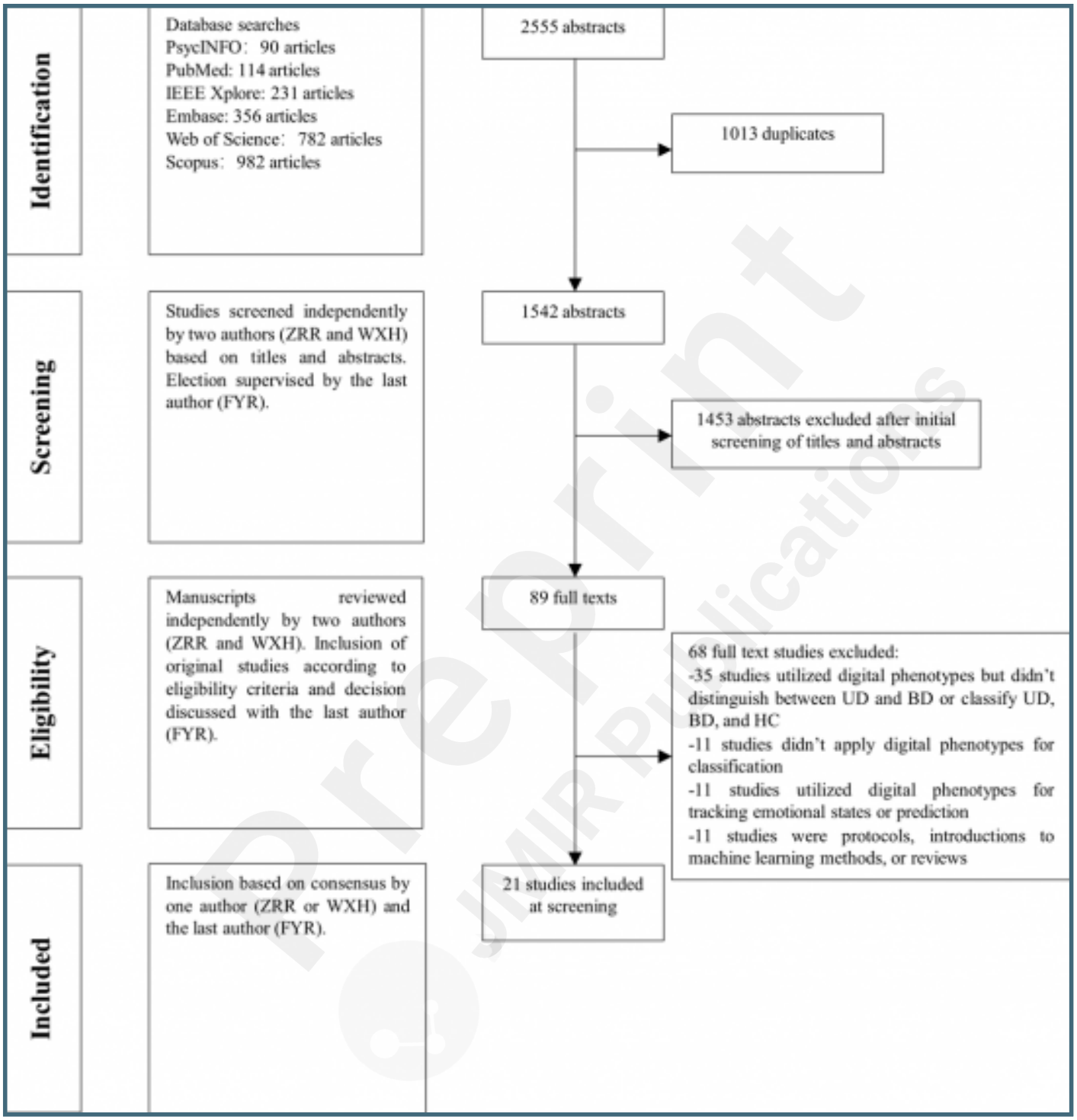
62. Sharon T. Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare. Philos Technol 2017 Mar 1;30(1):93–121. doi: 10.1007/s13347-016-0215-5

63. Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. Am J Public Health United States; 2013 May;103(5):777–780. PMID:23488489

Supplementary Files

Figures

Flowchart of review process and study selection.



Multimedia Appendixes

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
URL: http://asset.jmir.pub/assets/25ed1be94213dc8bddbe3d2ad716e78f.docx

Search strategy.
URL: http://asset.jmir.pub/assets/938366011a7869538163d30d8ecf62f5.docx

Risk of bias in studies included in the systematic review: Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2).
URL: http://asset.jmir.pub/assets/487371f3fee2c40b38ccfeb60c854007.docx

Methodological quality assessment of the studies included in the systematic review through the QUADAS-2.
URL: http://asset.jmir.pub/assets/e74accde6e7800ef00b8adee74ce098f.docx