algorithms

MDPI

Review

Using Wearable Technology to Detect, Monitor, and Predict Major Depressive Disorder—A Scoping Review and Introductory Text for Clinical Professionals

Quinty Walschots [1], Milan Zarchev [2,3], Maurits Unkel [4] and Astrid Kamperman [2,3,*]

[1] Faculty of Medicine, Leiden University, P.O. Box 9500, 2300 RA Leiden, The Netherlands; quintywalschots@gmail.com
[2] Department of Psychiatry, Epidemiological and Social Psychiatric Research Institute (ESPRi), Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands; m.zarchev@erasmusmc.nl
[3] Department of Psychiatry, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands
[4] Department of Neuroscience, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands; m.unkel@erasmusmc.nl
* Correspondence: a.kamperman@erasmusmc.nl

Abstract: The rising popularity of wearable devices allows for extensive and unobtrusive collection of personal health data for extended periods of time. Recent studies have used machine learning to create predictive algorithms to assess symptoms of major depressive disorder (MDD) based on these data. This review evaluates the clinical relevance of these models. Studies were selected to represent the range of methodologies and applications of wearables for MDD algorithms, with a focus on wrist-worn devices. The reviewed studies demonstrated that wearable-based algorithms were able to predict symptoms of MDD with considerable accuracy. These models may be used in the clinic to complement the monitoring of treatments or to facilitate early intervention in high-risk populations. In a preventative context, they could prompt users to seek help for earlier intervention and better clinical outcomes. However, the lack of standardized methodologies and variation in which performance metrics are reported complicates direct comparisons between studies. Issues with reproducibility, overfitting, small sample sizes, and limited population demographics also limit the generalizability of findings. As such, wearable-based algorithms show considerable promise for predicting and monitoring MDD, but there is significant room for improvement before this promise can be fulfilled.

Keywords: depression; wearables; prediction; mental healthcare

Citation: Walschots, Q.; Zarchev, M.; Unkel, M.; Kamperman, A. Using Wearable Technology to Detect, Monitor, and Predict Major Depressive Disorder—A Scoping Review and Introductory Text for Clinical Professionals. *Algorithms* **2024**, *17*, 408. https://doi.org/10.3390/a17090408

Academic Editor: Edward Rolando Núñez-Valdez

Received: 9 August 2024
Revised: 30 August 2024
Accepted: 6 September 2024
Published: 12 September 2024

1. Introduction

In the past two decades, there has been increasing interest in wearable technology [1,2]. These wearables include any type of technology one can wear on their person, such as pedometers, headphones, and virtual reality headsets. The category of smartwatches is the most common, making up roughly 30% of all wearables [3]. The FitBit and Apple Watch are prominent examples. With smartwatches, it is possible to collect a wide variety of data points about the user's health, such as heart rate and physical activity. Their displays or companion apps make these data easily accessible for the user. There is a rising interest in monitoring personal fitness and mental health [4]. However, mobile mood-tracking apps are often used infrequently, and users tend to avoid entering negative moods [5]. Wearables allow for continuous collection of a large variety of detailed information and have previously been used to identify what emotions the user is experiencing [6,7]. Furthermore, wearables can collect these data while remaining unobtrusive, making them especially suitable for long-term measurements with minimal impact on the user. The use of wearables

in mental health care has been a topic of interest for some time. While existing reviews focus on exploring the various wearable sensors and features that can inform predictive models [8–11], this review focuses more on the practical applications and their feasibility. Specifically, this review aims to evaluate the role of wearables in a clinical setting, in particular, how they may be used to identify and monitor symptoms of major depressive disorder (MDD) in a way that is accessible to both researchers and clinical professionals.

1.1. Major Depressive Disorder

Major depressive disorder (MDD) is one of the most prevalent mood disorders worldwide, affecting an estimated 280 million people (3.8% of the population) as reported in 2023 [12]. In reality, these numbers may be even higher, as it is not uncommon for MDD to stay undiagnosed [13]. Furthermore, MDD is becoming more prevalent as its global burden has increased by more than 50% compared to 35 years ago [14]. It is characterized by persistent feelings of sadness and hopelessness, as well as a generally depressed mood and anhedonia. For detailed diagnostic criteria, refer to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [15]. Diagnosis and monitoring of MDD rely heavily on subjective and retrospective reports by the patient or observations made by others. Wearables offer a method to relate depressive symptoms to objective measures and monitor their development over time. Furthermore, they offer additional insights to the casual user about their own mental health, which could prompt them to seek professional help at earlier time points. This could help reduce the number of undiagnosed patients as well as facilitate earlier diagnosis or intervention, which is typically associated with better treatment outcomes and reduces the burden of the disease both emotionally and financially [16]. Recent studies have been looking into how wearable technology may be used for such a purpose.

1.2. Measurable Features of MDD

While there is no single or definitive cause for MDD, disruptions of several biological processes have been related to the disease. One prominent example is the circadian clock. Using external time cues, this system modulates certain processes in the body (most notably the sleep-wake cycle) to follow a roughly 24 h rhythm [17]. It has been widely reported that the circadian clock is disrupted in patients with depression, as reflected, among other things, by the prevalence of insomnia and disrupted sleep reported by this population [18,19]. Another major regulatory mechanism found to be disrupted in MDD is the hypothalamic pituitary adrenal (HPA) axis [20,21]. This system is crucial for stress adaptation and is normally activated in response to emotional or physical stressors [22]. The HPA axis is involved in the regulation of blood pressure, metabolism, and the fight–flight response, which in turn may increase heart rate, body temperature, and perspiration when activated [23–25].

Both the circadian clock and the HPA axis involve physiological processes that can be measured using sensors available in most wearables. Accelerometers can further measure physical activity, which can also be used to quantify dimensions of sleep. Existing research has shown that depressed patients generally spend more time sedentary compared to healthy controls [26], and people with lower activity levels are more likely to show signs of depression [27]. The level of physical activity has been found to increase as treatment progresses, making it a promising feature for monitoring depressive symptoms [28,29]. Furthermore, the activity of patients was found to be shifted towards later in the day, with a dampened diurnal amplitude [30]. Photoplethysmogram (PPG) sensors can measure heart rate. Measured heart rate, and specifically lowered heart rate variability (HRV), has been found to be associated with poor sleep and increased risk of depressive symptoms, and HRV was found to increase following successful treatment [31,32]. Additionally, there appears to be a relation between HRV and disease severity, as the reduction in HRV was found to be more prominent when depressive symptoms were worse [33,34]. Electrodermal activity (EDA) sensors can measure perspiration. Multiple studies have found reduced

EDA in depressed patients [35–37], and one even claimed to be able to differentiate between depressed and non-depressed individuals within their population based on whether skin conductance was above or below a certain threshold [38]. Finally, many smartwatches include temperature sensors. Body temperature is commonly elevated in patients with depression [39–42]. This was related to the circadian rhythm, as it was specifically the nocturnal core temperature [39,40] and evening peripheral temperature [42] that were higher in depressed patients compared to non-depressed participants. As such, these two well-understood mechanisms are often invoked to explain why it might be possible to predict mood from wearables in theory. It should be noted, however, that these studies identified associations between isolated physiological features and depression. Individually, these associations are too small to meaningfully predict a complex psychological outcome such as depression in a broader population. Instead, we found that studies aiming to predict depression typically combine a number of these features in an attempt to model the intricacies of how they interact.

1.3. Introduction to Predictive Algorithms

Research on the prediction of depression makes use of technical jargon unique to the machine learning literature. A brief discussion of commonly encountered concepts is thus needed to cover some recurring terms and methods. In general, a predictive algorithm is the predetermined set of computer instructions for how the data will be processed to ultimately come to the desired predictions. To assess how successful an algorithm is, a measure of the true outcome is measured when possible, or a gold standard within the expert field is used as the "ground truth" or ideal expected result. The goal for the prediction is to be as close as possible to this ground truth. For example, the ground truth of whether a person is depressed can be assessed by a clinician or, somewhat less optimally, by a screener survey, independently from the algorithm. The time between the last collected data and the predicted outcome can vary and is henceforth referred to as the prediction horizon.

After the instructions are coded, the model is trained. This is typically conducted using cross-validation, randomly splitting the data into folds of training and testing sets. For each fold, model parameters are optimized iteratively by exploring combinations of features. The model trained on each training set will predict the outcome found in each test set. Very little input is required at this stage. To conclude the training, the parameters for all folds are integrated into a unified model. The quality of this unified model is then assessed by applying it to yet-unused data that were put aside from the start and comparing the predictions to the ground truth. Cross-validation is a crucial tool to prevent overfitting. This happens when the model is so fine-tuned to the training data that it loses generalizability to new populations [43]. Other ways to minimize the chances of overfitting include regularization and feature selection.

The exact way model performance is assessed depends on the type of prediction it makes. For classification models, performance is commonly assessed using the error matrix and accompanying area under the receiver operator characteristic (AUROC) curve. The error matrix is built up of true (T)/false (F) positives (P)/negatives (N). Then, sensitivity (TP/(TP + FN)), i.e., the proportion of positive cases correctly predicted, and specificity (TN/(TN + FP)), i.e., the proportion of negative cases correctly predicted, are formulated to create the ROC. We obtain the AUROC by taking the integral of this curve, which ranges between 0 and 1, where 0 would mean all outcomes wrongly classified, 0.5 would mean no discriminatory power or similar to random guessing, and 1 would mean all correctly classified. For regression models, to assess performance with AUROC, a threshold is set on the outcome scale to define everything above as positive and below as negative. When no gold standard is present, relative quality measures such as the mean square error and variations on it are used. Generally, the lower the relative quality measure, the better a given model is able to fit a dataset.

Finally, one should be aware of the fundamental difference between the statistical modeling encountered in classical epidemiological research and the predictive modeling

in wearables research. In classical epidemiology, we build models to remove bias from specific estimates, which we then interpret individually in terms of direction and effect size [44,45]. These interpretations often rely on statistical assumptions about the data. In predictive modeling, the focus is instead exclusively on achieving more accurate predictions as measured by the metrics described above. Biased relationships are not a problem and can paradoxically be exploited for the purpose of better prediction [46]. Individual associations are irrelevant, as only the performance of all the variables taken together is assessed. Prioritizing predictive performance allows for very flexible models to exploit both the smallest associations and the most complex interactions in the data without the need for strict statistical assumptions. This flexibility comes at the cost of interpretability, meaning it is not always clear exactly how variables interact with each other to produce the observed results [47]. For instance, a Random Forest algorithm combines logical decision trees to show exactly which features and their values were used to classify a sample, whereas non-linear or complex algorithms are more like a black box. Predictive modelers thus have to make a trade—they can build sophisticated algorithms to predict mental health outcomes from wearable technology, but they surrender the ability to understand *how* physiological features might relate to depression.

2. Search Method

To select papers that describe studies that focused on the use of wearables to predict the presence and severity of depressive symptoms, we conducted a search in PubMed (from the time of inception to June 2023) using the following terms: "wearables", "depression", "predictive algorithms", and variations of these terms. In addition, the reference lists of selected papers were checked for relevant publications. We chose PubMed in particular because the articles found in this database likely reflect best what clinical research professionals will encounter when looking to research the applications of wearables in healthcare. We included original research studies. Review studies were considered not eligible. We considered studies written in English only. Given the scoping nature of this review, the list of included studies is not comprehensive but instead represents the range of methodologies and applications of wearables in this field of research. As such, studies were included if they met each of the following three criteria: (1) using a predictive algorithm to (2) detect or predict depressive symptoms according to clinical standards using (3) data collected using a wearable device. Furthermore, wearables were selected to be wrist-worn and feasible to be used in day-to-day life. Studies using smartphone data were not excluded as long as wrist-worn wearable data were used as well. The reasoning behind this was that the requirement for smartphones would not significantly hinder practical implementation due to the wide availability of smartphones to the general public. Any algorithm training method was acceptable as long as the reported aim was to detect or predict symptoms of depression using wearable data. This includes algorithms detecting current depression or symptom severity, as well as the prediction of future symptom severity or onset of depression. In-person clinical interviews with a mental health care professional (i.e., a psychiatrist or resident) constituted the current gold standard for diagnosis and were therefore preferred as ground truth. In order to maintain our focus on clinical depression, studies that instead used screeners to determine true mood were only included if classifications were made using the clinical cut-off scores.

3. Results

A total of 15 studies were selected for the current review. An overview of all included studies can be found in Table 1. Among the collected studies, there was substantial diversity in the populations that were recruited and how the ground truth was determined. Additionally, the studies varied considerably in how the wearables were used to predict depression. More specifically, we noticed that a wide array of different features was selected using diverse types and brands of wearables. Also, we found a variety of prediction goals and ground truth determination methods. An overview of these methodological differences,

along with the reported quality of each resulting algorithm, can be found in Table 2. The following sections highlight the exemplary methodological features that emerged from this set of studies in more detail.

Table 1. Overview of the selected studies.

Reference	Title	Population (Baseline)	Duration of Study Period
Bai et al., 2021 [48]	Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study	261 outpatients with MDD	12 weeks
Chikersal et al., 2021 [49]	Detecting Depression And Predicting Its Onset Using Longitudinal Symptoms Captured By Passive Sensing: A Machine Learning Approach With Robust Feature Selection	138 college students (20 $\geq$ mild MDD, 118 HC at baseline)	16 weeks (1 semester)
Cho et al., 2019 [50]	Mood Prediction of Patients With Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study	55 (18 MDD, 18 BD-I, 19 BD-II)	2003 days total, over a 2-year study period
Ghandeharioun et al., 2017 [51]	Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data	12 patients with MDD	8 weeks
Griffiths et al., 2022 [52]	Investigation Of Physical Activity, Sleep, And Mental Health Recovery In Treatment Resistant Depression (TRD) Patients Receiving Repetitive Transcranial Magnetic Stimulation (rTMS) Treatment	17 patients with TRD	5 weeks
Horowitz et al., 2022 [53]	Using Machine Learning With Intensive Longitudinal Data To Predict Depression And Suicidal Ideation Among Medical Interns Over Time	2459 first-year training physicians (7.9% MDD, 3.6% suicidal ideation at baseline)	92 days
Kim et al., 2019 [54]	Depression Prediction by Using Ecological Momentary Assessment, Actiwatch Data, and Machine Learning: Observational Study on Older Adults Living Alone	47 elderly ($\geq$mild symptoms of MDD, baseline SGDS $\geq$ 5)	14 days
Lee et al., 2022 [55]	Prediction Of Impending Mood Episode Recurrence Using Real-Time Digital Phenotypes In Major Depression And Bipolar Disorders In South Korea: A Prospective Nationwide Cohort Study	95 patients with MDD	23,459 days total over a 4 year study period.

Table 1. *Cont.*

Reference	Title	Population (Baseline)	Duration of Study Period
Lu et al., 2018 [56]	Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multitask Learning	103 college students (39, 64 HC)	3 months
Mahendran et al., 2019 [57]	Sensor-Assisted Weighted Average Ensemble Model for Detecting Major Depressive Disorder	450 (with complaints of mood swings)	7 days
Makhmutova et al., 2022 [58]	Predicting Changes In Depression Severity Using The PSYCHE-D (Prediction Of Severity Change-Depression) Model Involving Person-Generated Health Data: Longitudinal Case-Control Observational Study	4036 (38.7% HC/minimal symptoms, 61.3% ≥ mild symptoms; 20.7% with severity change after 3 months)	3 months
Mullick et al., 2022 [59]	Predicting Depression in Adolescents Using Mobile and Wearable Sensors: Multimodal Machine Learning–Based Exploratory Study	55 adolescents (≥mild MDD)	24 weeks
Pedrelli et al., 2020 [60]	Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors	31 (with MDD)	8 weeks
Rykov et al., 2021 [61]	Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling	267 (110 ≥ mild MDD/157 HC)	14 days, consecutive
Tazawa et al., 2020 [62]	Evaluating Depression With Multimodal Wristband-Type Wearable Device: Screening And Assessing Patient Severity Utilizing Machine-Learning	86 (30 ≥ mild MDD, 15 BD, 41 HC)	5250 days

BD (-I/II) = Bipolar Disorder (type I/II). HC = Healthy controls. MDD = Major Depressive Disorder. SGDS = Short Geriatric Depression Scale. TRD = Treatment-Resistant Depression.

Table 2. Details of study setups and results.

Reference	Wearable Specifications	Predictive Features	Prediction Outcome	Ground Truth Determination	Quality Measures (Best Model)
Bai et al., 2021 [48]	Mi band-2 Smartphone: OS n.s.	Activity Sleep Heart rate Phone usage	Detect current mood state (stable/swinging)	Change in PHQ-9 score over time more or less than 5 points, and score above the cut-off of 11 (≥moderate depression) or below 5 (remission)	Accuracy ranges [1] = 75.64–84.27% Sensitivity ranges [1] = 85.33–93.33%
Chikersal et al., 2021 [49]	FitBit Flex 2 Smartphone: Android, iOS	Activity Sleep Location Phone use	Predict post-semester depression (depression yes/no)	BDI-II above or below the cut-off of 14 for ≥mild depression	Accuracy = 81.3% F1 = 0.75
			Predict post-semester change in severity (symptoms worsened, yes/no)	Did the BDI-II level (minimal, mild, moderate, severe) change, yes or no	Accuracy = 88.1% F1 = 0.81

Table 2. *Cont.*

Reference	Wearable Specifications	Predictive Features	Prediction Outcome	Ground Truth Determination	Quality Measures (Best Model)
Cho et al., 2019 [50]	FitBit Charge HR or 2 Smartphone: Android, OS n.s.	Activity Sleep Heart rate Light exposure	Predict future mood state (stable/biased)	Mood state classification based on mood score distribution within the study population. 'Biased' for the highest 30%, 'stable' for the lowest 70%.	Accuracy = 67% Sensitivity = 39% Specificity = 74% AUC = 0.67
			Predict future mood episodes (occurs yes/no)	Mood episodes were identified retrospectively by the clinician based on interview and eMoodchart app data (app developed by research group)	Accuracy = 73.1% Sensitivity = 67.2% Specificity = 63.7% AUC = 0.79
Ghandeharioun et al., 2017 [51]	Empatica E4 wristband (both wrists) Smartphone: Android	EDA Temperature Sleep Activity Phone use Location	Predict the current HDRS-17 score	Biweekly HDRS-17	RMSE = 4.5
Griffiths et al., 2022 [52]	FitBit: model n.s.	Activity Sleep	Detect current depression severity (severe yes/no)	PHQ-9 above or below the clinical cut-off of 20 for severe depression	Accuracy = 82% Sensitivity = 82% Specificity = 81% F1 = 0.81
Horowitz et al., 2022 [53]	FitBit Charge 4	Daily mood Activity Sleep Heart rate	Predict end-of-quarter depression (yes/no)	PHQ-9 above or below the cut-off of 10 for ≥moderate depression	AUC = 0.749 All variables: AUC = 0.750
			Predict end-of-quarter suicidal ideation (present yes/no)	Based on the final question of PHQ-9, yes or no	AUC = 0.736 All variables: AUC = 0.699
Kim et al., 2019 [54]	Actiwatch Spectrum PRO	Activity Sleep Light exposure EMA	Detect current depression (depressed yes/no)	Classified as depressed if HDRS $\geq$ 8 (mild) and SGDS $\geq$ 7 (mild or worse)	Accuracy = 91% Sensitivity = 88% Specificity = 94% F1 = 0.90 AUC = 0.96
Lee et al., 2022 [55]	FitBit Charge HR, or 2, or 3 Smartphone: Android, iOS, possibly other OS n.s.	Activity Sleep Heart rate Light exposure	Predict future mood episodes (occurs yes/no)	Mood episodes were identified retrospectively by the clinician based on interview and eMoodchart app data (app developed by research group)	Accuracy = 93.8% Sensitivity = 91.5% Specificity = 94.3% AUC = 97.3%
Lu et al., 2018 [56]	FitBit Charge HR Smartphone: Android, iOS	Heart rate Activity Sleep Location	Predict the current QIDS score	Self-reported QIDS score	4-task model: R2 = 0.36 F1 = 0.77
			Predict current symptom severity level (which category as rated by a clinician)	Only if diagnosed by the clinician at baseline through interview per DSM-5 and QIDS: severity categorized by the clinician (stable, mild, moderate, severe)	

Table 2. *Cont.*

Reference	Wearable Specifications	Predictive Features	Prediction Outcome	Ground Truth Determination	Quality Measures (Best Model)
Mahendran et al., 2019 [57]	Mi band-3	Selected 9 questions of the HDRS-21 Activity	Detect current depression (depressed yes/no)	HDRS-21	Accuracy = 99.01% Sensitivity = 98.4% Specificity = 98.87% F1 = 0.98 PPV = 97.54%
Makhmutova et al., 2022 [58]	FitBit: model n.s.	PHQ-9 (true and predicted) Lifestyle Changes Activity Sleep	Predict current PHQ-9 score	PHQ-9 measured at three-month intervals	Quadratic weighted Cohen κ = 0.476 Adjacent accuracy = 77.6%
			Detect change in severity (change since three months ago yes/no)	PHQ-9 severity level is categorized according to clinical cut-offs (minimal, mild, moderate, moderately severe, severe)	Sensitivity = 55.4% Specificity = 65.3% AUPRC = 0.31
Mullick et al., 2022 [59]	FitBit Inspire HR; software version 1.84.5 Smartphone: Android, iOS	Heart rate Sleep Steps Phone use Location	Predict current PHQ-9 score	Weekly PHQ-9	MAE = 2.39 MSE = 10.28 MAPE = 0.27 RMSE = 2.83
			Detect change in severity level (how many PHQ-9 points up/down)	PHQ-9 severity levels determined with clinical cut-offs (minimal, mild, moderate, moderately severe, severe)	MAE = 3.12 MSE = 20.14 MAPE = 7.16 RMSE = 4.48
Pedrelli et al., 2020 [60]	Empatica E4 wristbands (both wrists) Smartphone: Android	EDA Heart rate Activity Phone use	Predict the current HDRS-17 score	HDRS-17 score derived from biweekly HDRS-28 measurements	MAE = 4.08 RMSE = 5.35 R = 0.56
Rykov et al., 2021 [61]	FitBit Charge 2	Activity Sleep Circadian rhythms	Detect current depression (depressed yes/no)	Depression is classified based on PHQ-9 with a cut-off of 8 (mild or worse)	Accuracy = 80% Sensitivity = 82% Specificity = 78%
Tazawa et al., 2020 [62]	Silmee W20 wristband	Activity Sleep Heart rate Skin temperature UV light exposure	Detect current depression	Assessed by a clinician using HDRS-17, MADRS, or YMRS, and self-reported BDI-II and PSQI.	Accuracy = 76% Sensitivity = 73% Specificity = 79%
			Predict the current HDRS-17 score	HDRS-17 by clinician	R = 0.61 R2 = 0.37 MAE = 4.94

AUC = Area Under Curve. AUPRC = Area Under Precision Recall Curve. BDI-II = Beck Depression Inventory-II. DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th edition. EDA = Electrodermal Activity. EMA = Ecological Momentary Assessment. HDRS-17/21 = Hamilton Depression Rating Scale with 17 or 21 items. MADRS = Montgomery Åsberg Depression Rating Scale. MAE = Mean Absolute Error. MAPE = Mean Absolute Percentage Error. MSE = Mean Squared Error. n.s. = not specified. OS = Operating system. PHQ-9 = Patient Health Questionnaire-9. PPV = Positive Predictive Value. PSQI = Pitsburg Sleep Quality Index. QIDS = Quick Inventory of Depressive Symptomatology. RMSE = Root Mean Square Error. SGDS = Short Geriatric Depression Scale. UV = Ultraviolet. YMRS = Young Mania Rating Scale. [1] Multiple models were made to distinguish between different subgroups. The reported ranges encompass the quality measures of the best model for each distinction.

3.1. Variation in Available Predictive Features

Wearables are able to collect vast amounts of personal data, and most studies use different combinations of the same commonly selected features. For wearables, these include activity (often measured as a number of steps or time in motion), sleep (derived from activity), and heart rate. When smartphones were used, additional features commonly included phone use (apps, communication) and location. However, some research groups utilized less common features. For example, Chikersal et al. [49] used Bluetooth to infer how many different people the user met. Based on how frequently devices were scanned to be in proximity, devices were categorized as "self" (most frequently scanned) and "other" (less frequently scanned). Moreover, as the study population consisted entirely of college students who lived on campus, common locations on campus were labeled and visits to these locations were tracked. Both features offered additional insight into the social network and activities of the participants. These types of features are difficult to translate to the general public but could be interesting in targeting populations that spend the majority of their time frequenting the same general area.

Ghandeharioun et al. [51] had each participant wear two Empatica E4 wristbands, one on each wrist. This allowed consideration of the symmetry of electrodermal activity between the left and right wrist. The number of skin conductance responses and magnitude of EDA showed stronger asymmetry on days when mental health was poor. The study population was small, so these results should be considered preliminary. A follow-up study with a slightly larger sample size used the same two-wristband approach and found that two of the top ten features were related to EDA, particularly the left/right asymmetry [60]. For practical application, two wristbands are less convenient for the user, but EDA asymmetry in depression may still be interesting to investigate for short-term application or from a fundamental research perspective.

Less commonly, survey data were included as a feature alongside wearable data. For example, one study used a selection of nine Hamilton Depression Rating Scale (HDRS)-21 questions in addition to wearable activity data to predict HDRS-21 scores. This produced an incredibly accurate model (99% accuracy), but this should be considered with caution as using part of the ground truth to predict the ground truth likely inflated the accuracy [57]. Another study first created an algorithm using a combination of wearable-collected sleep data, patient demographic and lifestyle survey data, and Patient Health Questionnaire (PHQ)-9 data to predict PHQ-9 severity levels in between survey moments. The static demographic data were found to be most important for this prediction. A second algorithm then used all available data to predict the PHQ-9 score 3 months after baseline. The previously predicted PHQ-9 development was one of the most important features of this second algorithm. The authors acknowledged that the need for surveys should be minimized to optimize future applications of such algorithms [58].

3.2. Feature Selection

From the wide variety of available features, each study makes a selection of features to use for their model. Consumer-grade wearables differ in the amount and sophistication of sensors they can provide, which is the first limiting factor researchers encounter in which features they select. However, sometimes, pre-existing hypotheses about the underlying mechanisms of depression also influence feature selection. For example, Cho et al. [50] and Lee et al. [55] chose to look specifically at features involving the circadian rhythm. Alternatively, data-driven approaches can be applied for selection. An often necessary first step in this process is to remove features that are heavily correlated with each other and thus provide redundant information [57–59]. Another data-driven approach is to create several model iterations using different selections of features and choosing the selection that results in the best prediction accuracy. This method was used, for example, by Griffiths et al. [52], to choose how many features to include in their model. After identifying the features with the highest importance, they created models using different numbers of features (top 5, top 10, etc.). As determined by cross-validation, the model using the top

10 features performed best, letting them predict symptom severity with up to 82% accuracy. Similarly, Pedrelli et al. [60] made models using different groups of features (smartphone only, wearable only, or both) to estimate HDRS scores. They then used two cross-validation approaches. The first simulated the tracking of single patients by using the first weeks for training and the last weeks for testing. The second focused on generalizability, splitting the participants into a training and test group. They found the smartphone-only model to be more accurate for the more personalized approach, whereas the wearable-only model allowed for better generalizability.

It is important to also realize how feature selection can lead to overfitting when conducted incorrectly. For example, Kim et al. [54] first classified their population as either depressed or not depressed, then compared these two groups and selected the features that were most predictive for classifying individuals. Since it is not reported otherwise, it must be assumed they made this comparison without first splitting the data into a training and test group. While these features may be particularly relevant within this population, it is impossible to know whether this would be the same in a new and unseen population. In a similar manner, Rykov et al. [61] explicitly state they based their feature selection on the statistical associations found over the entire sample. This again greatly compromises generalizability to new populations.

3.3. Prediction Outcomes and Horizons

A predictive algorithm can be used in many ways to anticipate depressive symptoms. Variation may lie in the type of outcome that is predicted or the prediction horizon. The prediction outcome may be a continuous score indicating symptom severity or a categorical outcome classifying by either depression diagnosis (yes or no) or clinical cut-off (above or below a clinical cut-off). With a smaller prediction horizon, a model may predict whether symptoms are being experienced in the present moment or whether symptoms have changed over recent weeks. With a larger prediction horizon, a model may predict whether depressive symptoms will change in the near future or even whether they will arise several weeks in advance. Each of these outcomes can be useful as long as the appropriate context is considered.

The following will highlight some studies to exemplify the various options. Firstly, Ghandeharioun et al. [51] looked at a continuous outcome, aiming to predict daily HDRS-17 scores for a period of 8 weeks. Predicting daily mood scores this way may be useful for more detailed symptom tracking in between regular treatment visits. Conversely, Griffiths et al. [52] used a binary outcome, as their model predicted whether the current PHQ-9 scores of patients with treatment-resistant depression (TRD) were above or below the clinical cut-off (PHQ-9 $\geq$ 20) for severe depression. Such a binary outcome reduces nuance but makes interpretability more direct in the clinic. In fact, in clinical practice, the category is often sufficient without knowing an exact severity score. The authors note that this type of prediction for symptom monitoring is likely redundant for TRD patients due to their frequent in-person monitoring. However, with some adjustments, predicting on which side of the clinical cut-off patients fall can be useful for monitoring treatment progress, for example.

Bai et al. [48] investigated the binary outcome of whether symptoms changed over time. This was classified based on the difference between the highest and lowest of three consecutive PHQ-9 scores, measured at least one week apart. The distinction was made primarily between swinging mood (symptom change) and stable mood (no symptom change), and groups were further specified as moderate or drastic swing and stable remission or depression for a total of four outcome groups. A plethora of models were created that decided which of two groups an individual belonged in, each using different features and algorithm training methods. A major disadvantage of this method is that each model assumes a binary outcome (e.g., stable depression or moderate swing), thereby ignoring the possibility that the individual belongs to one of the other groups (e.g., stable remission). The highest accuracy was reached with models deciding between stable remission and

moderate swing and between stable remission and drastic swing. These particular models may be useful for patients known to be in stable remission who wish to be alerted early of potential relapses.

Exploring a slightly larger prediction horizon, Cho et al. [50] aimed to predict mood states and impending mood episodes in patients with pre-existing mood disorders (MDD, Bipolar disorder type I/II). The algorithm used 18 days' worth of data to predict mood state and occurrence of mood episodes over the following three days. The ground truth was determined using the absolute mood score (AMS) as self-reported in the eMoodChart app that was newly designed for this study. The ground truth distinction between stable (neutral) and biased (depressed, manic, hypomanic) was made using AMS percentiles rather than cut-off scores. For example, the best model considered the mood state of the 30% highest AMS to be biased and the lowest 70% to be stable. As such, the distinction relies on group distribution rather than true mood score. The occurrence of mood episodes was instead confirmed through in-person interviews, as is the clinical standard. A similar study was conducted by the same group with a larger sample size [55]. They predicted mood episodes over the same prediction horizon, removing the reliance on AMS distribution for the ground truth. The reported accuracy for predicting depressive episodes in MDD patients was very high. However, for lack of reporting otherwise, presumably, all recorded episodes were used for retrospective feature selection, resulting in overfitting. The prediction of impending mood episodes a few days before they occur would give patients forewarning, allowing them to contact their clinician and be better prepared to deal with the episode.

Further expanding the prediction horizon, Chikersal et al. [49] aimed to predict symptoms of depression several weeks in advance. Specifically, over the course of a 16-week study period, two binary outcomes were predicted each week using the data available at that time: whether participants would have depression at the final check-up and whether their symptoms then would be worse compared to baseline. Depression and worsening symptoms were predicted with acceptable accuracy (> 80%) up to 9 and 14 weeks in advance, respectively. A caveat is that no mood data were collected outside of the Beck Depression Inventory-II (BDI-II) at the start and end of the 16-week study period, so it is unknown when exactly symptoms started developing. As such, it is unclear exactly how far in advance the model is able to predict the onset of depression. However, because it is unlikely that all 36 new cases would have been classified as newly depressed after the first two weeks, it is safe to assume that the number of cases was indeed predicted correctly several weeks in advance. A major benefit of predicting this far in advance is that these passive sensing models may be able to pick up on early signs of depression that participants themselves may not be aware of yet and that would, therefore, be missed through regular self-reported measures. This allows earlier detection and, most importantly, the possibility of earlier intervention.

Similarly, Horowitz et al. [53] also aimed to predict whether depression occurred several weeks in advance. Specifically, the two outcomes they predicted were the presence of depression and suicidal ideation at the end of a quarter (13 weeks). Each day, a new iteration of the models was made using the data available up to that time. To predict depression with an acceptable accuracy (AUC > 0.7), a minimum of two weeks' worth of data was needed, whereas prediction of suicidal ideation required at least 7 weeks' worth of data. Predicting participants' mood status was thus possible 6–11 weeks in advance. Model variations were created that used only the mood variables based on the self-reported daily mood scores or using all variables, including mood and wearable data. For predicting depression, accuracy was comparable across variations, while suicidal ideation was predicted slightly better using only mood variables. This implies wearable data are not necessary for accurate predictions as long as participants submit mood data. However, a major benefit of wearables is that they do not require user input and are, therefore, less obtrusive than daily mood surveys.

3.4. The Duration of Data Collection

Aside from feature selection and prediction outcomes, another interesting consideration is that different prediction outcomes require different durations of data collection. Since most studies only looked at a single type of outcome, this was not typically investigated in a direct manner. Tazawa et al. [62] did look at two outcomes. First to detect current depression (binary), and second to predict current severity score (continuous). They compared algorithms using 3 and 7 days' worth of data to investigate if a longer period was beneficial. While the prediction of depression showed no difference, the severity score was predicted significantly better with 7 days' worth of data. This implies that more complex predictions (continuous score as opposed to binary classification) may particularly benefit from the additional measurements. As mentioned before, Chikersal et al. [49] also looked at two types of predictions and found that predicting the onset of depression required 7 weeks' worth of data, whereas predicting worsening symptoms only required a minimum of 2 weeks. In a similar comparison, Horowitz et al. [53] found that predicting future depression was accurate after only 2 weeks, whereas suicidal ideation was not accurately predicted until 7 weeks' worth of data were available. The difference in time needed to predict future depression is likely due to the different model training methods and feature selection used, which both heavily influence the data requirements. In the current review, not enough studies focus on this aspect to draw a conclusion on how much data is required for specific types of predictions. Nonetheless, it is advisable to remember that some predictions and algorithm training methods benefit from longer periods of data collection. A practical consideration when making such a model should, therefore, be whether that duration is suitable for the model's intended application.

3.5. Types of Machine Learning Models (Ensemble vs. Single Type)

There are many ways of training a predictive algorithm. Some examples include linear regression, decision trees, or gradient boosting. Additionally, it is possible to combine several training models into an ensemble model. Such ensemble models have been found by multiple studies to have superior accuracy when compared to single-type models. For example, Lu et al. [56] used multitask learning to create their predictive algorithms that made two predictions: the current Quick Inventory of Depressive Symptomatology (QIDS) score for all participants (healthy and depressed) and the current severity of depressive symptoms for patients. The tasks for the multitask models were based on the type of prediction, (1) QIDS score, and (2) severity, as well as the operating system of the user's smartphone, (1) iOS and (2) Android. Two-task models each did one type of prediction for the two different operating systems, and four-task models combined QIDS score and severity prediction for both operating systems. Overall, the multitask models outperformed the single-task ones, and the four-task models generally performed better than the two-task models, showing the potential of combined predictive models to enhance each other. Additionally, the models using FitBit and smartphone data performed slightly better than those using smartphone data alone, showing that the wearable data can also improve the predictive capabilities of these models.

Another study that used a combination of different training methods was conducted by Mahendran et al. [57]. The algorithms that were compared used logistic regression, random forest, and a weighted average ensemble that combined the two. The accuracy reported for logistic regression and for random forest was generally high, but they found the weighted average ensemble model to be the most accurate of all. As mentioned earlier, the accuracy may be inflated due to the use of some HDRS-21 data to predict HDRS-21 scores, but this was the case for all three model variations. As such, this study again indicates that ensemble models can improve predictive quality. While the accuracy of ensemble models seems better than that of singular models, they also come with some disadvantages. Firstly, more complex models require more computing power and resources, which may not be available or desirable. Secondly, the increasing complexity of a model increases the risk of overfitting, whereas simpler models are more robust. Finally, due to the

complexity of ensemble models, it is much more difficult to rationalize and understand how predictions are made compared to single-type models. Depending on the data complexity and prediction goal, a fitting model complexity should be chosen.

3.6. Common Limitations

Finally, while wearables offer a lot of potential and many of the discussed models can predict depressive symptoms with impressive accuracy, it is also important to go over a number of common limitations and pitfalls that should be avoided when further researching this topic. Some of these have been mentioned throughout the previous sections. First, when best practices are not followed, there is a high risk of overfitting a model during feature selection, thus limiting generalizability [54,55,61]. Second, multiple studies used very small groups of participants, which is also likely to limit generalizability [50–52,54,59,60]. Third, some models relied on active user input through mood-scoring apps or surveys [53,54,57]. This directly conflicts with one of the main strengths of wearables: the ability to collect data without active user contribution. Fourth, none of the studies reviewed tested their model on independently collected validation datasets, thus increasing the risk of bias. This is an issue also highlighted in a recent review of models using smartphone data to predict depression [63]. These limitations combined make it highly likely that the impressive results observed in individual studies might fall short if the algorithms were applied to the real world. The following section will discuss some common limitations found in the reviewed studies, as well as general limitations that need to be considered when researching wearable predictive models.

One of these recurring limitations was that study populations had limited generalizability due to the overrepresentation of certain demographics. For example, many studies used populations that were predominantly female [51,52,56,58–61]. Additionally, people over the age of 65 were only included in a single study [54]. This is unfortunate, as this type of research may be particularly useful for the elderly population, which is at an increased risk of experiencing depression [64]. Lastly, most studies looked at participants who were predominantly White [56,58–60] or Asian [48,54,61,62], whereas Black people were in the minority if they were included at all. The latter is especially relevant because one of the common wearable sensors, the photoplethysmography (PPG) sensor, relies on how green light is absorbed by the skin and blood flowing underneath to measure heart rate. The accuracy of this sensor was found to be dependent on skin color, with less accuracy for darker skin tones [65]. Moreover, the fact that Black people are often underrepresented in validation studies is a systemic issue that should be taken into account when considering the use of wearables for health tracking in these populations.

Some other methodological challenges include the lack of standard protocols due to the novelty of this type of research. This is illustrated by the fact that many studies use different methods to establish the ground truth. Some relied on mood-scoring methods newly designed for their study [50], others used in-person visits with a psychiatrist [62], and the majority used one or more in-person or self-reported questionnaires such as the Hamilton Depression Rating Scale, Patient Health Questionnaire-9, or Beck Depression Inventory-II [49,54,58]. This large variety of methods complicates direct comparisons between studies. The reporting of methodological details is not always clear and complete. For example, the model of the wearable is not always specified [52,58], and the operating systems of smartphones used are not always disclosed [48,50,55]. Similarly, the way the predictive quality of the algorithms is reported also varies. In part, this is because binary classifications are simply assessed in a different way than when the outcome is a numeric score. However, even within the same types of predictions, the reported quality outcomes vary. This makes direct comparisons especially complicated when common measures (e.g., specificity, F1-score, mean absolute error) are not reported [48,51,56]. The variability in methods and reporting quality has been reported on before in more detail [66]. A last note on the methods is that patients are commonly not considered as separate individuals but rather as a collection of data that can be compiled with all the other patients' data

into one big dataset for the algorithm to analyze [55,56]. Alternatively, it also occurred that repeated measurements from the same patient were pooled together and treated as independent observations [58,62]. This can artificially increase the sample size and mislead the certainty with which algorithms make predictions about personal trajectories. It should finally be noted that the search strategy for this review did not include databases focused on computer science or engineering and instead used PubMed as the primary database in order to reflect the studies clinical research professionals would most likely encounter. Readers should be aware that this has likely led to the exclusion of interesting articles written with different disciplinary backgrounds and with potentially different approaches and limitations.

A final practical consideration is that the companies that make these consumer-grade wearables are continuously updating their models and releasing new versions of them. While new versions presumably are of similar or improved quality, this does add another factor that can vary when comparing the results of different studies. Additionally, not all studies specify which version they used and instead only report the brand of the device [52,58].

4. Discussion

The papers reviewed here altogether indicate wearable data have a lot of potential to be used for the detection, monitoring, and prediction of depressive symptoms. These predictions can be used for a wide variety of purposes, ranging from monitoring existing symptoms to facilitating early interventions. Exactly how much data is needed varies for each of these purposes and should be fine-tuned to the type of prediction a given model aims to make. Models can be made with diverse complexity, and more complex models tend to be slightly more accurate. However, simpler models are less likely to be overfitted, and these predictions are easier to grasp by humans. Certain types of wearable features (e.g., activity, sleep) were frequently used in the literature to successfully predict depression outcomes, although novel features (e.g., electrodermal activity, phone use) were also occasionally encountered. These features hold the potential to either better tune model predictions or inspire new directions altogether in research on how depression operates.

Application of Wearables in Mental Health Care

The widespread application of wearables in mental healthcare is the obvious end goal for this type of research. However, there are a number of real-world complexities yet to be solved. To start with, wearables collect massive amounts of personal data, which usually means they are reliant on cloud storage and capacity. This not only requires a lot of resources but can also have a significant environmental impact [67]. Moreover, the collection, transmission, and storage of large amounts of personal data need to meet a high standard of data security [68,69]. Similarly, there is the non-trivial issue of data privacy, which is an important ethical consideration but also a deciding factor for whether patients would trust these types of algorithms enough to use them [70,71]. There would be much to gain by independent systematic evaluation of the many different data collectors and health apps, preferably through international collaboration [72]. Finally, from a socioeconomic point of view, it is important to consider that access to wearables is not universal, and the accessibility of predictive algorithms is based in large part on the user's economic status and insurance. There is also considerable sampling and selection bias, as health-conscious people are more likely to participate in both studies and applications [73]. It should be noted that the US Food and Drug Administration (FDA) has outlined plans to regulate artificial intelligence and machine learning algorithms for health applications [74]. Their primary focus is to ensure that algorithms implemented in healthcare are explainable, transparent, secure to cyberattacks, and not biased towards certain sociodemographic groups. It remains to be seen how future academic research on mental healthcare responds to institutional regulation on wearables and machine learning.

Given the importance of the doctor–patient relationship, it is unlikely that wearables will ever fully replace the diagnosis by a clinician [75]. Instead, wearables may prove more useful as a complementary tool. For example, they could be used to monitor recovery progress in between visits and for early detection of relapse for patients with a previous diagnosis. By contrast, algorithms that predict the future onset of depression several weeks in advance seem able to detect physiological or behavioral traits indicating prodromal symptoms of depression, symptoms that go under the radar of traditional diagnoses. Since both reviewed studies with a longer prediction horizon looked at highly specific populations they considered to be higher risk [49,53], it would be interesting to initially apply these models to specific groups of at-risk individuals. The reviewed studies both looked at first-year medical students or trainees, but other groups could include people who are grieving or who have recently become unemployed [76,77]. Predicting this far in advance is admittedly a much less common approach in research, so there is still a lot of progress to be made in this direction, but the results so far are promising.

Alternatively, a more research-focused approach can also be explored. Most studies simply observe which features are used by the algorithm, and some do report which features have the biggest predictive value for the model. A question that is often left open is why these features are so informative. This is a very interesting and unexplored area of research that may provide novel insights into the underlying mechanisms of depression. Furthermore, these important features could potentially inspire new or more personalized interventions. For example, if a patient is classified as depressed based most importantly on their excessive late-night activity, they might benefit from a treatment that focuses more on their sleeping habits. However, the efficacy of this approach is merely speculative, as no studies were found looking at such personalized treatment adjustments.

5. Conclusions

All in all, the fact remains that the algorithms were able to predict depressive symptoms with better accuracy than purely random prediction. Even with the noted limitations, these are promising results. The use of consumer-grade wearables and machine learning in this manner is still a very novel approach that still has a long way to go before any of the widespread applications mentioned above would be feasible. Improvement is especially constructive in the development of more standardized methods and clearer guidelines for conducting and reporting these kinds of studies. This would naturally improve reproducibility, comparability across studies, and, most importantly, the predictive quality and validity of the models. The next steps would include larger-scale studies, ideally proportionally more studies with longer prediction horizons, an exploration of potential intervention strategies, and more focus on the research implications of these algorithms in order to truly reach the full potential of wearable technology in mental health care.

Author Contributions: Conceptualization, Q.W. and A.K.; methodology, Q.W., M.Z., M.U. and A.K.; validation, M.Z. and M.U.; investigation, Q.W.; data curation, Q.W.; writing—original draft preparation, Q.W. and A.K.; writing—review and editing, Q.W., M.Z., M.U. and A.K.; supervision, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: Astrid Kamperman and Milan Zarchev are funded by the Epidemiological and Social Psychiatric Research Institute (ESPRi), a consortium of academic and non-academic research groups at the following institutes of mental health care (GGz): Parnassia Psychiatric Institute Antes, GGz Breburg, GGz Delfland, GGz Westelijk Noord-Brabant and Yulius.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Omura, J.D.; Carlson, S.A.; Paul, P.; Watson, K.B.; Fulton, J.E. National physical activity surveillance: Users of wearable activity monitors as a potential data source. *Prev. Med. Rep.* **2017**, *5*, 124–126. [CrossRef] [PubMed]

2. Thompson, W.R. Worldwide Survey Of Fitness Trends For 2019. *ACSM's Health Fit. J.* **2018**, *22*, 10–17. [CrossRef]
3. Wearables—Statistics & Facts. Available online: https://www.statista.com/topics/1556/wearable-technology/-topicOverview (accessed on 10 May 2023).
4. Cruz, S.; Lu, C.; Ulloa, M.; Redding, A.; Hester, J.; Jacobs, M. Perceptions of Wearable Health Tools Post the COVID-19 Emergency in Low-Income Latin Communities: Qualitative Study. *JMIR Mhealth Uhealth* **2024**, *12*, e50826. [CrossRef] [PubMed]
5. Schueller, S.M.; Neary, M.; Lai, J.; Epstein, D.A. Understanding People's Use of and Perspectives on Mood-Tracking Apps: Interview Study. *JMIR Ment. Health* **2021**, *8*, e29368. [CrossRef]
6. Saganowski, S.; Dutkowiak, A.; Dziadek, A.; Dzieżyc, M.; Komoszyńska, J.; Michalska, W.; Polak, A.; Ujma, M.; Kazienko, P. Emotion Recognition Using Wearables: A Systematic Literature Review—Work-in-progress. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020.
7. Shu, L.; Yu, Y.; Chen, W.; Hua, H.; Li, Q.; Jin, J.; Xu, X. Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet. *Sensors* **2020**, *20*, 718. [CrossRef]
8. Abd-Alrazaq, A.; AlSaad, R.; Aziz, S.; Ahmed, A.; Denecke, K.; Househ, M.; Farooq, F.; Sheikh, J. Wearable Artificial Intelligence for Anxiety and Depression: Scoping Review. *J. Med. Internet Res.* **2023**, *25*, e42672. [CrossRef]
9. Ahmed, A.; Aziz, S.; Alzubaidi, M.; Schneider, J.; Irshaidat, S.; Abu Serhan, H.; Abd-Alrazaq, A.; Solaiman, B.; Househ, M. Wearable devices for anxiety & depression: A scoping review. *Comput. Methods Programs Biomed. Update* **2023**, *3*, 100095. [CrossRef]
10. Kang, M.; Chai, K. Wearable Sensing Systems for Monitoring Mental Health. *Sensors* **2022**, *22*, 994. [CrossRef] [PubMed]
11. Lee, S.; Kim, H.; Park, M.J.; Jeon, H.J. Current Advances in Wearable Devices and Their Sensors in Patients with Depression. *Front. Psychiatry* **2021**, *12*, 672347. [CrossRef]
12. Depressive Disorder (Depression). Available online: https://www.who.int/news-room/fact-sheets/detail/depression (accessed on 15 May 2023).
13. Williams, S.Z.; Chung, G.S.; Muennig, P.A. Undiagnosed depression: A community diagnosis. *SSM Popul. Health* **2017**, *3*, 633–638. [CrossRef]
14. Liu, Q.; He, H.; Yang, J.; Feng, X.; Zhao, F.; Lyu, J. Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *J. Psychiatr. Res.* **2020**, *126*, 134–140. [CrossRef] [PubMed]
15. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5™*, 5th ed.; American Psychiatric Publishing, a Division of American Psychiatric Association: Washington, DC, USA, 2013.
16. Halfin, A. Depression: The benefits of early and appropriate treatment. *Am. J. Manag. Care* **2007**, *13*, S92–S97. [PubMed]
17. Vitaterna, M.H.; Takahashi, J.S.; Turek, F.W. Overview of circadian rhythms. *Alcohol Res. Health* **2001**, *25*, 85–93. [PubMed]
18. Boyce, P.; Barriball, E. Circadian rhythms and depression. *Aust. Fam. Physician* **2010**, *39*, 307–310.
19. Soria, V.; Urretavizcaya, M. Circadian rhythms and depression. *Actas Esp. Psiquiatr.* **2009**, *37*, 222–232.
20. Coppen, A.; Abou-Saleh, M.; Milln, P.; Metcalfe, M.; Harwood, J.; Bailey, J. Dexamethasone suppression test in depression and other psychiatric illness. *Br. J. Psychiatry* **1983**, *142*, 498–504. [CrossRef]
21. Green, H.S.; Kane, J.M. The dexamethasone suppression test in depression. *Clin. Neuropharmacol.* **1983**, *6*, 7–24. [CrossRef]
22. Tsigos, C.; Chrousos, G.P. Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. *J. Psychosom. Res.* **2002**, *53*, 865–871. [CrossRef]
23. The Connection between Anxiety and Body Temperature. Available online: https://www.calmclinic.com/anxiety/symptoms/body-temperature (accessed on 22 June 2023).
24. Cortisol. Available online: https://my.clevelandclinic.org/health/articles/22187-cortisol (accessed on 22 June 2023).
25. Vedder, H. Physiology of the Hypothalamic-Pituitary-Adrenocortical Axis. In *NeuroImmune Biology*, 1st ed.; Rey, A.D., Chrousos, G., Besedovsky, H., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Volume 7, pp. 154–196.
26. Helgadóttir, B.; Forsell, Y.; Ekblom, Ö. Physical activity patterns of people affected by depressive and anxiety disorders as measured by accelerometers: A cross-sectional study. *PLoS ONE* **2015**, *10*, e0115894. [CrossRef]
27. De Mello, M.T.; de Aquino Lemos, V.; Antunes, H.K.M.; Bittencourt, L.; Santos-Silva, R.; Tufik, S. Relationship between physical activity and depression and anxiety symptoms: A population study. *J. Affect. Disord.* **2013**, *149*, 241–246. [CrossRef]
28. Raoux, N.; Benoit, O.; Dantchev, N.; Denise, P.; Franc, B.; Alliale, J.-F.; Widlöcher, D. Circadian pattern of motor activity in major depressed patients undergoing antidepressant therapy: Relationship between actigraphic measures and clinical course. *Psychiatry Res.* **1994**, *52*, 85–98. [CrossRef] [PubMed]
29. Todder, D.; Caliskan, S.; Baune, B.T. Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression. *World J. Biol. Psychiatry* **2009**, *10*, 276–284. [CrossRef] [PubMed]
30. Minaeva, O.; Booij, S.H.; Lamers, F.; Antypa, N.; Schoevers, R.A.; Wichers, M.; Riese, H. Level and timing of physical activity during normal daily life in depressed and non-depressed individuals. *Transl. Psychiatry* **2020**, *10*, 259. [CrossRef] [PubMed]
31. Balogh, S.; Booij, S.H.; Lamers, F.; Antypa, N.; Schoevers, R.A.; Wichers, M.; Riese, H. Increases in heart rate variability with successful treatment in patients with major depressive disorder. *Psychopharmacol. Bull.* **1993**, *29*, 201–206.
32. da Estrela, C.; McGrath, J.; Booij, L.; Gouin, J.-P. Heart Rate Variability, Sleep Quality, and Depression in the Context of Chronic Stress. *Ann. Behav. Med.* **2021**, *55*, 155–164. [CrossRef]

33. Agelink, M.W.; Boz, C.; Ullrich, H.; Andrich, J. Relationship between major depression and heart rate variability.: Clinical consequences and implications for antidepressive treatment. *Psychiatry Res.* **2002**, *113*, 139–149. [CrossRef]
34. Kemp, A.H.; Quintana, D.S.; Gray, M.A.; Felmingham, K.L.; Brown, K.; Gatt, J.M. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biol. Psychiatry* **2010**, *67*, 1067–1074. [CrossRef]
35. Iacono, W.G.; Lykken, D.T.; Peloquin, L.J.; Lumry, A.E.; Valentine, R.H.; Tuason, V.B. Electrodermal activity in euthymic unipolar and bipolar affective disorders. A possible marker for depression. *Arch. Gen. Psychiatry* **1983**, *40*, 557–565. [CrossRef]
36. Mestanikova, A.; Ondrejka, I.; Mestanik, M.; Hrtanek, I.; Snircova, E.; Tonhajzerova, I. Electrodermal Activity in Adolescent Depression. *Adv. Exp. Med. Biol.* **2016**, *935*, 83–88. [CrossRef]
37. Storrie, M.C.; Doerr, H.O.; Johnson, M.H. Skin conductance characteristics of depressed subjects before and after therapeutic intervention. *J. Nerv. Ment. Dis.* **1981**, *169*, 176–179. [CrossRef]
38. Ward, N.G.; Doerr, H.O. Skin conductance. A potentially sensitive and specific marker for depression. *J. Nerv. Ment. Dis.* **1986**, *174*, 553–559. [CrossRef] [PubMed]
39. Avery, D.H.; Shah, S.H.; Eder, D.N.; Wildschindtz, G. Nocturnal sweating and temperature in depression. *Acta Psychiatr. Scand.* **1999**, *100*, 295–301. [CrossRef]
40. Avery, D.H.; Wildschiødtz, G.; Rafaelsen, O.J. Nocturnal temperature in affective disorder. *J. Affect. Disord.* **1982**, *4*, 61–71. [CrossRef]
41. Rausch, J.L.; Johnson, M.; Corley, K.; Hobby, H.; Shendarkar, N.; Fei, Y.; Ganapathy, V.; Leibach, F. Depressed patients have higher body temperature: 5-HT transporter long promoter region effects. *Neuropsychobiology* **2003**, *47*, 120–127. [CrossRef] [PubMed]
42. Tocchetto, B.F.; Ramalho, L.; Zortea, M.; Bruck, S.M.; Tomedi, R.B.; Alves, R.L.; da Silva Torres, I.L.; Fregni, F.; Caumo, W. Peripheral body temperature rhythm as a marker of the severity of depression symptoms in fibromyalgia. *Biol. Psychol.* **2023**, *177*, 108494. [CrossRef] [PubMed]
43. Molnar, C.; Freiesleben, T. *Supervised Machine Learning for Science: How to Stop Worrying and Love Your Black Box*, 1st ed.; Bookdown, 2024; Available online: https://ml-science-book.com/ (accessed on 24 April 2024).
44. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
45. Hernán, M.A.; Hsu, J.; Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **2019**, *32*, 42–49. [CrossRef]
46. Yarkoni, T.; Westfall, J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect. Psychol. Sci.* **2017**, *12*, 1100–1122. [CrossRef]
47. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed.; Bookdown, 2022; Available online: https://christophm.github.io/interpretable-ml-book (accessed on 24 April 2024).
48. Bai, R.; Xiao, L.; Guo, Y.; Zhu, X.; Li, N.; Wang, Y.; Chen, Q.; Feng, L.; Wang, Y.; Yu, X.; et al. Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study. *JMIR Mhealth Uhealth* **2021**, *9*, e24365. [CrossRef]
49. Chikersal, P.; Doryab, A.; Tumminia, M.; Villalba, D.K.; Dutcher, J.M.; Liu, X.; Cohen, S.; Creswell, K.G.; Mankoff, J.; Creswell, J.D.; et al. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach with Robust Feature Selection. *ACM Trans. Comput.-Hum. Interact.* **2021**, *28*, 3. [CrossRef]
50. Cho, C.H.; Lee, T.; Kim, M.-G.; In, H.P.; Kim, L.; Lee, H.-J. Mood Prediction of Patients With Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study. *J. Med. Internet Res.* **2019**, *21*, e11029. [CrossRef] [PubMed]
51. Ghandeharioun, A.; Fedor, S.; Sangermano, L.; Ionescu, D.; Alpert, J.; Dale, C.; Sontag, D.; Picard, R. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017.
52. Griffiths, C.; da Silva, K.M.; Leathlean, C.; Jiang, H.; Ang, C.S.; Searle, R. Investigation of physical activity, sleep, and mental health recovery in treatment resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rTMS) treatment. *J. Affect. Disord. Rep.* **2022**, *8*, 100337. [CrossRef]
53. Horwitz, A.G.; Kentopp, S.D.; Cleary, J.; Ross, K.; Wu, Z.; Sen, S.; Czyz, E.K. Using machine learning with intensive longitudinal data to predict depression and suicidal ideation among medical interns over time. *Psychol. Med.* **2022**, *53*, 5778–5785. [CrossRef]
54. Kim, H.; Lee, S.; Lee, S.; Hong, S.; Kang, H.; Kim, N. Depression Prediction by Using Ecological Momentary Assessment, Actiwatch Data, and Machine Learning: Observational Study on Older Adults Living Alone. *JMIR Mhealth Uhealth* **2019**, *7*, e14149. [CrossRef] [PubMed]
55. Lee, H.J.; Cho, C.-H.; Lee, T.; Jeong, J.; Yeom, J.W.; Kim, S.; Jeon, S.; Seo, J.Y.; Moon, E.; Baek, J.H.; et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: A prospective nationwide cohort study. *Psychol. Med.* **2022**, *53*, 5636–5644. [CrossRef]
56. Lu, J.; Shang, C.; Yue, C.; Morillo, R.; Ware, S.; Kamath, J.; Bamis, A.; Russell, A.; Wang, B.; Bi, J. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 21. [CrossRef]
57. Mahendran, N.; Vincent, D.R.; Srinivasan, K.; Chang, C.-Y.; Garg, A.; Gao, L.; Reina, D.G. Sensor-Assisted Weighted Average Ensemble Model for Detecting Major Depressive Disorder. *Sensors* **2019**, *19*, 4822. [CrossRef] [PubMed]

58. Makhmutova, M.; Kainkaryam, R.; Ferreira, M.; Min, J.; Jaggi, M.; Clay, I. Predicting Changes in Depression Severity Using the PSYCHE-D (Prediction of Severity Change-Depression) Model Involving Person-Generated Health Data: Longitudinal Case-Control Observational Study. *JMIR Mhealth Uhealth* **2022**, *10*, e34148. [CrossRef]
59. Mullick, T.; Radovic, A.; Shaaban, S.; Doryab, A. Predicting Depression in Adolescents Using Mobile and Wearable Sensors: Multimodal Machine Learning-Based Exploratory Study. *JMIR Form. Res.* **2022**, *6*, e35807. [CrossRef]
60. Pedrelli, P.; Fedor, S.; Ghandeharioun, A.; Howe, E.; Ionescu, D.F.; Bhathena, D.; Fisher, L.B.; Cusin, C.; Nyer, M.; Yeung, A.; et al. Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors. *Front. Psychiatry* **2020**, *11*, 584711. [CrossRef]
61. Rykov, Y.; Thach, T.-Q.; Bojic, I.; Christopoulos, G.; Car, J. Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling. *JMIR Mhealth Uhealth* **2021**, *9*, e24872. [CrossRef] [PubMed]
62. Tazawa, Y.; Liang, K.-C.; Yoshimura, M.; Kitazawa, M.; Kaise, Y.; Takamiya, A.; Kishi, A.; Horigome, T.; Mitsukura, Y.; Mimura, M.; et al. Evaluating depression with multimodal wristband-type wearable device: Screening and assessing patient severity utilizing machine-learning. *Heliyon* **2020**, *6*, e03274. [CrossRef] [PubMed]
63. Leaning, I.E.; Ikani, N.; Savage, H.S.; Leow, A.; Beckmann, C.; Ruhé, H.G.; Marquand, A.F. From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neurosci. Biobehav. Rev.* **2024**, *158*, 105541. [CrossRef]
64. Depression and Older Adults. Available online: https://www.nia.nih.gov/health/depression-and-older-adults (accessed on 10 September 2023).
65. Colvonen, P.J.; DeYoung, P.N.; Bosompra, N.-O.; Owens, R.L. Limiting racial disparities and bias for wearable devices in health science research. *Sleep* **2020**, *43*, zsaa159. [CrossRef]
66. De Angel, V.; Lewis, S.; White, K.; Oetzmann, C.; Leightley, D.; Oprea, E.; Lavelle, G.; Matcham, F.; Pace, A.; Mohr, D.C.; et al. Digital health tools for the passive monitoring of depression: A systematic review of methods. *NPJ Digit. Med.* **2022**, *5*, 3. [CrossRef] [PubMed]
67. Powell, D.; Godfrey, A. Considerations for integrating wearables into the everyday healthcare practice. *npj Digit. Med.* **2023**, *6*, 70. [CrossRef]
68. Gerke, S.; Minssen, T.; Cohen, I.G. Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare. In *Artificial Intelligence in Healthcare*, 1st ed.; Bohr, A., Memarzadeh, K., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; ISBN 9780128184387. [CrossRef]
69. Silva-Trujillo, A.G.; González, M.J.G.; Pérez, L.P.R.; Villalba, L.J.G. Cybersecurity Analysis of Wearable Devices: Smartwatches Passive Attack. *Sensors* **2023**, *23*, 5438. [CrossRef]
70. Sui, A.; Sui, W.; Liu, S.; Rhodes, R. Ethical considerations for the use of consumer wearables in health research. *Digit. Health* **2023**, *9*, 20552076231153740. [CrossRef]
71. Ford, E.; Curlewis, K.; Wongkoblap, A.; Curcin, V. Public Opinions on Using Social Media Content to Identify Users With Depression and Target Mental Health Care Advertising: Mixed Methods Survey. *JMIR Ment. Health* **2019**, *6*, e12942. [CrossRef]
72. Essén, A.; Stern, A.D.; Haase, C.B.; Car, J.; Greaves, F.; Paparova, D.; Vandeput, S.; Wehrens, R.; Bates, D.W. Health app policy: International comparison of nine countries' approaches. *NPJ Digit. Med.* **2022**, *5*, 31. [CrossRef]
73. Smuck, M.; Odonkor, C.A.; Wilt, J.K.; Schmidt, N.; Swiernik, M.A. The emerging clinical role of wearables: Factors for successful implementation in healthcare. *NPJ Digit. Med.* **2021**, *4*, 45. [CrossRef] [PubMed]
74. Food and Drug Administration (FDA). *Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP Are Working Together*; FDA: Silver Spring, MD, USA, 2024.
75. Zafar, F.; Alam, L.F.; Vivas, R.R.; Wang, J.; Whei, S.J.; Mehmood, S.; Sadeghzadegan, A.; Lakkimsetti, M.; Nazir, Z. The Role of Artificial Intelligence in Identifying Depression and Anxiety: A Comprehensive Literature Review. *Cureus* **2024**, *16*, e56472. [CrossRef] [PubMed]
76. Blomqvist, S.; Högnäs, R.S.; Virtanen, M.; LaMontagne, A.D.; Hanson, L.L.M. Job loss and job instability during the COVID-19 pandemic and the risk of depression and anxiety among Swedish employees. *SSM Popul. Health* **2023**, *22*, 101424. [CrossRef]
77. Zisook, S.; Shear, K. Grief and bereavement: What psychiatrists need to know. *World Psychiatry* **2009**, *8*, 67–74. [CrossRef] [PubMed]