Diagnosis of depressive symptoms using mobile devices: a systematic review

Diagnóstico de sintomas depressivos por meio de dispositivos móveis: uma revisão sistemática

Diagnóstico de síntomas depresivos mediante dispositivos móviles: una revisión sistemática

Matheus Damasceno[*], Paulo Abrantes[**], André Takahata[***], Antonio Netto[****]

ABSTRACT

Background: depression is a widespread mental health issue, often underdiagnosed and undertreated due to reliance on subjective self-reports and limited access to care. Mobile health (mHealth) technologies, utilizing smartphone and wearable sensors, offer innovative solutions for objective and scalable diagnostics. **Objectives**: this review examines the effectiveness of mobile device sensors in diagnosing depression, identifying relevant biosignals, and exploring diagnostic methods. **Methodology**: a systematic search following PRISMA-DTA guidelines was conducted in MEDLINE/PubMed and Embase, focusing on diagnostic accuracy of mHealth sensors validated against gold standards like DSM-5 or PHQ-9. **Results**: eleven studies showed that accelerometers and heart rate monitors are key in detecting movement, activity, and physiological patterns linked to depression. Machine learning algorithms, especially random forests, achieved high diagnostic accuracy. **Conclusion**: mHealth technologies hold promise for depression diagnostics, but improvements in methodological consistency, sample size, and external validation are necessary for broader clinical use.

Keywords: systematic review; depressive disorder; remote sensors; mHealth

*BSc., Center for Engineering, Modeling and Applied Social Sciences (CECS), Federal University of ABC (UFABC), São Paulo, Brazil – https://orcid.org/0000-0003-3015-509X

**PhD Student., Center for Engineering, Modeling and Applied Social Sciences (CECS), Federal University of ABC (UFABC), São Paulo, Brazil - https://orcid.org/0000-0003-4322-5675

***PhD., Center for Engineering, Modeling and Applied Social Sciences (CECS), Federal University of ABC (UFABC), São Paulo, Brazil - https://orcid.org/0000-0002-7701-6452

****PhD., Paulista School of Medicine, Federal University of São Paulo (UNIFESP), São Paulo, Brazil - https://orcid.org/0000-0001-9215-8531

Corresponding Author:
Antonio Netto
avnetto@unifesp.br

How to cite:
Damasceno, M., Abrantes, P., Takahata, A., & Netto, A. (2025). Diagnosis of depressive symptoms using mobile devices: a systematic review and metaanalysis. *Revista de Investigação & Inovação em Saúde, 8*(2), 1-15. https://doi.org/10.37914/riis.v8i2.439

Received: 11/12/2024
Accepted: 23/04/2025

RESUMO

Enquadramento: a depressão é um problema generalizado de saúde mental, frequentemente subdiagnosticado e subtratado devido à dependência de autorrelatos subjetivos e acesso limitado a cuidados. As tecnologias de saúde móvel (mHealth), utilizando sensores de smartphones e wearables, oferecem soluções inovadoras para diagnósticos objetivos e escaláveis. **Objetivos**: esta revisão examina a eficácia dos sensores de dispositivos móveis no diagnóstico da depressão, identificando biossinais relevantes e explorando métodos de diagnóstico. **Metodologia**: uma busca sistemática seguindo as diretrizes PRISMA-DTA foi conduzida no MEDLINE/PubMed e Embase, com foco na precisão diagnóstica de sensores mHealth validados contra padrões ouro como DSM-5 ou PHQ-9. **Resultados**: onze estudos mostraram que acelerômetros e monitores de frequência cardíaca são essenciais para detectar movimento, atividade e padrões fisiológicos associados à depressão. Algoritmos de aprendizado de máquina, especialmente florestas aleatórias, alcançaram alta precisão diagnóstica. **Conclusão**: as tecnologias mHealth são promissoras para diagnósticos de depressão, mas melhorias na consistência metodológica, tamanho da amostra e validação externa são necessárias para uso clínico mais amplo.

Palavras-chave: revisão sistemática; transtorno depressivo; sensores remotos; mHealth

RESUMEN

Marco contextual: la depresión es un problema de salud mental generalizado, a menudo infradiagnosticado y subtratado debido a la dependencia de autoinformes subjetivos y al acceso limitado a la atención. Las tecnologías de salud móvil (mHealth), que utilizan sensores de teléfonos inteligentes y portátiles, ofrecen soluciones innovadoras para diagnósticos objetivos y escalables. **Objetivos**: esta revisión examina la eficacia de los sensores de dispositivos móviles para diagnosticar la depresión, identificar bioseñales relevantes y explorar métodos de diagnóstico. **Metodología**: se realizó una búsqueda sistemática siguiendo las pautas PRISMA-DTA en MEDLINE/PubMed y Embase, centrándose en la precisión diagnóstica de los sensores mHealth validados contra estándares de oro como DSM-5 o PHQ-9. **Resultados**: once estudios mostraron que los acelerómetros y los monitores de frecuencia cardíaca son clave para detectar movimiento, actividad y patrones fisiológicos vinculados a la depresión. Los algoritmos de aprendizaje automático, especialmente los bosques aleatorios, lograron una alta precisión diagnóstica. **Conclusión**: las tecnologías mHealth son prometedoras para el diagnóstico de la depresión, pero se necesitan mejoras en la consistencia metodológica, el tamaño de la muestra y la validación externa para un uso clínico más amplio.

Palabras clave: revisión sistemática; trastorno depresivo; sensores remotos; mHealth

INTRODUCTION

Depression, a common mental health disorder worldwide, is characterized by symptoms such as sadness, loss of interest, fatigue, disturbed sleep or appetite, and poor concentration (Lim et al., 2018). The World Health Organization estimates that 5% of the global adult population suffers from depression, with 75% of individuals in low- and middle-income countries lacking proper treatment (Institute for Health Metrics and Evaluation, n.d.).

Malgaroli, Calderon & Bonanno (2021) state that the diagnostic criteria for Major Depressive Disorder (MDD) in the DSM-5 include nine core symptoms: (1) depressed mood, (2) loss of interest or pleasure, (3) significant changes in appetite or weight, (4) insomnia or hypersomnia, (5) psychomotor agitation or retardation, (6) fatigue or loss of energy, (7) feelings of worthlessness or excessive guilt, (8) difficulty concentrating or indecisiveness, and (9) recurrent thoughts of death or suicidal ideation. A diagnosis requires the presence of at least five of these symptoms, which allows for numerous possible symptom combinations and highlights the heterogeneity of the disorder.

Currently, the diagnosis of depression relies on clinician-administered assessments such as the Hamilton Depression Rating Scale (Hamilton, 1960) and the Beck Depression Inventory (Beck et al., 1996), both of which are based on self-reports and symptom evaluation. To address the subjectivity of these methods, researchers have explored alternative approaches that involve monitoring biological and physiological signals to enable a more objective diagnosis (Netto, 2024).

With the advancement of embedded technologies in mobile devices, such as smartphones and smartwatches, passive sensing has emerged as a promising solution for monitoring depression. These devices, equipped with various sensors, can collect a wide range of data, including step count, heart rate, sleep patterns, movement, and location, offering a more comprehensive and continuous method of assessing mental health (De Angel et al., 2022). Passive sensing, as defined by Winkler et al. (2022), refers to the non-invasive collection of behavioral data through smartphones or wearable devices.

This approach aligns with global trends in telemedicine and biomedical engineering, particularly within the scope of mobile health (mHealth). According to Netto (2020), mHealth refers to the use of mobile and wireless technologies, such as smartphones, smartwatches, remote patient monitoring devices, personal digital assistants, and mobile software applications, to support health-related objectives. As a subset of eHealth, mHealth is part of a broader effort to integrate information and communication technologies (ICT) into healthcare, promoting more accessible, continuous, and data-driven health monitoring and interventions (Netto & Petraroli, 2020).

Several literature reviews have explored the use of mobile devices for psychiatric evaluations. Cornet and Holden (2018) analyzed studies on the use of smartphone sensors for assessing health and wellbeing. Seppälä et al. (2019) reviewed studies linking sensors with psychiatric disorders, while De Angel et al. (2022) focused specifically on using passive data from mobile devices to monitor depression. In addition, Zarate et al. (2022) examined

various methods for digital data collection, emphasizing digital phenotyping (DP) for depression evaluation. Other authors (Highland & Zhou, 2022) explored the application of sensors, signal processing techniques, and algorithms in detecting depression and bipolar disorder. Despite their shared focus on using technology for mental health monitoring, each review offers unique perspectives on the topic.

The aim of this systematic review is to answer the research question: should mobile device sensors be employed for diagnosing clinical depression in the general population? The review focuses on studies that utilize mobile sensors for depression diagnosis, identifying biosignals relevant to this condition, and examining the methods used to collect these signals. To the best of our knowledge, this is the first systematic review dedicated to evaluating depression diagnosis through mobile device sensors.

METHODOLOGICAL REVIEW PROCEDURES

Study design

This study adopts systematic review design, grounded in the Cochrane methodology (Macaskill et al., 2010), which offers a rigorous and reproducible framework for synthesizing evidence from primary studies. The rationale for conducting a systematic review lies in its capacity to comprehensively identify, assess, and synthesize relevant studies from the global literature to answer a specific research question, in this case, to determine which mobile device sensors can detect depression through biosignals.

A systematic review allows researchers to aggregate data across diverse studies, considering variations in clinical trials, sample populations, and methodologies. This approach provides a structured synthesis of existing evidence, increasing the robustness and generalizability of findings. By following transparent and replicable procedures, systematic reviews contribute to evidence-based practice and guide future research and technology development.

To ensure transparent reporting of diagnostic accuracy studies, this review adheres to the PRISMA-DTA guidelines (Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies) (McInnes et al., 2018). Furthermore, to assess the reliability and strength of the gathered evidence, tools such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) may be employed, allowing for classification of evidence quality as high, moderate, low, or very low (Dijkers, 2013; Galvão & Pereira, 2015).

Ultimately, by systematically gathering and evaluating current literature on the use of mobile sensors for depression detection, this review aims to uncover emerging trends, assess diagnostic capabilities, and outline opportunities for future research in this interdisciplinary field.

Databases and search strategy

To identify primary studies for this systematic review, the search encompassed the following databases: MEDLINE/PubMed and Embase. These databases were selected due to their broad coverage of biomedical literature and their recognized relevance for retrieving studies in the field of health technologies, clinical research, and diagnostic methods.

In the pursuit of diagnostic test studies for depression involving sensors, the search strategy was structured around three key aspects to formulate the search string: type of study, target condition, and sensors. The search sequences described below were performed separately in each database (Table 1).

Table 1

Three key aspects to formulate the search string

Database	Query
MEDLINE / Pubmed	(sensitiv* OR specifici* OR "Sensitivity and Specificity"[Mesh] OR (predictive AND value*) OR "Predictive Value of Tests"[Mesh] OR accuracy* OR "False Negative*" OR "False Positive*") AND ("major depressive disorder" OR "major depression" OR "major depressive" OR "unipolar depression" OR "depressive disorder" OR "depression disorder" OR ("Mood disorders" and depression) OR ("Affective disorders" and depression) OR "bipolar depression") AND ("Biosensing Techniques" [Mesh], "Biosensing Technique" OR "Technique, Biosensing" OR "Techniques, Biosensing" OR "Biosensing Technics" OR "Biosensing Technic" OR "Technic, Biosensing" OR "Technics, Biosensing" OR Biosensors OR Biosensor OR "Electrodes, Enzyme" OR "Electrode, Enzyme" OR "Enzyme Electrode" OR "Enzyme Electrodes" OR Bioprobes OR Bioprobe OR Biosensing OR "Internet of Things" [Mesh] OR IoT OR "Wearable Electronic Devices" [Mesh] OR "Device, Wearable Electronic" OR "Devices, Wearable Electronic" OR "Electronic Device, Wearable" OR "Electronic Devices, Wearable" OR "Wearable Electronic Device" OR "Wearable Technology" OR "Technologies, Wearable" OR "Technology, Wearable" OR "Wearable Technologies" OR "Wearable Devices" OR "Device, Wearable" OR "Devices, Wearable" OR "Wearable Device" OR "Electronic Skin" OR "Skin, Electronic" OR "smartband" OR "fitness tracker" OR "smart watch" OR "smartphone" OR "Automatic exercise detection" OR "Automatic sleep monitoring" OR "Connected GPS" OR "Heart rate monitor" OR "Heart rate variability for stress scores" OR "Rep counting for gym exercises" OR "Sleep monitoring with Sleep Stages" OR "Sleep tracking" OR "SpO2 sensor / Oximetry" OR "Step tracking" OR "Steps and activity tracking" OR "Swim tracking" OR "VO2 Max" OR "Fatigue Monitoring" OR "Blood pressure" OR "Mileage recording Calories")
Embase	(sensitiv* OR specifici* OR "Sensitivity and Specificity" OR (predictive AND value*) OR "Predictive Value of Tests" OR accuracy* OR "False Negative*" OR "False Positive*") AND ("major depressive disorder" OR "major depression" OR "major depressive" OR "unipolar depression" OR "depressive disorder" OR "depression disorder" OR ("Mood disorders" and depression) OR ("Affective disorders" and depression) OR "bipolar depression") AND ("Biosensing Techniques", "Biosensing Technique" OR "Technique, Biosensing" OR "Techniques, Biosensing" OR "Biosensing Technics" OR "Biosensing Technic" OR "Technic, Biosensing" OR "Technics, Biosensing" OR Biosensors OR Biosensor OR "Electrodes, Enzyme" OR "Electrode, Enzyme" OR "Enzyme Electrode" OR "Enzyme Electrodes" OR Bioprobes OR Bioprobe OR Biosensing OR "Internet of Things" OR IoT OR "Wearable Electronic Devices" OR "Device, Wearable Electronic" OR "Devices, Wearable Electronic" OR "Electronic Device, Wearable" OR "Electronic Devices, Wearable" OR "Wearable Electronic Device" OR "Wearable Technology" OR "Technologies, Wearable" OR "Technology, Wearable" OR "Wearable Technologies" OR "Wearable Devices" OR "Device, Wearable" OR "Devices, Wearable" OR "Wearable Device" OR "Electronic Skin" OR "Skin, Electronic" OR "smartband" OR "fitness tracker" OR "smart watch" OR "smartphone" OR "Automatic exercise detection" OR "Automatic sleep monitoring" OR "Connected GPS" OR "Heart rate monitor" OR "Heart rate variability for stress scores" OR "Rep counting for gym exercises" OR "Sleep monitoring with Sleep Stages" OR "Sleep tracking" OR "SpO2 sensor / Oximetry" OR "Step tracking" OR "Steps and activity tracking" OR "Swim tracking" OR "VO2 Max" OR "Fatigue Monitoring" OR "Blood pressure" OR "Mileage recording Calories")

Eligibility criteria

To define the scope of the systematic review, only primary diagnostic accuracy studies were included, all primary studies of diagnostic testing used in the diagnosis of depression (target condition) through sensors (test under evaluation). Eligible studies had to include a formal diagnosis of depression, based either on internationally recognized diagnostic criteria such as the DSM-V (American Psychiatric Association, 2013) and/or the use of validated gold-standard instruments, such as the Beck Depression Inventory (BDI), the Hamilton Depression Rating Scale (HAM-D), or the Patient Health Questionnaire (PHQ-9) (Kroenke & Spitzer, 2002; Snaith & Taylor, 1985). This

requirement ensured that depression was not inferred solely from isolated symptoms, which could overlap with conditions like anxiety or stress, but was instead grounded in robust and widely accepted diagnostic frameworks.

To guarantee diagnostic reliability, studies were only included if they explicitly reported the reference standard used for diagnosing depression. This allowed for proper comparison between the sensor-based evaluation methods and a consistent, clinically validated benchmark.

In terms of technological scope, the review considered studies involving mHealth solutions, including wearable devices, mobile phones, and handheld technologies. Conversely, studies in which the diagnosis of depression relied solely on clinical or hospital-grade equipment (i.e., outside the mHealth domain) were excluded.

Additionally, studies were excluded if they met any of the following criteria:

- They focused exclusively on stress and/or anxiety without a confirmed diagnosis of depression.
- They involved only treatment or intervention strategies for individuals already diagnosed with depression, rather than aiming to evaluate diagnostic performance.
- They failed to present or describe the diagnostic method or reference standard used.

Screening and studies selection

The initial screening consisted of evaluating the titles and abstracts of all studies retrieved from the databases. At this stage, two researchers independently labeled the studies for inclusion or exclusion according to the previously stated eligibility criteria. After independent assessment, their decisions were compared, and studies not excluded by both reviewers proceeded to the next stage.

Subsequently, the full-text versions of the remaining studies were retrieved for a more detailed selection stage. Both researchers independently read the full articles to determine, based on the eligibility criteria, which studies would be included in the review. After this evaluation, the reviewers' decisions were compared.

In cases of disagreement during either the abstract or full-text screening stages, a consensus process was applied. When no agreement could be reached between the two reviewers, two senior researchers with PhDs were consulted to provide expert judgment and support the final decision regarding the study's eligibility. Agreement statistics between the initial reviewers were calculated using the Kappa Agreement Coefficient (k), with the interpretation categories proposed by Altman (1991): Poor (< 0.2), Fair (≥ 0.2 and < 0.4), Moderate (≥ 0.4 and < 0.6), Good (≥ 0.6 and < 0.8), and very good (≥ 0.8).

After running the search in the databases, it was possible to retrieve a set of 156 and 1638 results in the PubMed and Embase databases respectively, which included a total of 1794 search results. However, 41 studies were removed since they were duplicates. Thus, 1753 articles remained to be evaluated in the screening stage. From the analysis of the titles and abstracts of the search results, 1441 studies were excluded. The reasons included: Study not related to the diagnosis of depression; No use of sensors/mHealth; Diagnosis through genetic techniques; Exclusively intervention/treatment

studies; Literature review articles; and Systematic reviews and/or meta-analysis.

Data collection

The data from the included studies were extracted, including information on the publication year, country and continent of the research, study design, and demographic data such as age and sex. Additionally, the number of individuals with depression and sensor hits were recorded, and the use of the DSM as a confirmatory gold standard was verified. Two reviewers independently extracted the data, and any disagreements were resolved by consensus. Each evaluator calculated the pre-test probability (prevalence of depression), sensitivity, specificity, and associated measures (Macaskill et al., 2010) for diagnosing depression using sensors. Out of the 296 studies initially retrieved for reading, 284 were excluded for not meeting the inclusion criteria, leaving 11 studies that were included in the review. The flowchart of study selection, based on the PRISMA guidelines (Page et al., 2021), illustrates the study selection process.

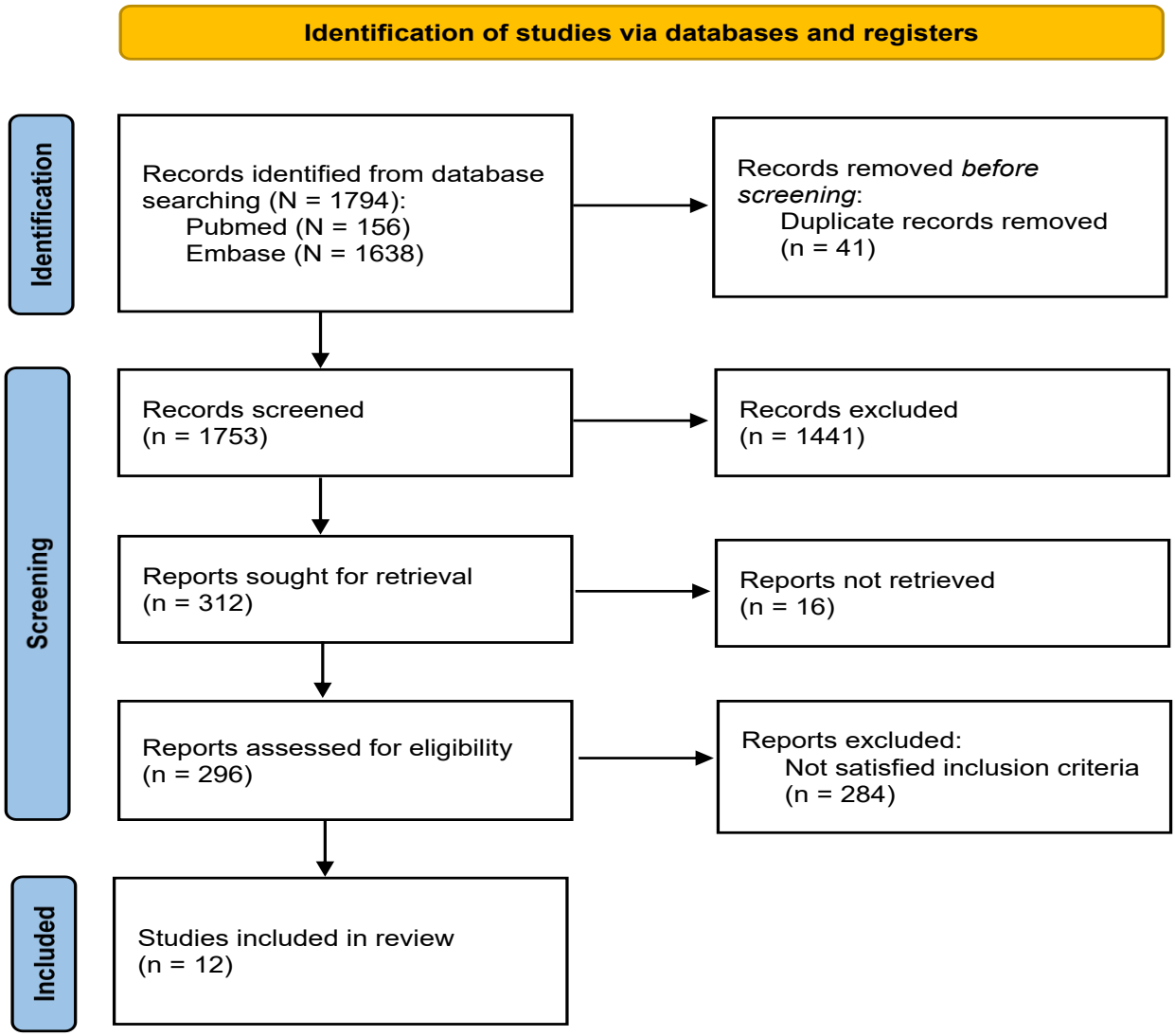

Figure 1

Flowchart illustrates the selection steps according to PRISMA

Evaluation of methodological quality and data analysis

The evaluation of the methodological quality of the studies was conducted using the Quality Assessment of Diagnostic Accuracy Studies - 2 tools, which focus on four key domains: patient selection, evaluation test, standard reference, and flow/time. Each domain is assessed for risk of bias, with the first three also considering concerns regarding study applicability (Schueler et al., 2012; Whiting et al., 2011). Additionally, PROBAST (Prediction model Risk of Bias ASsessment Tool) was employed to evaluate studies based on four domains: patient selection, predictors, outcome, and analysis (Wolff et al., 2019). The risk of bias was determined for each study by categorizing it as low, high, or unclear based on the evaluations of these domains. The GRADE system, supported by the GradePro GDT tool, was used to assess the quality of evidence, considering factors such as study design, bias risk, inconsistency, indirectness, imprecision, and publication bias. A 2x2 contingency table was constructed for each study to classify gold standard and sensor results, and the Diagnostic Odds Ratio (DOR) was calculated to measure diagnostic accuracy.

RESULTS

Sample information

According to the eligibility criteria, 11 studies were included. The sample was made up of adult individuals, between 18 and 69 years old, with studies executed in North and South America, Europe and mainly Asia. The duration of the studies ranged from 1 week (Jacobson et al., 2019) to over 2 years (Cho et al., 2019). Over the analyzed studies, the set of patients diagnosed with depression and the set control groups varied, from imbalanced in some studies to perfectly balanced in others. The PHQ-9 was the most used gold standard, being adopted in at least 50% of the selected studies. Considering depression prevalence, the values vary from 20.5% to 100%, while the lower absolute number of patients as 15 (Jacobson et al., 2019) and the max 1375 (Zanella-Calzada et al., 2019). A common characteristic between all the studies samples included adults.

Used methods information and machine learning methods

This review highlights the use of smartphones as the primary device for diagnosing depression, appearing in 9 out of 11 selected studies. This is expected, given the range of sensors in smartphones, such as accelerometers, GPS, light sensors, microphones, and cameras, along with their widespread accessibility (Chao, 2018). In addition to smartphones, smartwatches were also utilized to collect biosignals for depression diagnosis. Regarding machine learning techniques, five main algorithms were employed: random forest (RF), k-nearest neighbors (KNN), artificial neural networks (ANN), extreme gradient boosting (XGBoost), and support vector machine (SVM). As shown in Table 2, random forest was the most frequently used, followed by SVM and XGBoost, with KNN and neural networks each appearing in two studies. The choice of algorithms may depend on the type and volume of data available for training, as the performance of machine learning techniques is often influenced by the quality and quantity of the training data (Zhou et al., 2020).

Sensors

The most used sensor in the selected studies was the accelerometer (67%), followed by smartphone usage patterns (50%). Other sensors included global positioning system (GPS), light sensor, heart-rate monitor (HRM), and the smartphone's touchscreen. These sensors are typically associated with capturing signals related to movement and location patterns (using accelerometer and GPS), social interaction (based on smartphone usage patterns such as calls, messages, and social media usage), and sleep patterns (via light sensors and heart activity). The extracted features were primarily derived from statistical calculations applied to the sensor data, including maximum and minimum values, mean, standard deviation, kurtosis, and skewness. Table 3 summarizes the methodologies used in the selected studies.

Table 2

Sample information summary extracted from studies

Study	Sample	Patients' origin	N_sample / N_depression	Depression prevalence	Gold standard	Study execution year (Duration)
Jacobson et al., 2019	Outpatients taking serotonin reuptake inhibitor or tricyclic antidepressants	Porto Alegre Clinical Hospital, Brazil	15 / 15	100%	HAM-D	N/A (1 week)
Cho et al., 2019	27 women and 28 men diagnosed with a major mood disorder	Korea University Anam Hospital, Korea	55 / 19	34,5%	DSM-5	Mar/15 – Dec/17 (2 years)
Sarda et al., 2019	29 men and 17 women with diabetes	Aurangabad, India	46 / 30	65,2%	PHQ-9	2016 (N/A)
Narziev et al., 2020	College students	Inha University, Korea	21 / 16	76,2%	PQH-9 / BD-II	N/A
Masud et al., 2020	19 men and 14 women (18+ years old)	Dhaka, Bangladesh	33 / 9	27,3%	PHQ-9	Apr – Jun/18 (11 weeks)
Dogrucu et al., 2020	N/A	N/A	335 / N/A	N/A	PHQ-9	N/A (2 weeks)
Zebin et al., 2019	Adults between 40 and 69 years old	UK Biobank, United Kingdom	80 / 39	48,8%	N/A	2013-2015
Zanella-Calzada et al., 2019	N/A	Depresjon database	1375 / 682	49,6%	MADRS	N/A
Mastoras et al., 2019	Adults (up to 40 years old) without undergoing any medication treatment	Khalifa University, United Arab Emirates	25 / 11	44,0%	PHQ-9	Nov/18 – Mar/19 (124 days)
Faurholt-Jepsen et al., 2019	Patients' diagnosis according to ICD-10 non pregnant	Psychiatric Centre Copenhagen, Denmark	66 / 29	43,9%	HDRS-17 / YMRS	Oct/13 – Dec/14 (12 weeks)
Ware et al., 2020	College students	United States	88 / 18	20,5%	PHQ-9	N/A (8 months)

Table 3

Methods information summary extracted from studies

Study	Machine learning algorithms	Signal features extraction	Features selection	Used signal	Sensors	Device
Jacobson et al., 2019	Extreme gradient boosting	Square root, square and log transformations	Leave-one-out, permutation test	Oscillations and peak values of movement and luminosity	Accelerometer, Luminosity	Actiwatch
Cho et al., 2019	Random forest	N/A	N/A	Light exposure, steps, sleep and heart rate	Accelerometer, HRM, Luminosity	Smartphone / Smartwatch (Fitbit Charge)
Sarda et al., 2019	Extreme gradient boosting	N/A	Bagging and boosting trees	Activity rate, smartphone screen time, call patterns	Accelerometer, GPS, Luminosity, Smartphone usage	Smartphone
Narziev et al., 2020	SVM, random forest	Mean, stdev, max and min values, energy, kurtosis, skeweness, root-mean-square	N/A	Movement patterns, heart rate and calls	Accelerometer, HRM, Smartphone usage, Luminosity	Smartphone / Smartwatch
Masud et al., 2020	SVM, KNN, ANN	Distance variance, normalized entropy, quotidian movement	Wrapper, root-mean stdev	Location, movement and step patterns	Accelerometer, GPS	Smartphone
Dogrucu et al., 2020	KNN, SVM, random forest	Call frequency, audio features, distance locations	N/A	Location, speech and typing patterns	GPS, Smartphone usage	Smartphone
Zebin et al., 2019	DNN, random forest	Daily, weekly and overall acceleration, no wearing time.	Ensemble DNN	Type, intensity and duration of physical activity	Accelerometer	Accelerometer (Activity AX3)
Zanella-Calzada et al., 2019	Random forest	Mean, stdev, kurtosis, skeweness, coefficient of variation	N/A	Movement	Accelerometer	Actiwatch (AW4)
Mastoras et al., 2019	SVM, random forest, gradient boosting	Typing patterns, mean, stdev, kurtosis, skeweness	Select k-best with ANOVA	Typing metadata	Smartphone touchscreen	Smartphone
Faurholt-Jepsen et al., 2019	Extreme gradient boosting	N/A	Gradient boosting	Call, SMS and screen time patterns	Smartphone usage	Smartphone
Ware et al., 2020	N/A	Location clusters, activity data	N/A	Location variance and time spend in moving	GPS, Smartphone usage	Smartphone

Quadas analysis

Systematic reviews of diagnostic accuracy studies often exhibit significant heterogeneity due to differences in study design and execution. As noted by Whiting et al. (2011), the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool was introduced in 2003 and has since been widely adopted and recommended by organizations such as the Cochrane Collaboration and the Agency for Healthcare Research and Quality. The original version comprised 14 items aimed at assessing risk of bias, applicability, and reporting quality, with each item rated as “yes,” “no,” or “unclear.” However, users reported difficulties in interpreting some questions and noted overlapping between certain criteria. In response, the QUADAS-2 tool was developed as an enhanced version, incorporating user feedback and new evidence to provide a more structured and precise evaluation of bias and applicability in diagnostic accuracy studies.

Regarding patient selection, the sample of study participants was randomly or consecutively obtained, assuming that this was also the case in studies in which this information was not made explicit. A fact that may have introduced bias in relation to the sample was in the study of Faurholt-Jepsen et al. (2019), which not only was a case-control study, but also the control group consisted of patients from the hospital's blood bank, being a group with more favorable health conditions that may influence the comparison of the affective group. In addition to Faurholt-Jepsen et al. (2019), the study of Hess et al. (2015) also indicated that it was a case-control study, which can generate biased results. Studies of Jacobson et al., (2019), Narziev et al. (2020), Zanella-Calzada et al. (2019), Mastoras et al. (2019) and Ware et al., (2020) were not clear about one or more questions regarding patient selection.

Considering the index tests, in all studies, except for Dogrucu et al. (2020) and Mastoras et al. (2019), it was unclear whether the index test results were interpreted with knowledge of the reference standard results or if a pre-specified threshold was used.

Regarding the reference standard, except for study of Zebin et al. (2019), in which the use of a recognized gold standard was not specified, all studies adopted a reference standard capable of making the diagnosis of depression. Furthermore, considering that the data from the sensors collected in each study compose datasets, which are later used to train supervised machine learning models, that is, which, based on the input data, seek to correspond to a target (condition depression or not, or even the level of depression), the reference standard is obtained before the test-index, since it is a necessary data to carry out the training.

Another detail that possibly requires considering is the way in which the benchmark tests were performed between the different studies. Citing as an example the PHQ-9, which is a questionnaire that the patient answers, which according to the answers given is generated a score that allows the diagnosis of depression, in some of the studies this was performed by the patients in clinical evaluations, while in others, the form was only filled in electronically, using a smartphone application (Mastoras et al., 2019), e-mail/SMS (Masud et al., 2020), online survey (Dogrucu et al., 2020) or even by telephone (Narziev et al., 2020), for example.

Considering that the machine learning models used to classify depression are generated based on data collected on an ongoing basis and with the reference standard being obtained at regular periods in most studies, it makes no sense to evaluate the time bias. introduced between the test and the reference. Regarding the adoption of the same standard of reference for all participants, the study of Zebin et al. (2019) used as reference the self-declarations of patients in having or not having depression, which may have introduced bias since it is not possible to guarantee that all participants adopted the same reference standard, or even if it was a recognized standard. As for the studies of Sarda, et al. (2019), Zanella-Calzada et al. (2019) and Mastoras et al. (2019), there is a possibility that they have introduced bias in the analysis of the results, due to the fact that not all of the selected participants were included in the analysis. presents a summary of the QUADAS-2, where green cells indicate “low risk,” yellow cells indicate “unclear risk,” and red cells indicate “high risk.”.

DISCUSSION

Risk of bias and applicability

As the PROBAST analyzes specific issues in studies involving predictive models, it has also been applied for bias and applicability assessment. This serves as a complement to the analysis conducted through Quadas-2.

In the first domain concerning participants, seven studies exhibited a high risk of bias, primarily attributed to the study type. Among them, four were cross-sectional studies (Mastoras et al., 2019; Sarda et al., 2019; Ware et al., 2020; Zebin et al., 2019), two were case-control studies (Faurholt-Jepsen et al., 2019; Zanella-Calzada et al., 2019), and one was an experimental study (Narziev et al., 2020). In three studies (Dogrucu et al., 2020; Jacobson et al., 2019; Masud et al., 2020), the risk of bias was deemed unclear due to insufficient evidence regarding the data sources used or the inclusion and exclusion criteria.

With respect to domains 2 (predictors) and 3 (outcomes), all studies exhibited a low risk of bias. In domain 4 (analysis), four studies (Jacobson et al., 2019; Mastoras et al., 2019; Masud et al., 2020; Ware et al., 2020) were deemed to have a high risk of bias due to their limited sample size and inadequate assessment of performance measures. Considering applicability, the selected studies were deemed to have a low risk across all domains. This suggests that they provide pertinent and compatible data for the research question of this systematic review. As none of the studies conducted external validation, all of them were regarded with an overall high risk of bias, including the study by Cho et al. (2019), which showed a low risk of bias in all domains.

Table 4

Summary of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) - 2

Study	Patient Selection			Index Test		Reference Standard		Flow and Timing			
	P1	P2	P3	P1	P2	P1	P2	P1	P2	P3	P4
Jacobson et al., 2019	🙂	😐	😐	😐	😐	🙂	🙂	😐	🙂	🙂	🙂
Cho et al., 2019	🙂	🙂	🙂	😐	🙂	🙂	🙂	🙂	🙂	🙂	🙂
Sarda et al., 2019	🙂	🙂	🙂	😐	🙂	🙂	🙂	😐	🙂	🙂	☹
Narziev et al., 2020	🙂	😐	😐	🙂	😐	🙂	🙂	😐	🙂	🙂	🙂
Masud et al., 2020	🙂	🙂	🙂	😐	😐	🙂	🙂	😐	🙂	🙂	🙂
Dogrucu et al., 2020	🙂	🙂	🙂	🙂	🙂	🙂	🙂	😐	🙂	🙂	🙂
Zebin et al., 2019	🙂	☹	🙂	😐	😐	☹	☹	☹	☹	☹	🙂
Zanella-Calzada et al., 2019	😐	😐	🙂	🙂	😐	🙂	🙂	😐	🙂	🙂	☹
Mastoras et al., 2019	🙂	😐	🙂	🙂	🙂	🙂	🙂	😐	🙂	🙂	☹
Faurholt-Jepsen et al., 2019	🙂	☹	🙂	🙂	😐	🙂	🙂	😐	🙂	🙂	🙂
Ware et al., 2020	🙂	😐	😐	😐	😐	🙂	🙂	😐	🙂	🙂	🙂

Depression test metrics and evidence quality assessment

Considering the values and confidence intervals for both sensitivity and specificity, the 3 studies that showed the best results were (Narziev et al., 2020), (Zebin et al., 2019) and (Zanella-Calzada et al., 2019). Each of these studies were based on movement patterns acquired by an accelerometer sensor embedded on a smartphone, smartwatch or both. Another common feature between these studies was the use of a random forest algorithm for training the machine learning model used for prediction. On the other hand, (Faurholt-Jepsen et al., 2019) presented discrepant results compared to the rest of the evaluated studies, since its specificity confidence interval that does not overlap with the other ones, indicating a smaller ability of the model to predict a negative result to the cases which the individuals patients do not have depression.

In line with the GRADE approach, the certainty of evidence from the primary studies for sensitivity and specificity was assessed as moderate. This implies that further research is likely to substantially impact on our confidence in the estimated effect and could potentially modify the overall estimate. The primary reasons for downgrading by one level were linked to the Risk of Bias identified in the PROBAST analysis, particularly concerning the study's design and the absence of external validation. No concerns were raised about inconsistency (statistical heterogeneity), indirectness (clinical heterogeneity), imprecision (presentation of results), or publication bias.

Simulations were conducted to evaluate the impact per 1000 patients tested, considering prevalences of 1% (reflecting a low prevalence), 5% (corresponding to the estimated prevalence), and 25% (representing a high prevalence scenario). With a 95% confidence interval and the estimated prevalence (5%), if 1000

individuals were tested, 45 individuals exhibiting depression symptoms would be correctly referred for treatment, while 5 would be missed. Among the remaining 950 individuals without depression symptoms, 826 would be correctly identified, and 124 would be erroneously identified. The evidence suggests that mobile devices may be effective in detecting signals of depression.

CONCLUSION

This systematic review analyzes studies on diagnosing depression using sensors in mobile devices, focusing on mHealth technologies. It highlights the potential applications of these devices, summarizing relevant algorithms, biosignals, sensors, sample data, and other key factors. Despite limitations such as study types and biases, addressed through QUADAS and PROBAST analysis, the review follows Cochrane guidelines and aims to contribute to clinical research. The findings suggest that motion sensors, especially accelerometers, along with smartphone usage patterns, show promise for developing diagnostic applications.

Data from everyday sensors in phones and smartwatches, such as step counter, heart-rate monitor, GPS, and device usage logs, can act like a quick "digital health check" for depression, since clinics can be warned when someone's patterns look risky, patients can see easy charts that help them notice mood changes, and accessibility options, such as adjusting fonts to larger sizes, using voice commands, and caregiver-sharing options, keep the tools friendly and useful for elderly people. These sensor signals may also give teachers ready examples for lessons in medicine, nursing, engineering, and data Science, and allow researchers to perform larger studies in more locations accross countries, create well-annotated datasets, track users over long periods, and measure real-world cost-benefit in clinical care. While further studies and clinical trials are needed to improve reliability, this review serves as a foundational resource for future research on depression diagnosis through mHealth devices.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™* (5th ed.). American Psychiatric Publishing Inc. https://psycnet.apa.org/doi/10.1176/appi.books.9780890425596

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory.* Harcourt Brace Jovanovich.

Chao, C. N. J. (2018). Traditional vs internet vs mobile: which is more effective way to reach potential customers. *Journal of Business Administration Research, 7*(2). https://doi.org/10.5430/jbar.v7n2p1

Cho, C. H., Lee, T., Kim, M. G, In, H. P, Kim, L. & Lee, H. J. (2019). Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *Journal of Medical Internet Research, 21*(4). https://doi.org/10.2196/11029

Cornet, V. P. & Holden, R. J. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics*, 77, 120-132. https://doi.org/10.1016/j.jbi.2017.12.008

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., & Mohr, D. C. J. N. (2022). Digital health tools for the passive monitoring of depression: a systematic review of methods. *npj Digital Medicine, 5*(3). https://doi.org/10.1038/s41746-021-00548-8

Dijkers, M. (2013). Introducing GRADE: a systematic approach to rating evidence in systematic reviews and

to guideline development. *KT Update*, *1*(5). https://ktdrr.org/products/update/v1n5/dijkers_grade_ktupdatev1n5.html

Dogrucu, A., Perucic, A., Isaro, A., Ball, D., Toto, E., Rundensteiner, E. A., Agu, E., Davis-Martin, R., & Boudreaux, E. (2020). Moodable: on feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health, 17*, 100118. https://doi.org/10.1016/j.smhl.2020.100118

Faurholt-Jepsen, M., Busk, J., Þórarinsdóttir, H., Frost, M., Bardram, J. E., Vinberg, M., & Kessing, L. V. (2019). Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Australian and New Zealand Journal of Psychiatry, 53*(2). https://doi.org/10.1177/0004867418808900

Galvão, T. F., & Pereira, M. G. (2015). Avaliação da qualidade da evidência de revisões sistemáticas. *Epidemiologia e Serviços de Saúde*, *24*(1), 173-175. https://doi.org/10.5123/S1679-49742015000100019

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56-62. https://doi.org/10.1136/jnnp.23.1.56

Hess, S., Fischel, T., Dolfin, D., Hermesh, H., Horowitz, I., Sela, K., Liberman, A., Nevo, U., & Weizman, A. (2015). The use of smartphone app for early detection of manic or depressive episodes in affective patients. *European Neuropsychopharmacology*, *25*(Suplemento 2), S392. https://doi.org/10.1016/S0924-977X(15)30514-9

Highland, D., & Zhou, G. (2022). A review of detection techniques for depression and bipolar disorder. *Smart Health*, *24*, 100282. https://doi.org/10.1016/j.smhl.2022.100282

Institute for Health Metrics and Evaluation. (n.d.). *Global health data exchange (GHDx)*. http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2019). Using digital phenotyping to accurately detect depression severity. *Journal of Nervous and Mental Disease*, *207*(10), 893–896. https://doi.org/10.1097/nmd.0000000000001042

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509-515. https://doi.org/10.3928/0048-5713-20020901-06

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. *Scientific Reports*, *8*(1), 2861. https://doi.org/10.1038/s41598-018-21243-x

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., & Takwoingi, Y. (2010). Chapter 10: Analysing and presenting results. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* version 1.0. The Cochrane Collaboration. https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/uploads/Chapter%2010%20-%20Version%201.0.pdf

Malgaroli, M., Calderon, A., & Bonanno, G. A. (2021). Networks of major depressive disorder: a systematic review. *Clinical Psychology Review*, *85*, 102000. https://doi.org/10.1016/j.cpr.2021.102000

Mastoras, R. E., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Kassie, S., Alsaadi, T., Khandoker, A., & Hadjileontiadis, L. J. (2019). Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific Reports*, 9, 13414. https://doi.org/10.1038/s41598-019-50002-9

Masud, M. T., Mamun, M. A., Thapa, K., Lee, D. H., Griffiths, M. D., & Yang, S. H. (2020). Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. *Journal of Biomedical Informatics*, *103*, 103371. https://doi.org/10.1016/j.jbi.2019.103371

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., & The PRISMA-DTA Group. (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. *JAMA*, *319*(4), 388–396. https://doi.org/10.1001/jama.2017.19163

Narziev, N., Goh, H., Toshnazarov, K., Lee, S. A., Chung, K. M., & Noh, Y. (2020). STDD: Short-Term Depression Detection with Passive Sensing. *Sensors*, *20*(5), 1396. https://doi.org/10.3390/s20051396

Netto, A. V. (2020). Application of blended care as a mechanism of action in the construction of digital therapeutics. *einstein (São Paulo)*, *18*. https://10.31744/einstein_journal/2020MD5640

Netto, A. V., & Petraroli, A. G. (2020). Modelagem de um sistema para o telemonitoramento de idosos com

condição crônica baseado em biotelemetria, *Journal of Health Informatics*, *12*(1). https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/691/379

Netto, A. V. (2024). Proposal for the application of a blended care platform for patients with depressive symptoms. *Revista Saúde Multidisciplinar*, *16*(1). https://doi.org/10.53740/rsm.v16i1.742

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. J. S. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(89). https://doi.org/10.1186/s13643-021-01626-4

Sarda, A., Munuswamy, S., Sarda, S., & Subramanian, V. (2019). Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study. *JMIR Mhealth Uhealth*, *7*(1), e11041. https://doi.org/10.2196/11041

Schueler, S., Schuetz, G. M., & Dewey, M. (2012). The revised QUADAS-2 tool [letter]. *Annals of Internal Medicine*, *156*(4). https://doi.org/10.7326/0003-4819-156-4-201202210-00018

Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., Feldman, Y., Grasa, E., Corripio, I., & Berdun, J. J. J. (2019). Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: systematic review. *JMIR Mental Health*, *6*(2), e9819. https://doi.org/10.2196/mental.9819

Snaith, R., & Taylor, C. (1985). Rating scales for depression and anxiety: a current perspective. *British Journal of Clinical Pharmacology*, *19*(S1), 17S-20S. https://doi.org/10.1111/j.1365-2125.1985.tb02737.x

Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Bi, J., Kamath, J., Russel, A., Bamis, A., & Wang, B. (2020). Predicting depressive symptoms using smartphone data. *Smart Health*, *15*, 100093. https://doi.org/10.1016/j.smhl.2019.100093

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A., Bossuyt, P. M., & Medicine, Q. G. J. A. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8). https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. on behalf of the PROBAST group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1). https://doi.org/10.7326/M18-1376

Zanella-Calzada, L. A., Galván-Tejada, C. E., Chávez-Lamas, N. M., del Carmen Gracia-Cortés, M., Magallanes-Quintanar, R., Celaya-Padilla, J. M., Galván-Tejada, J. I., & Gamboa-Rosales, H. (2019). Feature extraction in motor activity signal: towards depression episodes detection in unipolar and bipolar patients. *Diagnostics*, *9*(1), 8. https://doi.org/10.3390/diagnostics9010008

Zarate, D., Stravopoulos, V., Ball, M., Collier, G., & Jacobson, N. (2022). Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry*, *22*, 421. https://doi.org/10.1186/s12888-022-04013-y

Zebin, T., Peek, N., & Casson, A. J. (2019). Physical activity-based classification of serious mental illness group participants in the UK Biobank using ensemble dense neural networks [Abstract]. *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, pp. 1251–1254. https://doi.org/10.1109/EMBC.2019.8857532

Zhou, Q., Lu, S., Wu, Y., & Wang, J. (2020). Property-oriented material design based on a data-driven machine learning technique. *The Journal of Physical Chemistry Letters*, *11*(10). https://doi.org/10.1021/acs.jpclett.0c00665