Aalto University
School of Science
Bachelor's Programme in Science and Technology

Machine learning in psychiatry: prediction and detection of mood disorders

Bachelor's Thesis

April 16, 2021

Lauri Anttila

Aalto University
School of Science
Bachelor's Programme in Science and Technology

ABSTRACT OF BACHELOR'S THESIS

Author:	Lauri Anttila
Title of thesis:	Machine learning in psychiatry: prediction and detection of mood disorders
Date:	April 16, 2021
Pages:	32
Major:	Computer Science and Engineering
Code:	SCI3027
Supervisor:	Professor Eero Hyvönen
Instructor:	D.Sc., Talayeh Aledavood (Department of Computer Science)

The large prevalence of smart devices has enabled the collection of multivariate and granular personal data. Recently, machine learning has been used to find markers of mental health disorders from passively collected data, which could lead to more accurate diagnoses and personalized treatment.

This thesis is a literature review of 19 studies, which use machine learning with passive data to predict or detect mood disorders. Common themes are extracted, categorized and analyzed.

The goal of the thesis is to find effective methods of machine learning in this context, explore the potential for digital phenotyping and discuss the important aspects of future research in this domain.

The literature was sought with following criteria: 1) Machine learning is used as the statistical tool, 2) Passive data is used for input, 3) The goal of the study is to predict or detect mood disorders, or their indicators. In the end, 19 valid papers from 2015 - 2021 were included in the review.

Multiple approaches were seen in these 19 studies. Even though similarities can be found within single analysis categories (e.g. model selection), deeper patterns between categories are not found. This is attributed to the fact that the research is still in its infancy and effective methodology has not yet emerged, thus the approaches are more experimental. Furthermore, no significant differences were found between the methods used in prediction and detection.

The most common models used were Support Vector Machine and Random Forest. Reasons for model selection were not elaborated in the studies, but the strengths of the models and potential reasoning for use are discussed in this thesis.

The collected data was high-dimensional, which lead to large feature vectors. A majority of the studies used feature reduction methods, but with largely different approaches.

Model validation consisted mainly of leave-one-out cross-validation and k-fold cross-validation, which are the traditional methods of validation.

The challenges reported in the studies were small sample size, high dimensionality, problems of data collection and generalizability of the findings.

Most significant sources of data were location, activity, sleep and circadian rhythm. Studies concluded that these sources could be used to predict or detect mood disorders with machine learning.

In conclusion, prediction and detection of mood disorders is possible with machine learning, but there are challenges to address before the research could be applied in a clinical application. Future research should focus on robust data collection methods, effective pre-processing and feature reduction methods as well as addressing the privacy aspects.

Keywords:	machine learning, mood disorder, digital phenotyping, review
Language:	English

Contents

1 Introduction

Mood disorders are a subset of mental health disorders, which affect the mood of an individual (e.g. depression). The 2017 Global Burden of Disease Study estimated mental disorders to affect 971 million people globally [19]. The current clinical practice of diagnosis and treatment of mood disorders consists of human expert assessments of patient behaviour and experiences. This is problematic as mood disorders can have overlapping symptoms [28] and psychiatric patient's retrospective recall can be biased [5]. In order to get more accurate diagnoses to improve patient outcomes, more accurate biological indicators, commonly referred to as "biomarkers", for psychiatric disorders are needed [6].

The recent searches for psychiatric biomarkers have focused on blood biomarkers and behavioral biomarkers [6]. Extracting behavioral information from patients is commonly done via clinical assessment from interviews [1], where the patient may have to recall events from hours, days, or even weeks, earlier which can be subjective and biased. Digital phenotypes, i.e. digital behavioral patterns or markers that can be correlated with certain disorders or symptoms, could provide more reliable behavioral information and more accurate assessments.

Most people today have, and carry with them everywhere, a device that generates reliable and temporal information: a smartphone. The daily use of smartphones generates a behavioral fingerprint which could potentially be used to develop digital phenotypes and markers to diagnose psychiatric illnesses [25]. Onnela and Rauch [25, p. 1691] define digital phenotyping as "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices", whereas Insel [17, p. 1215] defines it as "measuring behaviour from smartphone sensors, keyboard interactions and various features of voice and speech".

Ben-Zeev et al. [4] demonstrated that smartphone data allows passive and unobtrusive collection of data that can be used for indicators of mental health. Smartphones generate complex multivariate data with potential to extract intricate behavioral patterns [25]. However, such high-dimensional data is difficult to summarize, and the models for summarization must be carefully crafted to leverage the full potential of the data [25].

Machine learning is a statistical tool that can be used to automatically detect meaningful patterns in high-dimensional data [30]. In recent years, studies have emerged on applying machine learning to mental health [31]. Due to the learning nature of machine learning methods, larger data sets are suitable, even required. Passive data collection can be used to gather granular data with large sample sizes [25]. This would imply machine learning could be a valuable statistical tool in predicting and classifying mood disorders and finding potential digital phenotypes of mood disorders.

This thesis reviews machine learning studies focused on predicting and classifying mood disorders from passive digital data. 107 amount of papers were queried and screened to find 19 relevant papers. Out of these papers, common themes and patterns in methods are categorized, analyzed and discussed.

The goals of the thesis consist of finding effective practices of machine learning methods in the context of mood disorders, finding potential digital phenotypes of mood disorders, and discussing future direction of research in this domain.

2 Background

This chapter introduces the necessary background information for the thesis. Mood disorders and the current state of mental health treatment, data collection and digital phenotyping as well as machine learning in mental health are covered.

2.1 Mood disorders and the current state of mental health

Mood disorders are a subset of mental health disorders, which affect the mood of an individual. They include disorders such as major depressive disorder, bipolar disorder and dysthymia (persistent depressive disorder) [23]. The 2017 Global Burden of Disease Study estimated mental disorders to affect 971 million people globally, and from these, 264 million were credited to depressive disorders [19]. Even though these disorders are rather prevalent, people may not seek help due to public or personalized stigma [11]. Furthermore, the current methods of treatment and diagnosis have problems to be addressed.

Mental heath treatment and diagnosis are traditionally based on human expert assessment. Evaluations are usually conducted via interviews and questionnaires such as the Patient Health Questionnaire-9 (PHQ-9), and carried out in a clinical setting, a hospital or a treatment center, instead of the individual's natural environment. These methods of assessment are often met with various biases, as the recall of a patient may not be accurate and the motivation of the individual, or lack thereof, could direct their answers to minimize their problems or exacerbate them [4]. This can cause misdiagnosis and thus incorrect treatment. For example, Singh and Rajput [32] found that 69% of bipolar disorder cases are initially wrongly diagnosed, and receiving the correct diagnosis after a misdiagnosis can take on average 5-7 years.

2.2 Data collection and digital phenotyping

The abundance of smartphones in the modern world opens up new avenues for data collection [17]. Smartphones have various sensors which can be polled rapidly, generating highly granular and temporal data from e.g. GPS, accelerometer, phone calls and texts, on-screen-time and Bluetooth [25].

Digital phenotyping is a relatively new phenomenon [17], where the digital behaviour of an individual is measured from passively collected data. Digital phenotyping could enable finding correlating behavioral patterns to various mental health disorders [4].

In the future, digital phenotyping and machine learning could be used for effective monitoring of patients after treatment to detect deterioration or improvement of mental health. Furthermore, if the efficacy of machine learning models in this domain is shown to be highly performant and accurate, diagnosis could evolve from clinical human assessment into a partially or fully automated process.

2.3 Machine learning in mental health

Machine learning is a suitable tool for individual clinical predictions, which could show benefits in early diagnosis, selection between drug treatments, treatment adjustment and tailored psychiatric care [7]. Compared to more classical statistical tools, which offer more retrospective results with in-sample estimates, machine learning success is evaluated by the prediction power in new individuals, out-of-sample estimates, and thus could be a candidate for future personalized psychiatry [7].

Machine learning is rarely used clinically in mental health, and a Scopus search with the query `("machine learning" and "mental health")` shows that the research in this area is still in the beginning stages, breaking 10 publications per year in 2012 and 100 publications per year in 2018. There are multiple aspects to be researched and validated, but machine learning seems to be a prospect of enabling various new approaches in the field of mental health.

3 Methodology

This thesis reviews studies that apply machine learning to passive digital data with the purpose of predicting or detecting mood disorders, mood disorder indicators or mood states. The methods of querying, screening, extracting and discussing the material are outlined in this chapter. All themes that are extracted are covered and explained in "Extraction" subsection.

3.1 Querying

Relevant studies are sought with following keywords "computational", "machine learning", "deep learning", "neural network", "predict*", "detect*", "classif*", "mood disorder", "MDD", "depression", "major depressive disorder", "BD", "bipolar", "passive data", "smartphone", "wearable", "phenotyping", "digital phenotyping".

The following fields were used in search queries: title, keywords, abstract.

The following query was constructed and input to Scopus and Google Scholar.

> ("machine learning" or "deep learning" or "neural network" or "computational") and ("predict*" or "detect*" or "classif*") and ("mood disorder" or "MDD" or "depression" or "major depressive disorder" or "BD" or "bipolar") and ("passive data" or "smartphone" or "wearable")

Google Scholar did not accept such long query, thus ("passive data" or "smartphone" or "wearable") phrase was removed from it. However, Scholar only returned a single result.

From these searches, a total of 107 papers were identified as potentially relevant.

3.2 Screening

Identified papers were then screened to satisfy the following parameters: 1) The study used passive data from wearable devices or sensors, 2) The study used machine learning as their analysis tool, 3) The study tried to predict, detect or classify mood disorder(s), indicators, or mood states.

Expanding on the last screening parameter: Papers are included if they predict, detect or classify mood disorders, mood states related to a mood disorder or indicators of a mood disorder, e.g. depressive state.

After screening, 15 valid studies were identified. The citations in these studies were also screened for more studies. This second screening yielded 4 additional selected papers. The second screening process was not exhaustive and could have resulted in more valid studies, but had to be halted due to time constraints.

After the screening process, a total of 19 valid studies for this review were found.

3.3 Extraction

The screened papers were analyzed and common themes and patterns in the used methods were extracted from them. The following common themes or steps in the process were identified in the machine learning methodologies of the selected papers.

Data collection includes the device(s) and the sensors or sources used. E.g. GPS from a smartphone and actigraph from a wrist sensor.

Ground truth is the act of labeling patients or data. Supervised machine learning models use these labels for training.

Pre-processing is split into two categories: **data quality** and **features**. Data quality covers the aspects of managing data quality in the data set: handling of missing data, invalid data, noisy data, etc. In other words, how data, that is not optimal for machine learning model training, is handled.

Feature pre-processing covers how features are extracted from the raw data. Refining and converting raw data into useful metrics is covered under pre-processing. This includes derived numerical features e.g. calculating location variance from raw longitudinal and latitudinal data, or creating more clinically relevant and interpretable semantic features e.g. "time patient stayed home".

Feature reduction covers the act of ranking and optimizing the feature space via algorithmic or manual approaches.

Model selection covers what machine learning models the studies have used. If multiple models were used, the flow of data between them is covered, i.e. whether the models are in series, parallel or separate.

Model optimization covers methods, tools and algorithms for optimizing model performance or hyper-parameters.

Model evaluation and validation is the act of measuring the performance of the model. The section covers validation techniques and evaluation metrics used.

Concerns and limitations cover aspects of the studies the authors have mentioned to be limiting or to be concerning. These can be factors of error in data collection, problems with recruitment, biases, problem sources in modeling,

4 Results

In this chapter the following identified common themes are covered: 1) ground truth, 2) data collection, 3) pre-processing: quality, 4) pre-processing: features, 4) feature reduction, 5) model selection, 6) model optimization, 7) model validation, 8) challenges.

None of the found papers attempted to classify between mood disorders, but rather if the subject has or does not have a certain mood disorder or mood state.

Majority (13/19) [27; 18; 8; 24; 22; 29; 12; 26; 14; 21; 15; 33; 34] of the papers attempted to detect mood disorders or states in the present. The rest (6/19) [9; 35; 10; 16; 13; 20] attempted to predict mood disorders.

I looked at the differences in the used methods between prediction and classification but no noticeable patterns were found, thus it is not differentiated in the following sections. From supervised machine learning point of view, prediction and detection are the same problem: correlate an input vector with outcome labels.

4.1 Ground truth

Most used (9/19) [35; 8; 22; 29; 12; 13; 21; 15; 34] source of ground truth was PHQ-9. It was either implemented in a clinical setting before and after the study, or submitted via a smartphone application. All studies used some sort of active questionnaire for establishing ground truth. Some (5/19) [10; 13; 14; 20; 33] accompanied the questionnaire with a separate clinical assessment for more reliable ground truth results.

Validated and tested questionnaires used, aside from PHQ-9, were Hamilton Depression Rating Scale (HDRS) [27; 20; 16; 33], Beck Depression Inventory-II (BDI-II) [9], Quick Inventory of Depressive Symptomatology–Self-report (QIDS-SR) [26], Montgomery-Asberg Depression Rating Scale (MADRS) [14]. Some researches created their own questionnaires [10; 18; 24; 20].

4.2 Data collection

Data source	Used in
Location (GPS or WiFi)	[9; 27; 18; 8; 24; 22; 29; 12; 16; 13; 26; 15; 34]
Activity (e.g. accelerometer, steps)	[9; 10; 24; 22; 29; 12; 16; 13; 14; 20; 33; 34]
Call and / or SMS logs	[9; 27; 18; 24; 16; 34]
Phone usage (e.g. apps, screenlock)	[27; 35; 24; 12; 16; 34]
Light exposure (ambient, UV)	[10; 24; 12; 20; 33]
Heartrate	[10; 18; 24; 33; 27]
Sleep	[9; 10; 16]
Weather	[27; 18]
Skin temperature	[33; 27]
Electrodermal activity	[16; 27]
Bluetooth	[9]
Internet traffic	[35]
Conversation audio	[12]
Typing patterns	[21]
Calendar events	[34]

Table 1: Data sources used in the final 19 reviewed studies

Almost all papers (16/19) [9; 27; 35; 10; 18; 8; 24; 22; 29; 12; 16; 13; 26; 21; 15; 34] used smartphone for data collection, which is expected as passive data from wearables was a requirement for inclusion in the thesis and smartphones are the most prevalent wearable device. Out of these, four papers [9; 27; 24; 16] also combine this with some other wearable device, usually wrist-based, such as Fitbit or a smartwatch. Three [14; 20; 33] studies used only a wrist-based device for data collection.

Majority of the studies (12/19) [9; 27; 10; 18; 24; 29; 12; 16; 13; 20; 33; 34] used a combination of sensors or data e.g. GPS or location and outgoing calls. Most common (13/19) [9; 27; 18; 8; 24; 22; 29; 12; 16; 13; 26; 15; 34] source of data was location (via GPS or WiFi), followed by activity (12/19) [9; 10; 24; 22; 29; 12; 16; 13; 14; 20; 33; 34].

4.3 Pre-processing: quality

Quality control is an important aspect to consider in machine learning, as invalid data can skew the learning results. 15 out of 19 papers [9; 27; 35; 8; 24; 22; 12; 16; 13; 26; 21; 20; 33; 29] had some sort of data quality control. From the rest, two studies [14; 15] had used a public data set ("The Depresjon Dataset" used by Frogner et al. [14], "StudentLife Dataset" used by Gerych et al. [15]), which are most likely pre-processed already. Farhan et al. [13] also used the StudentLife dataset, but included pre-processing in their study. Wahle et al. [34] did not report the use of quality control related pre-processing.

Missing data was present in all studies. Samples with missing data was either excluded [9; 27; 35; 10; 24; 13; 20; 33] or imputed [9; 8; 16; 26]. Canzian and Musolesi [8] replaced missing data with the mean of other samples, if the amount of missing data was small. Ghandeharioun et al. [16] utilized self-reports to impute missing data in clinical HDRS measures.

Reasons for missing data were speculated to be due to technical difficulties, forgetting to carry their phone or wearables and forgetting to fill in ground truth questionnaires. Invalidating results which seemed infeasible to be real was common as well, e.g. questionnaire answered in under a second.

Certain data sources may include noise, which should be filtered out for the best performance. Yue et al. [35] used low-pass and high-pass filters to reduce noise in internet traffic data and isolate meaningful data. Farhan et al. [12] used Wavelet transform to denoise their conversation data to obtain smoother time-series data retaining peaks and trends.

4.4 Pre-processing: features

This subsection outlines the approaches of extracting features from the raw data. Individual features were not always mentioned, and if they were, they were in the hundreds, if not thousands, which is why they are not reported here.

The extraction methods of features can be divided into two categories: features that were derived from the data via algorithmic means ("derived features") e.g. calculating variance of samples, and features which were calculated using raw values to infer more clinically relevant and interpretable variables ("semantic features") e.g. how long a patient stayed home.

18 studies [9; 35; 10; 16; 13; 20; 18; 27; 8; 24; 22; 29; 12; 26; 21; 15; 33; 34] out of 19 used derived features for their model. The same 18 studies [9; 35; 10; 16; 13; 20; 18; 27; 8; 24; 22; 29; 12; 26; 21; 15; 33; 34] also used semantic features. The one study [14] that did not derive any features used a pre-processed data set and their model of choice was a neural network, which can be trained with raw data without explicitly creating features.

In 7 studies, features were grouped and analyzed in temporal segments: weekdays / weekend [9; 26], 24 hours [9; 16; 18; 27], 4x6 hours [9; 27; 16; 35], daytime / nighttime [10].

Clustering was used to identify common stationary locations from location data. DBSCAN [9; 22; 13] and k-means [26; 29; 15; 12] were the preferred methods of clustering location data.

4.5 Feature reduction

Reduction method	Used in
Logistic Regression	[9; 20]
Recursive Feature Elimination	[35; 27]
Wrapper Methods	[22; 26]
Classification prior the final model	[9; 24]
Joint Mutual Information	[35]
Principal Component Analysis	[16]
Boruta Algorithm	[27]
Multiview Bi-Clustering Method	[12]
Select-k-best with ANOVA	[21]
Autoencoder	[15]
Gini Importance	[33]

Table 2: Used feature reduction methods in the final 19 reviewed studies

Smartphones and wearables can output high-dimensional data, which would result in a complex model and would require significantly more processing power. Feature ranking algorithms were used to reduce the dimensionality while retaining most if not all of the relevant features and information.

More than half (11/19) [9; 27; 35; 22; 12; 16; 26; 21; 15; 20; 33] of the studies used some sort feature ranking or reduction method. There was very little overlap between methods used, at most two studies used similar approaches, aside from simple correlation analysis.

4.6 Model selection

Model	Used in
Support Vector Machine	[35; 8; 24; 22; 12; 13; 21; 34]
Random Forest	[27; 10; 18; 24; 16; 21; 20; 34]
Logistic Regression	[9; 29; 20]
AdaBoost	[9; 16; 27]
Extreme Gradient Boost	[9; 16; 27]
Gradient Boost	[9; 21]
Linear Regression	[26; 29]
Neural Network	[14; 22]
Quadratic Discriminant Analysis Classifier	[26]
One Class Support Vector Machine	[15]
K-Nearest-Neighbors	[16]
Gaussian Boost	[16]
Lasso Regression	[16]
Ridge Regression	[16]
ElasticNet Regression	[16]
Decision Tree	[20]
Boosted Tree	[20]

Table 3: Models identified in the final 19 reviewed studies

There is large variety in models used aside from two models: Support Vector Machine (SVM) [35; 8; 24; 22; 12; 13; 21; 34] and Random Forest (RF) [27; 10; 18; 24; 16; 21; 20; 34].

12 out of 19 papers [9; 27; 35; 18; 24; 22; 29; 16; 26; 21; 20; 34] used multiple models. Models were used in series and in parallel to create more complex model pipelines, and also separately to compare their performance. Ghandeharioun et al. [16] had the largest pipeline, combining lasso, ridge, and elasticNet regressions, AdaBoost and Gaussian Boost, Random Forest and finally using k-nearest-neighbor (kNN) clustering to smooth the feature space and choosing suitable models based on the set of points from kNN.

Jacobson and Chung [18] used extreme gradient boosting model (XGBoost) in series with Random Forest (RF), first modeling intraindividual variability in data with XGBoost which was then used as secondary features in the prediction modeling with RF. Narziev et al. [24] used SVM and RF in parallel to first create symptom classifications from the original features and piping the output to SVM and RF again for the prediction. Chikersal et al. [9] had a similar approach, classifying feature sets of individual categories with Gradient Boost and Logistic Regression and then piping the output to AdaBoost.

Masud et al. [22] compared the performance of three models separately: SVM, kNN and neural network (NN). Results showed that the models had their own strengths and weaknesses, e.g. NN had best negative predictions, but SVM performed generally well in every category.

4.7 Model parameter optimization

Model parameter optimization methods were reported in 14 [9; 35; 16; 13; 18; 27; 8; 22; 29; 12; 14; 21; 33; 34] out of 19 studies.

Most common method, mentioned in four [9; 18; 21; 33] out of 19 studies, was grid-search. Five [35; 13; 8; 14; 12] other studies did not explicitly mention the use of grid-search, but their described process mirrored that of grid-search, where parameter values are iterated over.

Ghandeharioun et al. [16] used Theil-sen estimator, random sample consensus (RANSAC) and Huber algorithms for fitting classifier parameters in their multi-model pipeline.

Saeb et al. [29] used least squares optimization and Elastic-Net regularization.

Masud et al. [22] used sequential minimal optimization for SVM model parameter optimization. It solves a quadratic problem common with SVM. However, other papers did not utilize this method of optimization even though multiple papers used SVM.

Farhan et al. [12] used PCA for clustering and tuning of optimal parameters.

Frogner et al. used Adam optimizer for parameter fitting and also iterated over values to find optimal segment length.

Wahle et al. [34] used Nelder-Mead simplex optimization method to tune SVM parameters.

Pedrelli et al. [27] mentioned the use of 10-fold cross-validation for parameter fitting, but did not further specify how the parameters were tuned.

4.8 Model validation

All of the studies used some form of cross-validation, also referred to as out-of-sample validation, where model is validated on unseen data. This form of validation demonstrates the predictiveness of machine learning and the ability to extract generalizing patterns from the data, rather than making retrospective findings in-sample.

Two forms of validation methods were mainly seen used with training model parameters: leave-one-out cross-validation [9; 35; 8; 22; 29; 26; 14; 21; 15; 34] and k-fold cross-validation [16; 13; 12; 26; 33]. Other methods seen were rolling window through the data set [18; 10] and iterations of random data splits for test and training [20].

Split	Used in
65%/35%	[20]
70%/30%	[24; 12]
80%/20%	[27]
90%/10%	[16; 12]

Table 4: Observed training / testing splits in the final 19 reviewed studies

4.9 Challenges

A low sample size was mentioned in 9 studies [9; 35; 24; 29; 12; 26; 21; 20; 33]. 11/19 of the data sets consisted of under 50 people and the total mean was 51 people. Machine learning performance, especially neural networks, suffers greatly with very low sample sizes, as the models may not have enough data to recognize patterns.

Studies [35; 8; 29; 22; 24] also reported the concern of applicability to the general population. Most of the studies used a specific group of people, e.g. college students, and the results may not be applicable to different or larger population groups.

Overfitting was also mentioned in three studies [9; 27; 24]. A common cause for overfitting is a low sample size, especially if coupled with a high degree polynomial function and large feature space, which are both abundant in these studies.

5 Discussion

This chapter covers what common themes and effective practices were observed, what kind of challenges and limitations exist and what should be considered when conducting further research in this domain. The emphasis is set on data collection, feature engineering and reduction, and model selection as I believe these aspects have the largest impact on the outcome.

The studies have a wide variety of approaches in their methodology. While there are similarities in individual aspects of the methods (e.g. model selection), patterns were not observed between different aspects (e.g. data source and model selection) of the methods across studies. The pool of reviewed studies is relatively small, but this observation implies that the field is still in an experimental phase, where no validated and proven approaches are defined and researchers are not limited to any methodology and are more free to experiment with different methodology.

Expanding on this, differences in methods could not be pinpointed to the goal of the study. In other words, prediction and detection studies had generally similar approaches in individual aspects of the methods. This can be attributed to the fact that supervised machine learning does not differentiate between present and future data. It simply tries to correlate input vectors with output labels. This can be observed in Frogner et al. [14] where they classify depressed patients and predict MADRS-questionnaire scores with the same model architecture. Similarly, the most common machine learning models in these studies (SVM and RF) were seen in both prediction and detection categories.

5.1 Common themes and effective practices

5.1.1 Data collection

Most common sources of data in these studies were location data from GPS and WiFi sensors and activity data from accelerometer. Location and activity patterns are generally thought to be linked with mood disorders, e.g. depressed people may stay home more, thus it is expected to see the data being utilized for machine learning. Multiple studies [9; 10; 22; 29; 12; 16; 26] concluded that location data and activity can be used for detecting mood disorder characteristics. Activity was measured with both smartphones as well as smartwatches. Smartwatches may better capture activity data as I suspect accelerometer data from a wrist displays larger movement differences than the hips, where a phone usually is when carrying it in a pocket, and thus it would be easier to classify activities from wrist sensor data. Additionally, the location of the carried phone may vary (e.g. different pockets, purse, bag, jacket) whereas a wrist worn sensor is always on the wrist, although it can still vary between the left or right wrist. Some people may place their phone down or leave it home when engaging activities, whereas a smartwatch is more unobtrusive in that context and could be left on.

In fact, reliability of data from a smartphone was a limitation that was mentioned in multiple studies, as people may not carry their phone with them and thus the obtained data may not be fully representative of the individual's actions. Missing data was a problem that every study had to account for, and the approaches were to either remove the samples or intervals of missing data, or to impute them e.g. with mean of other values.

Irregular circadian rhythm and poor sleep quality are often associated with depression. Data from both categories were seen used and studies concluded that features derived from these categories could be indicative of mood disorders. Circadian rhythm was measured by finding daily repeating location and activity patterns. Sleep related data is difficult to capture with a smartphone as biological differences of good and poor sleep quality cannot be measured. The studies that measured sleep quality used a wrist worn sensor, which may not be suitable for widespread use outside of research, as people may not be inclined to sleep wearing a sensor. However, it is a step in the right direction, as previously sleep related research was done mainly in a clinical setting with polysomnography [3].

5.1.2 Pre-processing and feature reduction

Pre-processing and feature engineering are highly important aspects in machine learning. Machine learning models benefit from data that has minimal noise or erroneous and invalid data and properly extracted features, as the risk of learning patterns in the noise and invalid data decreases. Reducing the dimensionality of the input space while retaining the relevant information lessens the computing power needed and reduces the risk of overfitting, especially if using a low sample size.

Especially with the prevalent issue of missing data, methods of imputation and utilization of sparse intervals should be investigated, to enable wider use of the collected data.

When looking at highest performing studies, a trend of low number of features can be seen. Four highest performing studies [22; 24; 15; 20], based on accuracy and F1-score, all utilize some sort of feature ranking or reduction method. Feature reduction approaches had the most variability among the extracted themes, which leads me to believe that it is an important aspect, but multiple approaches are valid.

A lower number of features results in less risk of overfitting, which is a major issue in machine learning, especially when training with a small sample size. Most of the studies reported a small sample size as an limitation, and thus it is expected to see high performing studies utilize some feature engineering methods.

5.1.3 Model selection

Support Vector Machines (SVM) and Random Forest (RF) were the most popular choices of machine learning models among the papers. Aside from one study [22], none of the papers elaborated why they had chosen specific models. Comparisons between models were seen, but the initial selections were not explained. No correlation was seen between used data sources and model selection. However, there are advantages to both SVM and RF which could explain their use in this context.

Firstly, SVM and RF are both well-known and documented models, initially developed in the 90's, and they have been successfully applied in a variety of different problems as they can be used for classification as well as regression. Both SVM and RF are considered to be effective with high-dimensional data, which is a common characteristic in the studies. Masud et al. [22] explained their choice of SVM by its ability to handle overfitting with high-dimensional data.

SVM models can handle unbalanced data sets, which is suitable for mood disorder analysis as only a small part of a population suffer from mood disorders and thus can cause unbalanced data sets if the included people are randomly sampled. SVM is also suitable for sparse data sets, which is a significant advantage given that all of the studies had problems with missing data.

RF is a computationally cheap machine learning model and doesn't require a high-end computer to train, making it suitable for fast prototyping or experimenting. RF uses subsets of the features in the random decision trees it creates, thus performing implicit feature reduction and can handle input features that are intercorrelated. This causes Random Forest to be less prone to overfitting and may achieve more reliable performance than other classification methods.

Various boosting models were also seen (AdaBoost, Gaussian Boost, [Extreme] Gradient Boost). Boosting is a method where multiple weak classifiers are combined to create a stronger classifier. Boosting can be used to create non-linear classification models. This can be beneficial with passive wearable data as mood disorder related patterns can be intricate and most likely cannot be represented by a linear separation of the data points.

There were two studies [14; 22] that used a neural network. The low amount of neural network usage in the studies is most likely due to the attributes of the data in mental health: low sample size and high dimensionality. Neural networks, especially networks with high amount of layers, are susceptible to over-fitting. Neural networks also typically require a lot of resources to train.

Majority of the studies used only a single model in their pipeline, but I want to highlight two studies with a different approach.

Chikersal et al. [9] and Narziev et al. [24] had both had two models in series with the intent to first use a model to acquire a score from certain features or categories and then use said scores for the final classifier.

Chikersal et al. [9] split their features into seven categories and used Linear Regression and Gradient Boosting to acquire a score for each feature-set which were then fed into an AdaBoost classifier for the final prediction. They achieved accuracy of 85.7% with this method.

Narziev et al. [24] had a similar approach where they had 5 categories of features and used Support Vector Machine to calculate level of depression from each feature-set and fed the scores into a Random Forest model for the final prediction. They achieved a very high accuracy of 96%. However, they mentioned the concern of small sample size and personalized model which could cause overfitting and thus potentially skew the results.

These approaches are similar to the workings of a boosting model, where weaker classifiers are combined. This approach of aggregating models to be used for final classification could enable the models to more easily capture complex patterns and relationships, but would require further research to validate.

5.2 Privacy

Passive data collection is a valuable tool as it can collect multidimensional, granular data which can be used for more precise diagnosis, treatment and insights. However, it comes with a concern of privacy. Data in these studies is collected around the clock, from various sensors including location and phone usage. As the amount of data collected grows larger, so does the risk of identification from the data. The medical field is subject to strict regulations regarding patient data in most of the world, and thus the systems need to be safety-critical to comply with the laws and regulations and to protect the patients.

Studies with a smartphone as the main source of data used an application which recorded the data and sent it forward for processing. These applications often encrypted the data in the phone, before sending it to the researchers. A robust system for preserving privacy is essential in these applications to achieve trust between the patient and the reseachers.

Insel [17] discusses the aspects of privacy and mentions concerns related to data collection: who owns the data? how will the data be used, i.e. will it be used for identification and marketing or improve healthcare?

Insel [17] also mentions that currently no party exists that would set standards for value or ensuring trust. A set of quality standards should be created to handle efficacy, engagement and privacy concerns [17].

5.3 Future directions

As the research of machine learning in this domain is still relatively young and digital phenotyping is a new concept, future research should be done to identify more concretely the sources of data and features that can be identified. If the important sources of data can be identified, focus can be moved to creating more effective machine learning pipelines. In my thesis, location, activity, sleep and circadian rhythm were seen to correlate with mood disorders. These findings should be validated further, and deeper analysis should be done to identify the features derived from these categories that yield the best results.

Most studies were done on relatively small sets of participants. 11/19 of the studies had less than 50 people. This results in small data sets, which in turn results in uncertain performance and variance. In addition, most data sets were of similar population, meaning that the results may not be generalized. Larger studies should be conducted with data sets of different attributes (for example age, sex, cultural differences, educational background) to validate findings of smaller studies and find more generally applicable phenotypes.

Standard methodology for data collection and pre-processing should be created which would enable creating larger data sets from combining data of different studies and allow for more collaboration. Current limitation of machine learning in this domain is the lack of larger data sets to allow machine learning to utilize its full potential. Pre-processing pipelines should be refined and standardized to achieve uniformly processed quality data. A study [7] has noticed that when combining datasets, model performance could worsen, as different datasets can have heterogenous charasteristics (e.g. different noise and other patterns).

Aledavood et al. [2] conducted research regarding data collection in the domain of mental health with the objective of identifying key features for a digital data collection platform.

Their work concluded that important aspects include: 1) flexibility of access control, 2) flexibility of data sources, 3) first-order privacy protection [2].

More research of data collection methodology should be conducted to achieve robust methodology for data collection, as then the focus could be shifted more into optimizing the pre-processing of the data and machine learning model pipelines.

Solutions for privacy concerns should be researched to increase safety of the subject data and cultivate trust. Robust privacy protection is a necessity, especially when conducting studies of larger scale.

6 Conclusion

Digital phenotyping could enable the field of psychiatry to improve mental health treatment with more accurate diagnoses, more personalized treatment and more proactive action. Machine learning is a valuable tool in this improvement and is showing promising results in terms of mood disorder prediction and detection. The research in this domain is still in its infancy, and more research needs to be done before these methods can be used in a clinical application.

For future research, three things should be considered: data collection, pre-processing and feature reduction, and model selection.

More robust approaches to data collection should be identified and issues regarding reliability and missingness of the data should be improved. In general, a standardized methodology for data collection would enable easier collaboration, creation of larger data sets and improved privacy.

Pre-processing for passive data should be refined to provide quality input data for modeling, as wearable data was shown to have problems of missing data, technical problems resulting in invalid data, noise in the data. In addition, feature reduction techniques should be investigated further as wearable data tends to be high-dimensional and can lead to overfitting the learning models.

Model selection in the selected studies is not elaborated, and seems to be more out of convenience rather than the result of deeper analysis. Studies should be conducted to find optimal models and possibly combinations of models, as well as concrete reasoning for the use of said models, to achieve more robust pipelines and better performance in terms of accuracy.

References

[1] Ahmed Aboraya, Henry Nasrallah, Srinivas Muvvala, Ahmed El-Missiry, Hader Mansour, Cheryl Hill, Daniel Elswick and Elizabeth C Price. The standard for clinicians' interview in psychiatry (scip): A clinician-administered tool with categorical, dimensional, and numeric output-conceptual development, design, and description of the scip. *Innovations in clinical neuroscience*, 13:31–77, 5 2016. ISSN 2158-8333. URL `http://www.ncbi.nlm.nih.gov/pubmed/27800284http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5077257`.

[2] Talayeh Aledavood, Ana Maria Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä and Richard K Darst. Data collection for mental health studies through digital platforms: Requirements and design of a prototype. *JMIR Research Protocols*, 6:e110, 6 2017. ISSN 1929-0748. doi: 10.2196/resprot.6919. URL `https://www.researchprotocols.org/2017/6/e110`.

[3] Talayeh Aledavood, John Torous, Ana Maria Triana Hoyos, John A Naslund, Jukka-Pekka Onnela and Matcheri Keshavan. Smartphone-based tracking of sleep in depression, anxiety, and psychotic disorders. *Current psychiatry reports*, 21(7):1–9, 2019.

[4] Dror Ben-Zeev, Emily A. Scherer, Rui Wang, Haiyi Xie and Andrew T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38:218–226, 9 2015. ISSN 1559-3126. doi: 10.1037/prj0000130. URL `http://doi.apa.org/getdoi.cfm?doi=10.1037/prj0000130`.

[5] Edwin DE Beurs, Alfred Lange and Richard Van Dyck. Self-monitoring of panic attacks and retrospective estimates of panic: discordant findings. *Behav. Res. Ther.*, 30:411–413, 1992.

[6] Patricia Boksa. A way forward for research on biomarkers for psychiatric disorders. *J Psychiatry Neurosci*, 38, 2013. doi: 10.1503/jpn.130018. Biomarker usage.

[7] Danilo Bzdok and Andreas Meyer-Lindenberg. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3:223–230, 3 2018. ISSN 24519030. doi: 10.1016/j.bpsc.2017.11.007. https://www-sciencedirect-com.libproxy.aalto.fi/science/article/pii/S2451902217302069?via

[8] Luca Canzian and Mirco Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis.

2015. doi: 10.1145/2750858.2805845. URL http://dx.doi.org/10.1145/2750858.2805845.

[9] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K. Villalba, Janine M. Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel and Anind K. Dey. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Trans. Comput.-Hum. Interact.*, 28(1), January 2021. ISSN 1073-0516. doi: 10.1145/3422821. URL https://doi.org/10.1145/3422821.

[10] Chul-Hyun Cho, Taek Lee, Min-Gwan Kim, Hoh Peter In, Leen Kim and Heon-Jeong Lee. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. *J Med Internet Res*, 21(4):e11029, Apr 2019. ISSN 1438-8871. doi: 10.2196/11029. URL http://www.jmir.org/2019/4/e11029/.

[11] Daniel Eisenberg, Marilyn F. Downs, Ezra Golberstein and Kara Zivin. Stigma and help seeking for mental health among college students. *Medical Care Research and Review*, 66:522–541, 10 2009. ISSN 10775587. doi: 10.1177/1077558709335173. URL http://journals.sagepub.com/doi/10.1177/1077558709335173.

[12] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang and A. Bamis. Multi-view bi-clustering to identify smartphone sensing features indicative of depression. *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 264–273, 2016. doi: 10.1109/CHASE.2016.27.

[13] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis and B. Wang. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health (WH)*, pages 1–8, 2016. doi: 10.1109/WH.2016.7764553.

[14] Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen and Michael Alexander Riegler. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. *Proceedings of the 4th International Workshop on Multimedia for Personal Health amp; Health Care*, HealthMedia '19, page 9–15, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369145. doi: 10.1145/3347444.3356238. URL https://doi.org/10.1145/3347444.3356238.

[15] W. Gerych, E. Agu and E. Rundensteiner. Classifying depression in imbalanced datasets using an autoencoder- based anomaly detection approach. *2019 IEEE 13th*

International Conference on Semantic Computing (ICSC), pages 124–127, 2019. doi: 10.1109/ICOSC.2019.8665535.

[16] A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, C. Dale, D. Sontag and R. Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 325–332, 2017. doi: 10.1109/ACII.2017.8273620.

[17] Thomas R Insel. Digital phenotyping-technology for a new science of behavior published by jama, october 2017. 2017. doi: 10.1001/jama.2017.11295.

[18] Nicholas C. Jacobson and Yeon Joo Chung. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*, 20(12), 2020. ISSN 1424-8220. doi: 10.3390/s20123572. URL `https://www.mdpi.com/1424-8220/20/12/3572`.

[19] Spencer L. James, Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, Ibrahim Abdollahpour, Rizwan Suliankatchi Abdulkader, Zegeye Abebe, Semaw F. Abera, Olifan Zewdie Abil, Haftom Niguse Abraha, Laith Jamal Abu-Raddad, Niveen M.E. Abu-Rmeileh, Manfred Mario Kokou Accrombessi, Dilaram Acharya, Pawan Acharya, Ilana N. Ackerman, Abdu A. Adamu, Oladimeji M. Adebayo, Victor Adekanmbi, Olatunji O. Adetokunboh, Mina G. Adib, Jose C. Adsuar, Kossivi Agbelenko Afanvi, Mohsen Afarideh, Ashkan Afshin, Gina Agarwal, Kareha M. Agesa, Rakesh Aggarwal, Sargis Aghasi Aghayan, Sutapa Agrawal, Alireza Ahmadi, Mehdi Ahmadi, Hamid Ahmadieh, Muktar Beshir Ahmed, Amani Nidhal Aichour, Ibtihel Aichour, Miloud Taki Eddine Aichour, Tomi Akinyemiju, Nadia Akseer, Ziyad Al-Aly, Ayman Al-Eyadhy, Hesham M. Al-Mekhlafi, Rajaa M. Al-Raddadi, Fares Alahdab, Khurshid Alam, Tahiya Alam, Alaa Alashi, Seyed Moayed Alavian, Kefyalew Addis Alene, Mehran Alijanzadeh, Reza Alizadeh-Navaei, Syed Mohamed Aljunid, Ala'a Alkerwi, François Alla, Peter Allebeck, Mohamed M.L. Alouani, Khalid Altirkawi, Nelson Alvis-Guzman, Azmeraw T. Amare, Leopold N. Aminde, Walid Ammar, Yaw Ampem Amoako, Nahla Hamed Anber, Catalina Liliana Andrei, Sofia Androudi, Megbaru Debalkie Animut, Mina Anjomshoa, Mustafa Geleto Ansha, Carl Abelardo T. Antonio, Palwasha Anwari, Jalal Arabloo, Antonio Arauz, Olatunde Aremu, Filippo Ariani, Bahram Armoon, Johan Ärnlöv, Amit Arora, Al Artaman, Krishna K. Aryal, Hamid Asayesh, Rana Jawad Asghar, Zerihun Ataro, Sachin R. Atre, Marcel Ausloos, Leticia Avila-Burgos, Euripide F.G.A. Avokpaho, Ashish Awasthi, Beatriz Paulina Ayala Quintanilla, Rakesh Ayer, Peter S. Azzopardi, Arefeh

Babazadeh, Hamid Badali, Alaa Badawi, Ayele Geleto Bali, Katherine E. Ballesteros, Shoshana H. Ballew, Maciej Banach, Joseph Adel Mattar Banoub, Amrit Banstola, Aleksandra Barac, Miguel A. Barboza, Suzanne Lyn Barker-Collo, Till Winfried Bärnighausen, Lope H. Barrero, Bernhard T. Baune, Shahrzad Bazargan-Hejazi, Neeraj Bedi, Ettore Beghi, Masoud Behzadifar, Meysam Behzadifar, Yannick Béjot, Abate Bekele Belachew, Yihalem Abebe Belay, Michelle L. Bell, Aminu K. Bello, Isabela M. Bensenor, Eduardo Bernabe, Robert S. Bernstein, Mircea Beuran, Tina Beyranvand, Neeraj Bhala, Suraj Bhattarai, Soumyadeep Bhaumik, Zulfiqar A. Bhutta, Belete Biadgo, Ali Bijani, Boris Bikbov, Ver Bilano, Nigus Bililign, Muhammad Shahdaat Bin Sayeed, Donal Bisanzio, Brigette F. Blacker, Fiona M. Blyth, Ibrahim R. Bou-Orm, Soufiane Boufous, Rupert Bourne, Oliver J. Brady, Michael Brainin, Luisa C. Brant, Alexandra Brazinova, Nicholas J.K. Breitborde, Hermann Brenner, Paul Svitil Briant, Andrew M. Briggs, Andrey Nikolaevich Briko, Gabrielle Britton, Traolach Brugha, Rachelle Buchbinder, Reinhard Busse, Zahid A. Butt, Lucero Cahuana-Hurtado, Jorge Cano, Rosario Cárdenas, Juan J. Carrero, Austin Carter, Félix Carvalho, Carlos A. Castañeda-Orjuela, Jacqueline Castillo Rivas, Franz Castro, Ferrán Catalá-López, Kelly M. Cercy, Ester Cerin, Yazan Chaiah, Alex R. Chang, Hsing Yi Chang, Jung Chen Chang, Fiona J. Charlson, Aparajita Chattopadhyay, Vijay Kumar Chattu, Pankaj Chaturvedi, Peggy Pei Chia Chiang, Ken Lee Chin, Abdulaal Chitheer, Jee Young J. Choi, Rajiv Chowdhury, Hanne Christensen, Devasahayam J. Christopher, Flavia M. Cicuttini, Liliana G. Ciobanu, Massimo Cirillo, Rafael M. Claro, Daniel Collado-Mateo, Cyrus Cooper, Josef Coresh, Paolo Angelo Cortesi, Monica Cortinovis, Megan Costa, Ewerton Cousin, Michael H. Criqui, Elizabeth A. Cromwell, Marita Cross, John A. Crump, Abel Fekadu Dadi, Lalit Dandona, Rakhi Dandona, Paul I. Dargan, Ahmad Daryani, Rajat Das Gupta, José Das Neves, Tamirat Tesfaye Dasa, Gail Davey, Adrian C. Davis, Dragos Virgil Davitoiu, Barbora De Courten, Fernando Pio De La Hoz, Diego De Leo, Jan Walter De Neve, Meaza Girma Degefa, Louisa Degenhardt, Selina Deiparine, Robert P. Dellavalle, Gebre Teklemariam Demoz, Kebede Deribe, Nikolaos Dervenis, Don C. Des Jarlais, Getenet Ayalew Dessie, Subhojit Dey, Samath Dhamminda Dharmaratne, Mesfin Tadese Dinberu, M. Ashworth Dirac, Shirin Djalalinia, Linh Doan, Klara Dokova, David Teye Doku, E. Ray Dorsey, Kerrie E. Doyle, Tim Robert Driscoll, Manisha Dubey, Eleonora Dubljanin, Eyasu Ejeta Duken, Bruce B. Duncan, Andre R. Duraes, Hedyeh Ebrahimi, Soheil Ebrahimpour, Michelle Marie Echko, David Edvardsson, Andem Effiong, Joshua R. Ehrlich, Charbel El Bcheraoui, Maysaa El Sayed Zaki, Ziad El-Khatib, Hajer Elkout, Iqbal R.F. Elyazar, Ahmadali Enayati, Aman Yesuf Endries, Benjamin Er, Holly E. Erskine, Babak Eshrati, Sharareh Eskandarieh, Alireza Esteghamati, Sadaf Esteghamati, Hamed Fakhim, Vahid Fallah Omrani, Mahbobeh Faramarzi, Mohammad Fareed, Farzaneh Farhadi, Talha A. Farid, Carla Sofia Esá Farinha, Andrea Farioli, Andre Faro, Maryam S. Farvid, Farshad Farzadfar, Valery L. Feigin, Netsanet Fentahun, Seyed Mohammad Fereshtehnejad, Eduarda Fernandes, Joao C. Fernandes, Alize J. Ferrari, Garumma Tolu Feyissa, Irina Filip, Florian Fischer, Christina Fitzmaurice, Nataliya A. Foigt, Kyle J. Foreman, Jack Fox, Tahvi D. Frank, Takeshi Fukumoto, Nancy Fullman, Thomas Fürst, João M. Furtado, Neal D. Futran, Seana Gall, Morsaleh Ganji, Fortune Gbetoho Gankpe, Alberto L. Garcia-Basteiro, William M. Gardner, Abadi Kahsu Gebre, Amanuel Tesfay Gebremedhin, Teklu Gebrehiwo Gebremichael, Tilayie Feto Gelano, Johanna M. Geleijnse, Ricard Genova-Maleras, Yilma Chisha Dea Geramo, Peter W. Gething, Kebede Embaye Gezae, Keyghobad Ghadiri, Khalil Ghasemi Falavarjani, Maryam Ghasemi-Kasman, Mamata Ghimire, Rakesh Ghosh, Aloke Gopal Ghoshal, Simona Giampaoli, Paramjit Singh Gill, Tiffany K. Gill, Ibrahim Abdelmageed Ginawi, Giorgia Giussani, Elena V. Gnedovskaya, Ellen M. Goldberg, Srinivas Goli, Hector Gómez-Dantés, Philimon N. Gona, Sameer Vali Gopalani, Taren M. Gorman, Alessandra C. Goulart, Bárbara Niegia Garcia Goulart, Ayman Grada, Morgan E. Grams, Giuseppe Grosso, Harish Chander Gugnani, Yuming Guo, Prakash C. Gupta, Rahul Gupta, Rajeev Gupta, Tanush Gupta, Bishal Gyawali, Juanita A. Haagsma, Vladimir Hachinski, Nima Hafezi-Nejad, Hassan Haghparast Bidgoli, Tekleberhan B. Hagos, Gessessew Bugssa Hailu, Arvin Haj-Mirzaian, Arya Haj-Mirzaian, Randah R. Hamadeh, Samer Hamidi, Alexis J. Handal, Graeme J. Hankey, Yuantao Hao, Hilda L. Harb, Sivadasanpillai Harikrishnan, Josep Maria Haro, Mehedi Hasan, Hadi Hassankhani, Hamid Yimam Hassen, Rasmus Havmoeller, Caitlin N. Hawley, Roderick J. Hay, Simon I. Hay, Akbar Hedayatizadeh-Omran, Behzad Heibati, Delia Hendrie, Andualem Henok, Claudiu Herteliu, Sousan Heydarpour, Desalegn Tsegaw Hibstu, Huong Thanh Hoang, Hans W. Hoek, Howard J. Hoffman, Michael K. Hole, Enayatollah Homaie Rad, Praveen Hoogar, H. Dean Hosgood, Seyed Mostafa Hosseini, Mehdi Hosseinzadeh, Mihaela Hostiuc, Sorin Hostiuc, Peter J. Hotez, Damian G. Hoy, Mohamed Hsairi, Aung Soe Htet, Guoqing Hu, John J. Huang, Chantal K. Huynh, Kim Moesgaard Iburg, Chad Thomas Ikeda, Bogdan Ileanu, Olayinka Stephen Ilesanmi, Usman Iqbal, Seyed Sina Naghibi Irvani, Caleb Mackay Salpeter Irvine, Sheikh Mohammed, Shariful Islam, Farhad Islami, Kathryn H. Jacobsen, Leila Jahangiry, Nader Jahanmehr, Sudhir Kumar Jain, Mihajlo Jakovljevic, Mehdi Javanbakht, Achala Upendra Jayatilleke, Panniyammakal Jeemon, Ravi Prakash Jha, Vivekanand Jha, John S. Ji, Catherine O. Johnson, Jost B. Jonas, Jacek Jerzy Jozwiak, Suresh Banayya Jungari, Mikk Jürisson, Zubair Kabir, Rajendra Kadel, Amaha Kahsay, Rizwan Kalani, Tanuj Kanchan, Manoochehr Karami, Behzad Karami Matin, André Karch, Corine Karema, Narges Karimi, Seyed M.

Karimi, Amir Kasaeian, Dessalegn H. Kassa, Getachew Mullu Kassa, Tesfaye Dessale Kassa, Nicholas J. Kassebaum, Srinivasa Vittal Katikireddi, Norito Kawakami, Ali Kazemi Karyani, Masoud Masoud Keighobadi, Peter Njenga Keiyoro, Laura Kemmer, Grant Rodgers Kemp, Andre Pascal Kengne, Andre Keren, Yousef Saleh Khader, Behzad Khafaei, Morteza Abdullatif Khafaie, Alireza Khajavi, Ibrahim A. Khalil, Ejaz Ahmad Khan, Muhammad Shahzeb Khan, Muhammad Ali Khan, Young Ho Khang, Mohammad Khazaei, Abdullah T. Khoja, Ardeshir Khosravi, Mohammad Hossein Khosravi, Aliasghar A. Kiadaliri, Daniel N. Kiirithio, Cho Il Kim, Daniel Kim, Pauline Kim, Young Eun Kim, Yun Jin Kim, Ruth W. Kimokoti, Yohannes Kinfu, Adnan Kisa, Katarzyna Kissimova-Skarbek, Mika Kivimäki, Ann Kristin Skrindo Knudsen, Jonathan M. Kocarnik, Sonali Kochhar, Yoshihiro Kokubo, Tufa Kolola, Jacek A. Kopec, Soewarta Kosen, Georgios A. Kotsakis, Parvaiz A. Koul, Ai Koyanagi, Michael A. Kravchenko, Kewal Krishan, Kristopher J. Krohn, Barthelemy Kuate Defo, Burcu Kucuk Bicer, G. Anil Kumar, Manasi Kumar, Hmwe Hmwe Kyu, Deepesh P. Lad, Sheetal D. Lad, Alessandra Lafranconi, Ratilal Lalloo, Tea Lallukka, Faris Hasan Lami, Van C. Lansingh, Arman Latifi, Kathryn Mei Ming Lau, Jeffrey V. Lazarus, Janet L. Leasher, Jorge R. Ledesma, Paul H. Lee, James Leigh, Janni Leung, Miriam Levi, Sonia Lewycka, Shanshan Li, Yichong Li, Yu Liao, Misgan Legesse Liben, Lee Ling Lim, Stephen S. Lim, Shiwei Liu, Rakesh Lodha, Katharine J. Looker, Alan D. Lopez, Stefan Lorkowski, Paulo A. Lotufo, Nicola Low, Rafael Lozano, Tim C.D. Lucas, Lydia R. Lucchesi, Raimundas Lunevicius, Ronan A. Lyons, Stefan Ma, Erlyn Rachelle King Macarayan, Mark T. Mackay, Fabiana Madotto, Hassan Magdy Abd El Razek, Muhammed Magdy Abd El Razek, Dhaval P. Maghavani, Narayan Bahadur Mahotra, Hue Thi Mai, Marek Majdan, Reza Majdzadeh, Azeem Majeed, Reza Malekzadeh, Deborah Carvalho Malta, Abdullah A. Mamun, Ana Laura Manda, Helena Manguerra, Treh Manhertz, Mohammad Ali Mansournia, Lorenzo Giovanni Mantovani, Chabila Christopher Mapoma, Joemer C. Maravilla, Wagner Marcenes, Ashley Marks, Francisco Rogerlândio Martins-Melo, Ira Martopullo, Winfried März, Melvin B. Marzan, Tivani Phosa Mashamba-Thompson, Benjamin Ballard Massenburg, Manu Raj Mathur, Kunihiro Matsushita, Pallab K. Maulik, Mohsen Mazidi, Colm McAlinden, John J. McGrath, Martin McKee, Man Mohan Mehndiratta, Ravi Mehrotra, Kala M. Mehta, Varshil Mehta, Fabiola Mejia-Rodriguez, Tesfa Mekonen, Addisu Melese, Mulugeta Melku, Michele Meltzer, Peter T.N. Memiah, Ziad A. Memish, Walter Mendoza, Desalegn Tadese Mengistu, Getnet Mengistu, George A. Mensah, Seid Tiku Mereta, Atte Meretoja, Tuomo J. Meretoja, Tomislav Mestrovic, Naser Mohammad Gholi Mezerji, Bartosz Miazgowski, Tomasz Miazgowski, Anoushka I. Millear, Ted R. Miller, Benjamin Miltz, G. K. Mini, Mojde Mirarefin, Erkin M. Mirrakhimov, Awoke Temesgen Misganaw, Philip B. Mitchell, Habtamu Mitiku, Babak Moazen, Bahram Mohajer, Karzan Abdulmuhsin Mohammad, Noushin Mohammadifard, Mousa Mohammadnia-Afrouzi, Mohammed A. Mohammed, Shafiu Mohammed, Farnam Mohebi, Modhurima Moitra, Ali H. Mokdad, Mariam Molokhia, Lorenzo Monasta, Yoshan Moodley, Mahmood Moosazadeh, Ghobad Moradi, Maziar Moradi-Lakeh, Mehdi Moradinazar, Paula Moraga, Lidia Morawska, Ilais Moreno Velásquez, Joana Morgado-Da-Costa, Shane Douglas Morrison, Marilita M. Moschos, Seyyed Meysam Mousavi, Kalayu Brhane Mruts, Achenef Asmamaw Muche, Kindie Fentahun Muchie, Ulrich Otto Mueller, Oumer Sada Muhammed, Satinath Mukhopadhyay, Kate Muller, John Everett Mumford, Manoj Murhekar, Jonah Musa, Kamarul Imran Musa, Ghulam Mustafa, Ashraf F. Nabhan, Chie Nagata, Mohsen Naghavi, Aliya Naheed, Azin Nahvijou, Gurudatta Naik, Nitish Naik, Farid Najafi, Luigi Naldi, Hae Sung Nam, Vinay Nangia, Jobert Richie Nansseu, Bruno Ramos Nascimento, Gopalakrishnan Natarajan, Nahid Neamati, Ionut Negoi, Ruxandra Irina Negoi, Subas Neupane, Charles Richard James Newton, Josephine W. Ngunjiri, Anh Quynh Nguyen, Ha Thu Nguyen, Huong Lan Thi Nguyen, Huong Thanh Nguyen, Long Hoang Nguyen, Minh Nguyen, Nam Ba Nguyen, Son Hoang Nguyen, Emma Nichols, Dina Nur Anggraini Ningrum, Molly R. Nixon, Nomonde Nolutshungu, Shuhei Nomura, Ole F. Norheim, Mehdi Noroozi, Bo Norrving, Jean Jacques Noubiap, Hamid Reza Nouri, Malihe Nourollahpour Shiadeh, Mohammad Reza Nowroozi, Elaine O. Nsoesie, Peter S. Nyasulu, Christopher M. Odell, Richard Ofori-Asenso, Felix Akpojene Ogbo, In Hwan Oh, Olanrewaju Oladimeji, Andrew T. Olagunju, Tinuke O. Olagunju, Pedro R. Olivares, Helen Elizabeth Olsen, Bolajoko Olubukunola Olusanya, Kanyin L. Ong, Sok King Ong, Eyal Oren, Alberto Ortiz, Erika Ota, Stanislav S. Otstavnov, Simon øverland, Mayowa Ojo Owolabi, P. A. Mahesh, Rosana Pacella, Amir H. Pakpour, Adrian Pana, Songhomitra Panda-Jonas, Andrea Parisi, Eun Kee Park, Charles D.H. Parry, Shanti Patel, Sanghamitra Pati, Snehal T. Patil, Ajay Patle, George C. Patton, Vishnupriya Rao Paturi, Katherine R. Paulson, Neil Pearce, David M. Pereira, Norberto Perico, Konrad Pesudovs, Hai Quang Pham, Michael R. Phillips, David M. Pigott, Julian David Pillay, Michael A. Piradov, Meghdad Pirsaheb, Farhad Pishgar, Oleguer Plana-Ripoll, Dietrich Plass, Suzanne Polinder, Svetlana Popova, Maarten J. Postma, Akram Pourshams, Hossein Poustchi, Dorairaj Prabhakaran, Swayam Prakash, V. Prakash, Caroline A. Purcell, Manorama B. Purwar, Mostafa Qorbani, D. Alex Quistberg, Amir Radfar, Anwar Rafay, Alireza Rafiei, Fakher Rahim, Kazem Rahimi, Afarin Rahimi-Movaghar, Vafa Rahimi-Movaghar, Mahfuzar Rahman, Mohammad Hifz Ur Rahman, Muhammad Aziz Rahman, Sajjad Ur Rahman, Rajesh Kumar Rai, Fatemeh Rajati, Usha Ram, Prabhat Ranjan, Anna Ranta, Puja C. Rao, David Laith Rawaf, Salman Rawaf, K. Srinath Reddy, Robert C. Reiner, Nickolas Reinig, Marissa Bettay Reitsma, Giuseppe Remuzzi, Andre M.N. Renzaho, Serge Resnikoff, Satar Rezaei, Mohammad Sadegh Rezai, Antonio Luiz P. Ribeiro, Stephen R. Robinson, Leonardo Roever, Luca Ronfani, Gholamreza Roshandel, Ali Rostami, Gregory A. Roth, Ambuj Roy, Enrico Rubagotti, Perminder S. Sachdev, Nafis Sadat, Basema Saddik, Ehsan Sadeghi, Sahar Saeedi Moghaddam, Hosein Safari, Yahya Safari, Roya Safari-Faramani, Mahdi Safdarian, Sare Safi, Saeid Safiri, Rajesh Sagar, Amirhossein Sahebkar, Mohammad Ali Sahraian, Haniye Sadat Sajadi, Nasir Salam, Joseph S. Salama, Payman Salamati, Komal Saleem, Zikria Saleem, Yahya Salimi, Joshua A. Salomon, Sundeep Santosh Salvi, Inbal Salz, Abdallah M. Samy, Juan Sanabria, Yingying Sang, Damian Francesco Santomauro, Itamar S. Santos, João Vasco Santos, Milena M. Santric Milicevic, Bruno Piassi Sao Jose, Mayank Sardana, Abdur Razzaque Sarker, Nizal Sarrafzadegan, Benn Sartorius, Shahabeddin Sarvi, Brijesh Sathian, Maheswar Satpathy, Arundhati R. Sawant, Monika Sawhney, Sonia Saxena, Mete Saylan, Elke Schaeffner, Maria Inês Schmidt, Ione J.C. Schneider, Ben Schöttker, David C. Schwebel, Falk Schwendicke, James G. Scott, Mario Sekerija, Sadaf G. Sepanlou, Edson Serván-Mori, Seyedmojtaba Seyedmousavi, Hosein Shabaninejad, Azadeh Shafieesabet, Mehdi Shahbazi, Amira A. Shaheen, Masood Ali Shaikh, Mehran Shams-Beyranvand, Mohammadbagher Shamsi, Morteza Shamsizadeh, Heidar Sharafi, Kiomars Sharafi, Mehdi Sharif, Mahdi Sharif-Alhoseini, Meenakshi Sharma, Rajesh Sharma, Jun She, Aziz Sheikh, Peilin Shi, Kenji Shibuya, Mika Shigematsu, Rahman Shiri, Reza Shirkoohi, Kawkab Shishani, Ivy Shiue, Farhad Shokraneh, Haitham Shoman, Mark G. Shrime, Si Si, Soraya Siabani, Tariq J. Siddiqi, Inga Dora Sigfusdottir, Rannveig Sigurvinsdottir, João Pedro Silva, Dayane Gabriele Alves Silveira, Narayana Sarma Venkata Singam, Jasvinder A. Singh, Narinder Pal Singh, Virendra Singh, Dhirendra Narain Sinha, Eirini Skiadaresi, Erica Leigh N. Slepak, Karen Sliwa, David L. Smith, Mari Smith, Adauto Martins Soares Filho, Badr Hasan Sobaih, Soheila Sobhani, Eugène Sobngwi, Samir S. Soneji, Moslem Soofi, Masoud Soosaraei, Reed J.D. Sorensen, Joan B. Soriano, Ireneous N. Soyiri, Luciano A. Sposato, Chandrashekhar T. Sreeramareddy, Vinay Srinivasan, Jeffrey D. Stanaway, Dan J. Stein, Caitlyn Steiner, Timothy J. Steiner, Mark A. Stokes, Lars Jacob Stovner, Michelle L. Subart, Agus Sudaryanto, Mu'awiyyah Babale Sufiyan, Bruno F. Sunguya, Patrick John Sur, Ipsita Sutradhar, Bryan L. Sykes, Dillon O. Sylte, Rafael Tabarés-Seisdedos, Santosh Kumar Tadakamadla, Birkneh Tilahun Tadesse, Nikhil Tandon, Segen Gebremeskel Tassew, Mohammad Tavakkoli, Nuno Taveira, Hugh R. Taylor, Arash Tehrani-Banihashemi, Tigist Gashaw Tekalign, Shishay Wahdey Tekelemedhin, Merhawi Gebremedhin Tekle, Habtamu Temesgen, Mohamad Hani Temsah, Omar Temsah, Abdullah Sulieman Terkawi, Mebrahtu Teweldemedhin,

Kavumpurathu Raman Thankappan, Nihal Thomas, Binyam Tilahun, Quyen G. To, Marcello Tonelli, Roman Topor-Madry, Fotis Topouzis, Anna E. Torre, Miguel Tortajada-Girbés, Mathilde Touvier, Marcos Roberto Tovani-Palone, Jeffrey A. Towbin, Bach Xuan Tran, Khanh Bao Tran, Christopher E. Troeger, Thomas Clement Truelsen, Miltiadis K. Tsilimbaris, Derrick Tsoi, Lorainne Tudor Car, E. Murat Tuzcu, Kingsley N. Ukwaja, Irfan Ullah, Eduardo A. Undurraga, Jurgen Unutzer, Rachel L. Updike, Muhammad Shariq Usman, Olalekan A. Uthman, Muthiah Vaduganathan, Afsane Vaezi, Pascual R. Valdez, Santosh Varughese, Tommi Juhani Vasankari, Narayanaswamy Venketasubramanian, Santos Villafaina, Francesco S. Violante, Sergey Konstantinovitch Vladimirov, Vasily Vlassov, Stein Emil Vollset, Kia Vosoughi, Isidora S. Vujcic, Fasil Shiferaw Wagnew, Yasir Waheed, Stephen G. Waller, Yafeng Wang, Yuan Pang Wang, Elisabete Weiderpass, Robert G. Weintraub, Daniel J. Weiss, Fitsum Weldegebreal, Kidu Gidey Weldegwergs, Andrea Werdecker, T. Eoin West, Harvey A. Whiteford, Justyna Widecka, Tissa Wijeratne, Lauren B. Wilner, Shadrach Wilson, Andrea Sylvia Winkler, Alison B. Wiyeh, Charles Shey Wiysonge, Charles D.A. Wolfe, Anthony D. Woolf, Shouling Wu, Yun Chun Wu, Grant M.A. Wyper, Denis Xavier, Gelin Xu, Simon Yadgir, Ali Yadollahpour, Seyed Hossein Yahyazadeh Jabbari, Tomohide Yamada, Lijing L. Yan, Yuichiro Yano, Mehdi Yaseri, Yasin Jemal Yasin, Alex Yeshaneh, Ebrahim M. Yimer, Paul Yip, Engida Yisma, Naohiro Yonemoto, Seok Jun Yoon, Marcel Yotebieng, Mustafa Z. Younis, Mahmoud Yousefifard, Chuanhua Yu, Vesna Zadnik, Zoubida Zaidi, Sojib Bin Zaman, Mohammad Zamani, Zohreh Zare, Ayalew Jejaw Zeleke, Zerihun Menlkalew Zenebe, Kai Zhang, Zheng Zhao, Maigeng Zhou, Sanjay Zodpey, Inbar Zucker, Theo Vos and Christopher J.L. Murray. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: A systematic analysis for the global burden of disease study 2017. *The Lancet*, 392: 1789–1858, 11 2018. ISSN 1474547X. doi: 10.1016/S0140-6736(18)32279-7. URL `https://github.com/ihmeuw/`.

[20] Heejung Kim, SungHee Lee, SangEun Lee, Soyun Hong, HeeJae Kang and Namhee Kim. Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: Observational study on older adults living alone. *JMIR Mhealth Uhealth*, 7(10):e14149, Oct 2019. ISSN 2291-5222. doi: 10.2196/14149. URL `http://mhealth.jmir.org/2019/10/e14149/`.

[21] RE. Mastoras, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Kassie, T. Alsaadi, A. Khandoker and L. J. Hadjileontiadis. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci Rep*, 9, 13414 2019. doi: https://doi.org/10.1038/s41598-019-50002-9.

[22] Mohammed T. Masud, Mohammed A. Mamun, K. Thapa, D.H. Lee, Mark D. Griffiths and S.-H. Yang. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. *Journal of Biomedical Informatics*, 103:103371, 2020. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103371. URL https://www.sciencedirect.com/science/article/pii/S1532046419302916.

[23] John Hopkins Medicine. Mood disorders | johns hopkins medicine. URL https://www.hopkinsmedicine.org/health/conditions-and-diseases/mood-disorders.

[24] Nematjon Narziev, Hwarang Goh, Kobiljon Toshnazarov, Seung Ah Lee, Kyong-Mee Chung and Youngtae Noh. Stdd: Short-term depression detection with passive sensing. *Sensors*, 20(5), 2020. ISSN 1424-8220. doi: 10.3390/s20051396. URL https://www.mdpi.com/1424-8220/20/5/1396.

[25] Jukka Pekka Onnela and Scott L. Rauch. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health, 6 2016. ISSN 1740634X. URL www.neuropsychopharmacology.org.

[26] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin and M. De Vos. Detecting bipolar depression from geographic location data. *IEEE Trans Biomed Eng.*, 64(8):1761–1771, 2017. doi: https://doi.org/10.1109/tbme.2016.2611862. URL https://pubmed.ncbi.nlm.nih.gov/28113247/.

[27] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F. Ionescu, Darian Bhathena, Lauren B. Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, Lisa Sangermano, David Mischoulon, Johnathan E. Alpert and Rosalind W. Picard. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in Psychiatry*, 11:1413, 2020. ISSN 1664-0640. doi: 10.3389/fpsyt.2020.584711. URL https://www.frontiersin.org/article/10.3389/fpsyt.2020.584711.

[28] Mary L. Phillips and David J. Kupfer. Bipolar disorder 2 - bipolar disorder diagnosis: Challenges and future directions. *The Lancet*, 381:1663–1671, 2013. ISSN 1474547X. doi: 10.1016/S0140-6736(13)60989-7.

[29] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording and David C Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *J Med Internet Res*, 17(7):e175, Jul 2015. ISSN 1438-8871. doi: 10.2196/jmir.4273. URL http://www.jmir.org/2015/7/e175/.

[30] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, 2014. ISBN 987-1-107-05713-5.

[31] Adrian B.R. Shatte, Delyse M. Hutchinson and Samantha J. Teague. Machine learning in mental health: A scoping review of methods and applications, 7 2019. ISSN 14698978.

[32] Rajput Singh and Tanvir Rajput. Misdiagnosis of bipolar disorder. *American Journal of Managed Care*, 11:57–63, 2005. ISSN 10880224. URL /pmc/articles/PMC2945875/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2945875/.

[33] Yuuki Tazawa, Kuo ching Liang, Michitaka Yoshimura, Momoko Kitazawa, Yuriko Kaise, Akihiro Takamiya, Aiko Kishi, Toshiro Horigome, Yasue Mitsukura, Masaru Mimura and Taishiro Kishimoto. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*, 6(2):e03274, 2020. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2020.e03274. URL https://www.sciencedirect.com/science/article/pii/S2405844020301195.

[34] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer and Steffi Weidt. Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR Mhealth Uhealth*, 4(3):e111, Sep 2016. ISSN 2291-5222. doi: 10.2196/mhealth.5960. URL http://mhealth.jmir.org/2016/3/e111/.

[35] Chaoqun Yue, Shweta Ware, Reynaldo Morillo, Jin Lu, Chao Shang, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis and Bing Wang. Automatic depression prediction using internet traffic characteristics on smartphones. *Smart Health*, 18:100137, 2020. ISSN 2352-6483. doi: https://doi.org/10.1016/j.smhl.2020.100137. URL https://www.sciencedirect.com/science/article/pii/S2352648320300295.