Pervasive and Mobile Computing 51 (2018) 1–26

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Review

Mental health monitoring with multimodal sensing and machine learning: A survey

Enrique Garcia-Ceja [a,*], Michael Riegler [a,b], Tine Nordgreen [c,d], Petter Jakobsen [c,e], Ketil J. Oedegaard [f,g], Jim Tørresen [a]

[a] Department of Informatics, University of Oslo, Oslo, Norway
[b] Simula Metropolitan Center for Digital Engineering, Norway
[c] Division of Psychiatry, Haukeland University Hospital, Bergen, Norway
[d] Department of Clinical Psychology, Faculty of Psychology, University of Bergen, Bergen, Norway
[e] Department of Clinical Medicine, University of Bergen, Bergen, Norway
[f] NORMENT, Division of Psychiatry, Haukeland University Hospital, Bergen, Norway
[g] K.G. Jebsen Centre for Neuropsychiatric Disorders, Department of Clinical Medicine, University of Bergen, Bergen, Norway

ARTICLE INFO

Article history:
Received 17 November 2017
Received in revised form 17 August 2018
Accepted 15 September 2018
Available online 19 September 2018

MSC:
00-01
99-00

Keywords:
Mental health
Machine learning
Smartphones
Mental disorders
Sensors

ABSTRACT

Personal and ubiquitous sensing technologies such as smartphones have allowed the continuous collection of data in an unobtrusive manner. Machine learning methods have been applied to continuous sensor data to predict user contextual information such as location, mood, physical activity, etc. Recently, there has been a growing interest in leveraging ubiquitous sensing technologies for mental health care applications, thus, allowing the automatic continuous monitoring of different mental conditions such as depression, anxiety, stress, and so on. This paper surveys recent research works in mental health monitoring systems (MHMS) using sensor data and machine learning. We focused on research works about mental disorders/conditions such as: depression, anxiety, bipolar disorder, stress, etc. We propose a classification taxonomy to guide the review of related works and present the overall phases of MHMS. Moreover, research challenges in the field and future opportunities are also discussed.

Contents

1. Introduction 2
2. Related work 3
3. Taxonomy 3
   3.1. Study type 8
      3.1.1. Association 8
      3.1.2. Detection 8
      3.1.3. Forecasting 8
   3.2. Study duration 9
   3.3. Sensing types 9
      3.3.1. External sensors 9
      3.3.2. Wearable sensors 9
      3.3.3. Software and social media sensing 11

* Corresponding author.
E-mail address: enriqug@ifi.uio.no (E. Garcia-Ceja).

https://doi.org/10.1016/j.pmcj.2018.09.003

3.3.4. Discussion ... 12
4. Work flow of mental health monitoring systems ... 12
4.1. Prerequisites ... 13
4.2. Experiment design ... 13
4.3. Ethical approvals and participants' consent ... 14
4.4. Data acquisition and labeling ... 14
4.5. Exploratory data analysis and preprocessing ... 15
4.6. Machine learning model training ... 16
4.6.1. Types of models ... 16
4.6.2. Machine learning software tools ... 17
4.7. Machine learning model evaluation ... 18
4.8. Clinical evaluation ... 19
4.9. Deployment ... 19
5. Research challenges and opportunities ... 20
5.1. Data labeling ... 20
5.2. Inter-user variance ... 20
5.3. Intra-user variance ... 20
5.4. Sensor data fusion ... 20
5.5. Integration with other systems ... 21
5.6. Clinical validation ... 21
6. Limitations of this survey ... 21
7. Conclusions ... 22
Acknowledgment ... 22
References ... 22

1. Introduction

Mental health problems are common worldwide including changes in mood, personality, inability to cope with daily problems or stress, withdrawal from friends and activities, and so on. In 2010 mental health problems were the leading causes of years lived with disability (YLDs) worldwide with depressive and anxiety disorders among the most frequent disorders [1]. Polanczyk et al. estimated a worldwide prevalence of mental disorders in children and adolescents of 13.4% [2]. In the last years, their prevalence has increased even further [3,4]. Mental disorders can have a serious impact for the patients but also for their families, friends and society, since it is difficult to cope with the implications of someone close having a mental illness.

Dealing with a mental disorder can be physically, economically and emotionally demanding. Work impairment is one of the adverse consequences of mental illness [5] and is also the leading cause for hospital admissions [6]. According to the World Mental Health Survey Consortium, the proportion of respondents who received treatment for emotional or substance-use problems is much larger in developed than in less-developed countries [7]. Nonetheless, the unmet need for treatment of mental disorders is a major problem in both, developed and less-developed countries but being larger in the latter.

The chronic and relapsing nature of many mental health disorders are the rule and not the exception, thus, the need for long-term follow up and assessment methods become essential for patients' symptoms reduction and recovery. Traditional monitoring methods rely on retrospective reports which are subject to recall bias. This approach limits the ability to accurately characterize, understand, and change behavior in real world settings as pointed out by Shiffman et al. [8]. An alternative to retrospective reports, is the so called Ecological Momentary Assessment (EMA) which allows repeated sampling of thoughts, feelings and behaviors as close in time to the experience as possible in real-life situations [8]. EMA measurements have been shown to outperform paper and pencil reports in the assessment of some mental states in terms of sensitivity to detect changes [9]. The increasing capabilities of smartphones and wearable devices make them potential platforms for EMA measurements, monitoring mental illness, treatment, self-management and interventions, thus, reducing costs and expanding the coverage of mental health services to larger populations.

Smartphones and alike devices have been demonstrated to have potential in providing mental health interventions [10–12]. Wearable devices like smartphones, smart watches and fitness bands, have a vast variety of embedded sensors. These can include communication devices (WiFi, Bluetooth, etc.), inertial sensors (accelerometer, gyroscope, etc.), physiological sensors (heart rate, dermal activity, etc.) and ambient sensors (ambient pressure, temperature, etc.) to name a few. This opens the possibility of multimodal sensing applications in the healthcare domain [13]. By combining the data from subsets of those sensors, it is possible to infer contextual information such as physical activity [14], location [15], mood [16] and social relationships [17]; among others. Multimodal sensing settings have shown to produce better results in some applications, compared to single sensor modalities [18,19]. Knowing the contextual information about a user, can help in providing more fine grained personalized just-in-time services. On the other hand, general solutions do not take into account the individual characteristics of each person. Analyzing large amounts of sensor data is a complicated task (if not impossible) to do by hand, and this is where machine learning becomes important. By using machine learning methods, it is possible to extract meaningful information from sensor data and use it to continuously monitor the current users' state.

It has been shown that mental states can manifest through physiological and behavioral changes. For example, a systematic review [20] states that hypoactive electrodermal response is an established feature of patients affected by depression. In this review the authors also found evidence that monitoring electrodermal activity may be useful to differentiate phases of mood disorders. Chang et al. [21] found that bipolar mania is associated with cardiac autonomic dysregulation. Regarding behavioral changes, Berle et al. [22] found that motor activity recorded with actigraphs was significantly reduced in both schizophrenic and depressed patients compared to controls. Based on the evidence that there are associations between physiological/behavioral and mood states, wearable sensors have the potential to monitor mental conditions continuously and in an unobtrusive manner.

In this paper, we survey novel research works about mental health monitoring systems (MHMS) that make use of sensors and mainly machine learning. We focused on research works about mental disorders/conditions such as: depression, anxiety disorders, bipolar disorder, stress, epilepsy, etc. The goal of this work is to survey relevant work that illustrates the current use of technology (multimodal sensing and machine learning) for automatic and adaptive mental health monitoring.

The main contributions of this work are (i) an overview of the state-of-the-art of research in context of MHMS, (ii) a classification taxonomy for MHMS research works, (iii) an overview of sensors and the work flow of sensor based monitoring systems and (iv) a list of research challenges and opportunities in the field.

The paper is structured as following: Section 2 presents related work (surveys, review works) about MHMS and the use of wearable devices for automatic monitoring. In Section 3 we describe our proposed classification taxonomy of MHMS, covering study types, study duration and sensing types. Section 4 describes the general steps of sensor based monitoring systems, from the prerequisites to deployment. Section 5 introduces some of the research challenges and future opportunities in MHMS. In Section 6 we present the limitations of this work. Finally, we conclude this paper in Section 7.

2. Related work

In this section we present related work of MHMS, specifically, survey/review works that focus on monitoring human mental health by using multimodal sensing devices. This survey differs from the previous ones in the following aspects: Firstly, instead of focusing on a particular mental state, this work presents an overview of how different types of devices and sensing modalities have been used to monitor for a wide variety of different mental conditions. Secondly, we propose a classification taxonomy based on the abstraction of common characteristics across different MHMS. This taxonomy provides researchers in this emerging field with a general overview and can also serve as a preliminary guide to current researchers when designing their studies to monitor a particular mental state. Thirdly, while previous reviews focus exclusively on smartphones as sensing devices, this work uses the more general term *wearable devices* which encompasses the use of smartphones but also newer types of devices like smart watches, bands, flexible sensors, etc. This work also surveys *external sensors* and *software/social media* as sensing types. The possibility of capturing patient data in real-time and in real-life conditions represents a feasible tool for mental health monitoring and treatment. According to Torous et al. [11], there is an interest among patients in using mobile applications on a daily basis to monitor their mental health condition. Advances in different technological fields make it possible to design and build these types of systems. Table 1 presents a list of review and survey works regarding the use of wearable sensors for health care monitoring, assessment and treatment.

An overview of how mobile phones can be used as devices for mental disorder treatments was presented by Gravenhorst et al. [23]. In their work, they discussed several important aspects of MHMS such as human computer interfaces, practical implementations, legal issues, business models, collection of relevant data through sensors like voice, motion and location, etc. Given the nature of their work, they did not present a classification scheme that would help the reader understand the different types of works within the field which is one of the objectives of this work. Another similar work was presented by Mohr et al. [27] which is a review of personal sensing for mental health. They provided a framework for converting raw data into knowledge, however, previous steps required by MHMS were not covered; such as experiment design and how the raw data is acquired. This steps will be further discussed in this work (Section 4). Nicholas, J. et al. [24] conducted a systematic review of smartphone applications for bipolar disorder in Google Play and iOS stores. Their review included applications that provide information, but also applications that are management tools for screening and assessment, symptom monitoring, community support and treatment. They found that in general, the content of these applications is not in line with practice guidelines or established self-management principles. Donker et al. [12] conducted a systematic review of mental health applications and concluded that they have the potential to improve treatment accessibility and reduce symptoms, however the majority of applications available to the public lack scientific evidence about their efficacy. While most of the survey works focus on a particular mental condition (e.g., bipolar disorder [24], social anxiety disorder [34], etc.) this work surveys the use of different sensing types across different mental conditions with the advantage of giving a comprehensive overview and reveal synergies within the different conditions. In the next section we present our classification taxonomy.

3. Taxonomy

In this section, we present our proposed classification taxonomy for automatic MHMS works. Fig. 1 graphically depicts the classification taxonomy which is divided into three main levels, namely: *study type*, *study duration* and *sensing types*. Subsequently, each level has different types of categories. For instance, the *study type* can be association, detection and forecasting. The *study duration* can be either short-term or long-term. Finally, *sensing types* can be wearable, external and

Table 1
List of related survey/review papers.

Reference	Main considered devices	Description
Gravenhorst, F. et al. [23]	smartphones	Overview of how mobile phones can support the treatment of mental disorders.
Nicholas, J. et al. [24]	smartphones	A systematic review of android and iOS applications for bipolar disorder.
Pantelopoulos, A. and Bourbakis, N.G. [25]	wearable sensors	Surveys research and developments on wearable biosensor systems for health monitoring.
Donker, T. et al. [12]	smartphones/tablets	A systematic review of research evidence supporting the efficacy of mental health apps for mobile devices.
Bayndr, L. [26]	smartphones	A survey about using smartphones to detect human behavior including health related activities like physical exercise and sleeping.
Mohr, D.C. et al. [27]	smartphones and other wearable sensors	A review of personal sensing research related to mental health and a framework to convert raw data into knowledge.
Stephens, J. and Allen, J. [28]	smartphones	Systematic review of smartphone applications and text messaging in promoting weight reduction and physical activity.
Guntuku, S.C. et al. [29]	social media	A review of recent studies that analyze social media to detect depression and mental illness.
Wang, J. et al. [30]	smartphones	A review of smartphone interventions for long-term chronic condition management including mental health problems.
Huguet, A. et al. [31]	smartphones	Systematic review of self-help apps for people with depression and evaluate those that offer cognitive behavioral therapy or behavioral activation.
Alotaiby, T.N. et al. [32]	electroencephalogram	A survey of seizure detection and prediction challenges and algorithms.
Mosenia A. et al. [33]	wearable sensors	A survey of applications and architecture of wearable medical sensors systems.

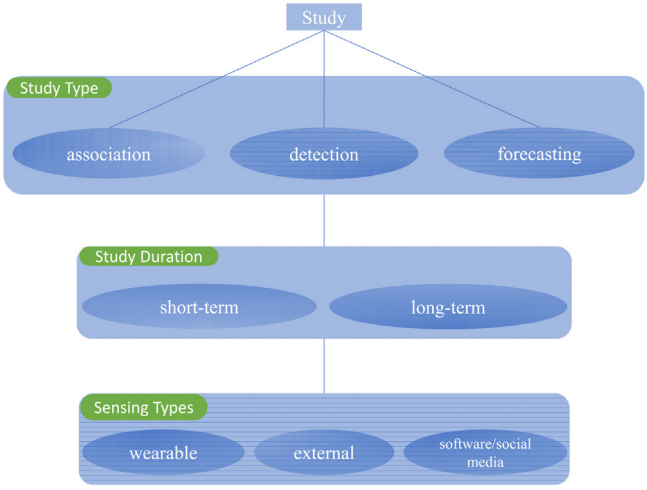

Fig. 1. Our proposed taxonomy based on study type, study duration and sensing types.

software/social media. The following subsections describe each level in detail. Table 2 presents what we have identified as the most representative works classified according to this taxonomy. From this table, we can see that for anxiety and stress the studies are often short-term and for bipolar disorder they are long-term. This is because the transition between states in some disorders (e.g. bipolar) takes more time, thus, long-term studies are needed to capture the state transitions. We can also observe that there are fewer works related to forecasting than related to association and detection. This might be due the fact that forecasting is a more challenging problem in the field of machine learning (Section 3.1.3). It can also be seen

Table 2
Classification of representative works based on our taxonomy.

Reference	Study Type (mental condition)	Study Duration (mean time/participant)	Sensing Types	Sensors	Study Participants	Main Results
Sampei, K. et al. [35]	association (fatigue)	short-term (60 min.)	wearable	eye sensor	3 males and 2 females. Ages 21-25.	3 out of 5 subjects showed correlation between workload and information from the eye detection system.
O'brien, J.T. et al. [36]	association (depression)	long-term (7 days)	wearable	wrist accelerometer	29 adults with depression and 30 healthy controls over 60 years.	Physical activity was reduced in the condition group and subjects showed slower fine motor movements.
Slater, M. et al. [37]	association (anxiety)	short-term (3 public talks)	external	VR headset, questionnaires	7 postgraduate students, 1 undergraduate, 2 faculty members at University College London. All subjects were in their 20s and 30s with only one woman.	Higher perceived audience interest increases self-rating and reduces public speaking anxiety.
Grillon, H. et al. [38]	association (social phobia)	long-term (10 weeks)	wearable	pupil-corneal reflection and head tracker	5 social phobic patients and 5 non phobic subjects aged 25–55.	Differences in salient facial feature avoidance and hyperscanning were found for some of the subjects.
Miranda D. et al. [39]	association (anxiety)	short-term (45 min)	wearable	eye, heart rate	4 graduate students with social anxiety and 6 without. 8 men and 2 females.	Found higher average heart rates after induced anxiety on the mild SAD group.
Faurholt-Jepsen, M. et al. [40]	association (bipolar disorder)	long-term (12 weeks)	wearable	sms, calls, screen, location	29 patients with bipolar disorder recruited from the Copenhagen Clinic for Affective Disorders, Psychiatric Center Copenhagen, Denmark.	Found correlations between bipolar states severity and screen on time, number of calls/day, cell tower ids.
Holmgard, C. et al. [41]	detection (stress)	short-term (6 gameplay sessions)	wear-able/external	dermal activity	14 male PTSD diagnosed veteran soldiers aged 22–32.	Found significant correlations between physiological responses and subjective evaluations.
Price, M. et al. [42]	association (social phobia)	8 sessions	wearable	head mounted display	41 subjects with social phobia 60% females.	Suggest that different components of presence are associated with the experience of fear and treatment response to VRE.

(*continued on next page*)

Table 2 (*continued*).

Reference	Study Type (mental condition)	Study Duration (mean time/participant)	Sensing Types	Sensors	Study Participants	Main Results
Mozos, O.M. et al. [43]	detection (stress)	short-term (17 min)	wearable	electrodermal activity, photoplethysmogram, heart rate variability, microphone, accelerometer	18 students from the School of Psychology at the University of Lincoln aged 18–39.	Classify stressful and neutral situations. Accuracy .94 precision .94 recall .96 with AdaBoost.
Leng, L.B. et al. [44]	detection (drowsiness)	short-term (1 day)	wearable	accelerometer, gyroscope, photoplethysmogram, galvanic skin response	Mentally healthy 15 males and 5 females. Mean age 35.	Accuracy of .98 with SVM.
Garcia-Ceja, E. et al. [45]	detection (stress)	long-term (8 weeks)	wearable	accelerometer	30 healthy subjects from two companies in Trentino, Italy. Mean age 37. 60% males.	Accuracy of .71 for low, moderate and high stress with Naive Bayes and Decision Trees.
Keshan, N. et al. [46]	detection(stress)	short-term	wearable	heart rate	10 drivers.	.88 accuracy in detecting 3 levels of stress with Random Tree
Miranda, D. et al. [47]	detection (anxiety)	short-term (3 sessions)	wearable	heart rate, dermal activity	8 males 2 females. Mean age 24.	Accuracy of .73 with Markov chain model.
Carneiro, D. et al. [48]	detection (stress)	short-term	wear-able/external	video cameras, accelerometers, pressure sensitive touchscreens	19 participants aged 20–57 at the Intelligent Systems Lab of the University of Minho.	.78 accuracy classifying stress and no stress with J48 tree.
Muaremi, A. et al. [49]	detection (stress)	long-term (4 months)	wearable	heart rate, audio, acceleration, gps, calls, contacts	35 subjects from three IT companies.	.61 accuracy for user specific models with Multinomial Logistic Regression.
Giakoumis, D. et al. [50]	detection (stress)	short-term	external	video, accelerometers, dermal activity, heart rate	17 males and 4 females at the Informatics and Telematics Institute, Centre for Research and Technology Hellas.	Accuracy of 1.0 when using all sensors and more extreme cases of stressed and not stressed with Linear Discriminant Analysis.
Lu, H. et al. [51]	detection (stress)	short-term (4 days)	wear-able/external	sound	10 females and 4 males with mean age 22. Thirteen were undergraduate students and one PhD student.	Accuracy of .81 with GMMs
Sano, A. and Picard, R.W. [52]	detection (stress)	short-term (5 days)	wearable	accelerometer, skin conductance, calls, sms, location, screen	15 healthy males and 3 females. Average age 28.	Accuracies over .75 with SVM and KNN.

(*continued on next page*)

Table 2 (*continued*).

Reference	Study Type (mental condition)	Study Duration (mean time/participant)	Sensing Types	Sensors	Study Participants	Main Results
Grünerbl, A. et al. [53]	detection (bipolar disorder)	long-term (12 weeks)	wearable	phone calls logs, microphone	10 bipolar patients in a rural area psychiatric hospital in Austria aged 18–65+.	.76 average recognition accuracy with Naive Bayes and .97 recall in state change detection.
Maxhuni, A et al. [54]	detection (bipolar disorder)	long-term (12 weeks)	wearable	accelerometer, microphone, questionnaires	Same as Grünerbl, A. et al. [53].	.85 classification accuracy with Bagging
Grünerbl, A. et al. [55]	detection (bipolar disorder)	long-term (12 weeks)	wearable	accelerometer, gps	Same as Grünerbl, A. et al. [53].	State change detection recall .94 and state recognition .80 accuracy with Naive Bayes.
Pagán, J. et al. [56]	forecasting (migraine)	long-term (4–6 weeks)	wearable	dermal activity, temperature, heart rate, spo2	2 patients	Predicted migraine attacks with a max horizon of 52 min with N4SID algorithm.
Andrew G. et al. [57]	forecast-ing/detection (depression)	short-term	social media	Instagram photos	166 total individuals. 71 viable participants aged 19–55.	Correctly identified 70% of all depressed cases with Random Forest.

that Support Vector Machines (SVM) and tree based classifiers seem to be among the most used models for detection which makes sense since they are well researched, easy to understand and included in many machine learning frameworks.

3.1. Study type

From the reviewed literature, we devised several different study types according to the main purpose of the work under consideration. Some of them focus on finding associations between predictor variables and the mental state while others have as main objective to build predictive models to detect different or particular mental states. In the remaining of this section we describe the three different study types: association, detection and forecasting. Association studies focus on finding correlation between the mental state and collected variables which then can be used to for example compare different groups. Detection refers to predicting a current state based on some input data while forecasting means predicting a future state based on some current input data. We will use the term *prediction* to refer to situations in which one wants to *predict* an outcome based on a number of input measures [58] (pp. 10). In regard to this, the term *prediction* can refer to both: detection and/or forecasting. Table 2 presents the *Study Type* of each considered work. The *Main Results* column presents the results and the used methods, if reported.

3.1.1. Association

In these types of studies the researchers aim to determine the relationship between one or more variables and the mental state. They can also look for differences between groups, e.g., a diagnosed group and a control group. The diagnosed group is the one that has the condition of interest (e.g., anxiety) and the control group is composed of participants with similar characteristics to the diagnosed group except that they do not have the condition. An example of this type of study is that of Miranda et al. [39] in which they monitored spontaneous blink rate and heart rate during a relaxed period and after inducing anxiety by asking the participants to give a presentation. They found higher average heart rates after induced anxiety spans on the mild social anxiety disorder group, but found no evidence of increased spontaneous blink rate. They argued this might be because when a person is concentrated, the blink rate decreases, so a different setup should be used to validate this result. In the work of Ferdous et al. [59], they investigated the correlation between verbal interactions and self-reported stress levels. They used the cellphones' microphone to extract the duration of verbal interactions, and found a positive correlation between self-reported stress levels and duration of verbal interactions for 60% of the subjects. This correlation was observed for 90% of highly stressed subjects. Association studies help to understand the relationship *between* variables and how they change within/between different levels. Some works conduct an association study before building their final detection algorithms. Doing so, serves as a basis to have a better understanding of the variables' importance. Some of the statistical methods used to find associations and differences are linear regression models [60], correlation analysis [61], t-tests [61], analysis of variance (ANOVA) [61], etc.

3.1.2. Detection

The aim of these type of works is to detect/recognize the current mental state based on real-time and/or previous data collected through different sensing modalities. Usually, the detection is performed by statistical and machine learning models. The models are built using previously collected and labeled training data and then used to detect the mental state type of newly unseen observations (more on this in Section 4). Unlike association works, detection works often use classification methods such as decision trees, Naive Bayes, etc. to automatically find patterns from the predictor variables. The learned patterns are then used to infer the response variable based on the observable predictor variables. For example, in the work of Carneiro et al. [48] they performed stress detection based on touch screen interactions. They used touch patterns such as intensity and duration to build a J48 decision tree classifier [62] that is able to differentiate between stressed and non-stressed touches with 78% accuracy. Grünerbl et al. [53] used phone call logs and sound collected through the microphone of smartphones to detect manic and depressed states in bipolar disorder patients using a Naive Bayes classifier [63] (pp. 90–97) achieving a recognition accuracy of 76%. In the same work, they proposed a state-change detection algorithm without explicitly recognizing the new state, i.e., detect when there is a change from a default state such that this could trigger a notification to visit a doctor for an exact diagnosis. Given the recent advances in wearable devices, communications and information technologies, it is becoming possible to envision systems capable of detecting mental states in an automatic and timely manner.

3.1.3. Forecasting

In forecasting systems, the objective is to predict a mental state before it actually happens or before its symptoms manifest in a way that can affect negatively the functioning of a person. In general, forecasting is by far more difficult to achieve than detection but with greater potential for health care applications since interventions can be performed in a timely manner, thus, allowing early interventions and more effective treatments. For example, Pagán et al. [56] proposed a multivariate patient-based model to forecast migraine attacks which was capable of predicting the events with an average forecast window of 47 min, and their method is based on state–space models (N4SID) [64]. Siirtola et al. [65] analyzed sleep time data and were able to predict migraine attacks one night prior by using a quadratic discriminant classifier. Another interesting work related to prediction is that of Mormann et al. [66] in which they analyzed electroencephalogram (EEG) signals to analyze the predictability of epileptic seizures. They compared different measures in terms of their suitability

for seizure prediction by testing their ability to discriminate a *preictal* stage from the *interictal* period. The Preictal stage is the time before the seizure and the interictal stage is the time between seizures [67]. Changes were found at least 240 min before seizures. In addition, a study on activation in bipolar disorder has observed early warning signs during a ten day period preceding a shift from depressive to hypomanic state [68]. All these examples give proofs to the existence of critical transition periods, where warning signals occur before substantial changes in the state of ecological and biological systems [69]. Looking into these early warning signs has the potential to both change the management of mental disorders (e.g., bipolar disorder), and to make it possible to develop functional automated monitoring systems [70]. In such scenarios, forecasting systems look more appealing than detection systems since they could trigger promptly interventions allowing the users and caregivers to take actions to mitigate or inhibit an unwanted state.

3.2. Study duration

From the reviewed literature, two main types of works were identified according to the study duration: *short-term* and *long-term*. In some cases, transitions between different mental states can occur in short periods of time. For example, in social anxiety disorder, several transitions between a non-anxious and an anxious state can occur within the same day. On the other hand, transition of states in other mental conditions such as bipolar disorder can take several days. Based on the type of condition, some studies collect sensor data for a couple of days while others require several months. For our classification purpose, we will consider *long-term* studies those that last seven or more days. Examples of short-term anxiety studies are the ones of Miranda et al. [39] and Slater et al. [37]. In the first one, the study lasted 45 min whereas in the second one it consisted of three public talks. Both studies were able to find differences within short time periods. Examples of long-term studies for bipolar disorder are the ones from Faurholt-Jepsen et al. [40] and Grünerbl et al. [53] and both studies lasted 12 weeks and used smartphone sensors such as the microphone, screen state, cell tower ids, sms, etc. The aim of the first study was to look for correlations between bipolar states and sensor values. In the second study they utilized machine learning to recognize bipolar states and achieved a recognition accuracy of 76%. Table 2 provides more examples with their respective details.

3.3. Sensing types

In this section, we present the three sensing modalities that have been used in automatic mental monitoring systems. *External*, *wearable* and *software/social media* sensing. However, the sensor data is not sensing the mental state itself, but can be described as the sensing of a behavior that is emerging from underlying physiological alterations [71]. For example, in the pathology of bipolar disorder, disrupted biological rhythms or disturbance in the circadian rhythms are considered common symptoms [72,73], and circadian rhythm disturbances have been shown in studies of activation in bipolar disorder [71]. The circadian system is an endogenous biological oscillator that synchronies all living light-sensitive organisms with the daily light–dark cycle, and regulates internal recurring rhythms like the sleep–wake cycle, rest–activity patterns and hormone production in humans [74]. Registrations of heart-rate variance and skin conductance gives information about the state of the autonomic nervous system, and this system is connected with both the circadian rhythm [75] and bipolar disorder [76]. Alterations of the autonomic nervous system can also be detected in the human voice, as revealed in several studies on stress [77].

3.3.1. External sensors

External sensors (sometimes also referred as ambient sensors) are generally installed in the environment and attached in a fixed position. Often, they do not require to be in direct contact with the user which can be an advantage since the user is freed of wearing devices. Examples of these type of sensors are video cameras, depth vision cameras, high quality microphones, motion sensors, etc. Sometimes, several sensors are installed in a common place usually called a *smart environment*. It can be a single room, a house or an entire building. In a smart environment, the information gathered by those sensors can be used to understand the context of that environment in order to provide assistance, recommendations and services to the inhabitants. For example, in [78], a sensor network setup that can be easily installed and used in different houses is presented. Their setup consists of 14 state change sensors located in doors, cupboards, refrigerator, etc. They used this sensor network to detect physical activities of daily living. In a follow up work, they used their proposed sensor network for a monitoring system for elderly care [79]. Generally, ambient sensors are of high quality and with many capabilities since they are not power consumption constrained. On the other hand, they can be expensive and difficult to install and calibrate for their proper functioning.

3.3.2. Wearable sensors

Wearable sensors are portable sensing devices that can be worn by the user during her/his daily life routines. Examples of wearable sensors are accelerometers, gyroscopes, heart rate, galvanic skin response monitors, etc. One of the advantages of wearable sensors is that the monitoring can be conducted at any place, unlike with external sensors which are usually restricted to a particular area. In environments with multiple residents, wearable sensors are preferable because with external sensors it becomes difficult to track and distinguish between different users. This was noted by Li et al. [80] who proposed a method to establish an association between wearable sensor data with corresponding external sensor data

Table 3
List of sensing capabilities commonly found in wearable devices.

Sensor	Description	Implementation	Privacy invasiveness
Accelerometer	Measures the acceleration force that is applied to a device.	hardware	low
Magnetometer	Measures the geomagnetic field strength.	hardware	low
Gyroscope	Measures a device's rate of rotation around each of the three physical axes (x, y, and z).	hardware	low
Ambient light	Measures ambient light level.	hardware	low
Proximity	How far away an object is from the phone's screen.	hardware	low
Touch state	Records movement, pressure and size of screen touch interaction.	hardware	medium
Screen state	Records every time the screen is turned on|off.	hardware	medium
Video	Captures video and pictures.	hardware	high
GPS	Provides user location coordinates.	hardware	high
Wifi	Provides data about the BSSID and signal strength of the nearby Wifi access points.	hardware	high
Cell towers	Provides information about the nearby cellphone towers.	hardware	high
Bluetooth	Detects nearby bluetooth capable devices.	hardware	high
Ambient temperature	Measures the ambient room temperature.	hardware	low
Pressure	Measures the ambient air pressure.	hardware	low
Galvanic Skin Response (GSR)	Measures electrical conductance of the skin.	hardware	medium
Electrocardiogram	Measures heart rate activity	hardware	medium
Skin temperature	Measures the temperature of the skin	hardware	medium
Phone call logs	Store phone calls meta-data: type, duration, time, etc.	software	high
App. usage	Stores app. usage information such as start time, time in foreground, etc.	software	high
SMS logs.	Store SMS information: number of sent/received messages, time, etc.	software	high
Social media	Information about uploaded photos, social network posts, likes, comments, etc.	software	high

from the same person in multi-person situations. With the assumption that individuals wear their own personal devices, it becomes straightforward to identify each individual. One of the important aspects of monitoring systems is that they should be sensitive, responsive, adaptive, transparent, ubiquitous and unobtrusive. With the recent advances in electronics miniaturization, it is common to find these type of sensors embedded in everyday life products such as smartphones, smart watches, fitness bracelets, actigraphy bands, tablet PCs, to name a few of the most common ones. Table 3 presents a list of sensing capabilities that can be found in recent wearable devices. The current and foregoing capabilities of these devices make them a natural choice for implementing continuous monitoring systems. Below, we present some of the portable devices with sensing capabilities:

- **Sensor units.** Sensor units are custom made prototypes that consist of a board with several assembled sensors. The sensors send the collected data to a central processing module such as a laptop or a Personal Digital Assistant (PDA). For example, Laerhoven & Cakmakci [81] installed accelerometers to a pair of pants and the board was wired to a laptop so the user had to carry the laptop while performing the experiments for activity monitoring. Another early work that

used a sensor unit was that of Lee & Mase [82]. They used an angular velocity sensor, digital compass sensor and an accelerometer to detect user behaviors. The sensors were connected to a linux-based PDA which is smaller than a laptop, making it less unobtrusive.

- **Smartphones.** With the ever growing popularity and capabilities of smartphones, several research works started to use them as a platform for data collection studies. Typical sensors that can be found in these devices are accelerometers, gyroscopes, ambient light sensors, proximity sensors, GPS, Bluetooth, microphone, video camera, magnetometer, etc. One of the advantages of smartphones is that a large amount the processing can be performed inside the phone and thus, removing the dependency on external processing units. However, performing all the processing locally may reduce battery life [83] and limit the complexity of the type of processing that can be carried out. Since these devices also offer communication, feedback and interaction capabilities, they have the potential to be used not only for monitoring but also as a tool for interventions [28,30] and treatment [23]. For example, Firth et al. [10] conducted a meta-analysis of the effects of psychological interventions through smartphones on anxiety symptoms, and significantly greater reductions in total anxiety scores were observed from smartphone interventions than control conditions. Lu et al. [51] developed a smartphone based system to recognize real-life stress in the human voice, and also demonstrated how fluctuating background noise from the environment can affect the accuracy of the classification. Many smartphone vendors provide Software Development Kits (SDKs) [84] for accessing the underlying hardware capabilities and creating custom programs (also known as applications or apps). At the time of this writing, the two main smartphone operating systems that provide such SDKs are Android [85] and iOS [86].
- **Smart watches/bands/bracelets.** With the further miniaturization of sensors, they began to be embedded in smaller devices such as bracelets and watches. Since bracelets and watches are in constant contact with the users' skin, physiological signals such as heart rate, galvanic skin response, body temperature, etc., can now be collected by such devices. By combining sensor data from smartphones and smart watches, it is possible to capture more details about users' behavior. For example, Zenonos et al. [87] used a wrist band with physiological sensors along with a smartphone to detect worker's mood at work.
- **Actigraphy devices.** Actigraphy is a method to monitor human activity levels using inertial sensors commonly worn on the wrist. Actigraphy devices are often used to monitor sleep patterns [88] and sleep disorders [89] but have also been applied in mental conditions such as in bipolar disorder [71,90], and to distinguish between symptoms of ADHD and bipolar disorder in children [91]. As opposed to smartphones, actigraphy devices have less computational capabilities but their battery life is usually longer which can make them more suitable for long term studies. These devices can register movement but also skin temperature, light, electrodermal activity and so on.
- **Electronic textiles.** Electronic textiles (e-textiles) are fabrics that incorporate electronics [92]. This emerging technology extends the functionality of common fabrics by using electronics directly embedded in the clothes [93–95]. This allows the development of textiles able to sense the environment, react and adapt [92]. López et al. [96] proposed an e-textile and wireless sensor network platform capable of monitoring different physiological signals such as heart rate, body temperature, activity index, and ECG in hospital environments and the measurements are gathered using a wearable smart shirt. Another example of this type of technology is a health-shirt based on e-textile that can monitor physiological parameters like blood pressure [97]. The authors in [98], proposed a t-shirt containing hearth-rate and respiration sensors that was able to recognize mood state with accuracy above 95% in a study on bipolar patients.
- **Flexible sensors.** These sensors benefit from printed electronics technologies and advances in materials offering advantages in terms of mechanical flexibility, thinness and weight reduction allowing them to be worn as patches on the skin [99–101]. In a recent work, traditional tattoo ink was replaced with biosensors that is capable to change color with variations in pH, glucose and sodium [102]. One example application of these type of sensors is a flexible stress monitoring patch [103] that can sense skin conductance, temperature and pulse wave.

3.3.3. Software and social media sensing

Some sensing data can also be obtained by software which does not necessarily requires a hardware sensor. For example, smartphones application usage data such as the time an application was first opened, the duration in the foreground (when the user is using it), etc. can be obtained by software. For example, Ferdous et al. [104] used smartphone application usage data to predict stress levels by grouping applications into four categories: social, utility, entertainment and browser. Then, they computed time and frequency spent by each user on each category per day and used a Support Vector Machine to get the final predictions with an accuracy of 75%. Moreover, phone calls and Short Message Service (SMS) logs can be collected through software APIs. Faurholt-Jepsen et al. [40] analyzed phone call logs of bipolar disorder patients and found that the more severe the depressive symptoms, fewer outgoing calls/day were made, and there were less answered incoming calls/day. LiKamWa et al. [105] used SMS, phone calls, emails, applications usage, web visits, etc., to infer mood states with smartphones achieving accuracies up to 93%. Recent works have also shown that user behavior captured by social media can reveal depression markers [57,106]. Reece and Danforth [57] analyzed uploaded images to Instagram and found that photos posted by depressed users were more likely to be bluer, grayer and darker. Analysis of social media content has also been used to detect stress [107]. These results show the potential of using digital traces to monitor mental health conditions.

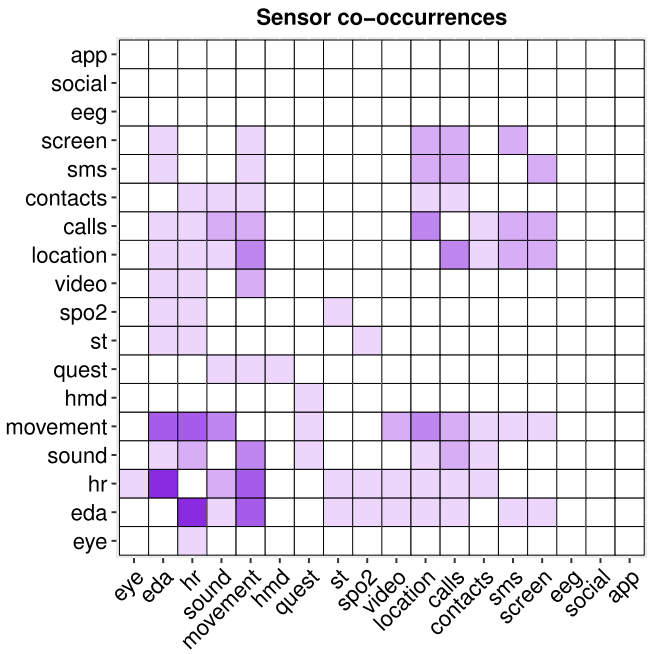

Fig. 2. Sensor co-occurrences. app: application usage, social: social media, eeg: electroencephalogram, screen: smartphone screen state, contacts: phone contact information, calls: phone call logs, location: gps and cell towers, spo2: blood oxygen saturation levels, st: skin temperature, quest: questionnaire, hmd: head mounted device, movement: acceleration, orientation, hr: heart rate, eda: electrodermal activity, eye: eye tracking.

3.3.4. Discussion

As we have seen, there are many heterogeneous types of sensing modalities from which behavioral data can be obtained. Each of those has its own strengths and weaknesses. For example, external sensors can be expensive and limited to a particular physical space. On the other hand, wearable devices are ubiquitous but offer limited computing power. Which sensing modalities to choose will depend on each application.

Table 2 lists the sensors used for each of the works. From this table it seems that inertial and physiological sensors are the most used for stress detection. For social anxiety/phobia, the most common used sensors are based on head mounted devices like virtual reality headsets and eye trackers but also physiological sensors are common, specially heart rate and dermal activity. For bipolar disorder, phone calls and logs have been used in several studies [53,55,108]. Given the limited number of studies, limited number of participants, limited exhaustive comparisons between different sensing modalities, and differences in experimental designs across studies, it is difficult to generalize what sensors will work best for a given application/mental-condition.

The performance of MHMS also heavily depends on user individual characteristics. For instance, a system that only relies on social media behavior analysis will not be suitable for users who do not use social media applications frequently. Furthermore, relying just on physiological sensor data may not be optimal since body changes can occur as a consequence of other factors like food/medication intake, physical exercise, illnesses, and so on. We believe that a robust solution should consider different sensing types and adapt itself to the individual characteristics of each user and the current situation.

From the literature, we can see that each research study has used different combinations of sensors. Fig. 2 shows a heatmap of pair-wise sensor co-occurrences from the 28 works that contained enough information to perform the analysis. This represents how many times two different sensors have been used together in different studies. A white color represents a zero count while a more solid color means higher number of counts. For instance, heart rate (hr) measurements are often used in combination with Electrodermal Activity (EDA) measurements. It can also be seen that movement sensors (accelerometers, gyroscopes) are likely to be combined with other types of sensors. App, social media and EEG were used in isolation (from the sampled works). These insights show some opportunities for future research work, for example, combining social media with physiological sensors.

4. Work flow of mental health monitoring systems

Building automatic monitoring systems for mental health involves a series of steps that go from the *prerequisites* step to the *deployment*. Fig. 3 shows the overall process. As opposed to generic sensing and monitoring pipelines [109,110], mental

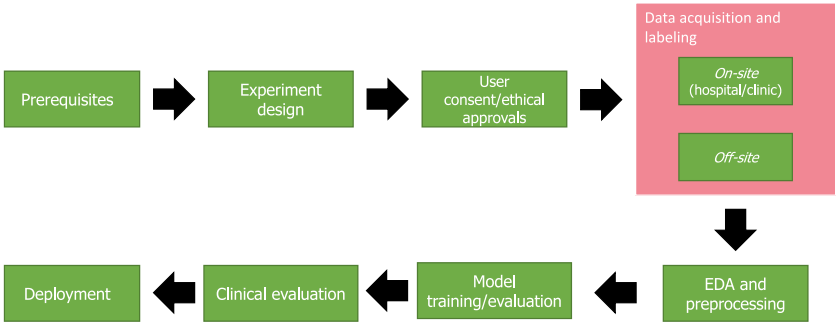

Fig. 3. Overall process for the realization of an automatic sensor based mental health monitoring system.

health applications require some additional steps given their nature. For example, the need of ethical approvals and user consent for data collection is imperative. Another difference is that data collection can be conducted *on-site* (at the hospital or clinic) and in *off-site* (naturalistic) settings. Regarding mental health applications, clinical evaluation is required as opposed to traditional machine learning work flows that may just require evaluation in terms of computational models. In the following subsections each of the phases will be detailed.

4.1. Prerequisites

Given the interdisciplinary nature of these type of studies, constant communication and interaction between the fields is important. As discussed in [111] interdisciplinary research comes with challenges and at the same time great opportunities, especially, in the intersection between medicine and computer science [112]. Therefore, the prerequisites are an important phase in the process. The most important steps/goals to achieve in this phase are:

(i) Common understanding of the task/research: It should be clear what is the goal of the research and what are the hypotheses to address. In an interdisciplinary project it can occur that medical partners and computer science partners are interested in completely different research questions during the same experiment (e.g., treatment efficiency vs. algorithm performance). Thus it is important that both parties understand each others fields to a certain extent, meaning a deeper understanding of the medical part from the technical researchers and vice versa for the medical researchers about some aspects of the technical part.

(ii) Common terms and way to communicate: It often occurs that the same terms are used to describe different things within the fields. For example gold standard in medicine and computer science can have different meanings. Therefore, it is important to conclude on a common terminology and also be aware of issues that could arise from this during the process.

(iii) Legal requirements: Legal requirements such as ethical approval of patient data and dissemination of results should be discussed and sorted out at this stage. Especially ethical approvals and data processing agreements can often take a long time to be approved.

(iv) Common understanding of experimental designs and publication cultures within the fields: Experiments have to be designed in ways that they fulfill both fields' standards. This includes power calculations and standards in clinical testing but also algorithmic evaluation (touching on dataset size, number of negative and positive samples, etc.). For publication, culture common practices should be communicated and discussed. For example, in medicine, conference papers are not common whereas in computer science conference papers can sometimes be more important than journal articles.

(v) Infrastructure, software and hardware requirements: At this phase also requirements such as IT infrastructure, software, hardware, etc., should be decided. Due to common limitations in hospitals, it can often be not possible to use cloud services or specific software. This can influence the design of experiments and algorithms and should be addressed early in the process to avoid later problems or even game stoppers.

Following the presented recommendations will increase the chance of a successful process. It is also important to continue communication and discussion during the complete pipeline to address problems identified later on [112].

4.2. Experiment design

The experiment design step is the cornerstone of the whole process towards the realization of the monitoring system and decisions made during this phase will impact all subsequent phases. During this phase, several aspects need to be defined such as the type of study (Section 3.1), duration (Section 3.2), type of sensors (Section 3.3), target audience, environment (on-site or off-site), data collection protocol, etc. Wearable sensors are becoming common for out-of-lab studies [40,45,49] due to their ubiquitous and portable properties. The type and number of sensors should be selected based on battery consumption, obtrusiveness, relevance and privacy [83]. During the experiment design, sensor sampling rates are also defined, i.e., the

frequency at which the data is collected. Sampling rates affect battery life and storage capacity requirements. High sampling rates will require more computational resources but will be able to capture more fine grained pattern details. However, this can lead to participants to drop out since this will reduce their device battery life [83]. It is also important to implement mechanisms to keep the compliance of participants and reduce dropout rates in long term studies. This can be achieved through different types of incentives [113] such as gift cards, discount cards, letting the participant to keep the device after the study, etc. Other aspects to take into account during this phase are the type of study: longitudinal, cohort, cross-sectional, qualitative, quantitative etc. In MHMS the type of study is usually performed as longitudinal and quantitative. An explanation for the focus on quantitative studies using sensors and machine learning might be the fact that for numerical analysis, data including measurements is needed to build automatic predictive models. The focus on longitudinal cohort studies is related to the fact that sensors typically measure a certain state over a period of time following a specific set of patients. For example, activity measurements of bipolar patients [53,114,115].

One of the key aspects specific of MHMS is that clinicians, health research and patients use standard self-report scales such as the Hamilton Depression Scale (HAMD) and the Young Mania Rating Scale (YRMS), among others and the assessment frequency needs to be defined. For example, in [116] the psychological tests were performed every three weeks to reduce memory effect, which can bias the evaluation outcome. Similarly, in [40], bipolar symptoms were clinically assessed every second week using the Hamilton Depression Rating Scale and the Young Mania Rating Scale.

4.3. Ethical approvals and participants' consent

An important aspect for MHMS before starting the data acquisition phase is the user consent and ethical approvals. The core principle for conducting medical research on human subjects is that all processes must be in accordance to the ethical principles of the World Medical Association Declaration of Helsinki [117]. Initially, the patients' health and best interest rules, and all planned medical research on humans need to get a research protocol approved by an ethical committee before data acquisition can start. Only patients voluntarily agreeing to be investigated can provide data, and a written informed consent must be obtained. Hence it is essential that all included participants' are adequately informed on aims, methods, source of founding, potentially harmful effects and other relevant features of the study before they sign the consent, as well as their right to withdraw the consent at any time. It should also be noted that research on vulnerable patient groups is specifically protected by the Helsinki Declaration, a protection that applies to psychiatric patients, and for that reason, mental health data is often regarded as more sensitive than other health data.

4.4. Data acquisition and labeling

This phase corresponds to the actual data collection by the participants. During the experiment design phase the data collection protocol is defined. For MHMS, the data acquisition and labeling phase often consists of two types: *on-site* and *off-site* (Fig. 3). In the on-site setting, the patient is required to be at the hospital or clinic while the data collection and/or labeling (clinical assessment) takes place. In the off-site setting, the data collection and labeling occurs while the participants perform their daily routines at home/work/etc. For example Faurholt-Jepsen et al. [40] assessed bipolar patients every second week in a on-site setting with face-to-face patient-clinician encounters. Their MONARCA application also allowed them to collect smartphone sensor data in an off-site setting but also conduct self-assessments related to mood, sleep, activity level and so son. The following, is a list of some considerations to take into account regarding the data collection step:

- **Privacy.** Before, during, and after the data acquisition phase, sensitive data about the participants must be protected and secured to safeguard their privacy. Participants need to be informed about the type of data being collected, the intended use, and the implications. Some of the ways to keep participants' privacy is to anonymize the data, encrypt it, transmit it through secure connections, and store it in secure servers. For example, in [108], all the collected speech data was encrypted and transferred securely for analysis to ensure that the integrity and privacy of the collected data is not compromised.
- **Storage.** This is how the data is going to be handled and stored during and after the data collection process. When using smartphones, the data is usually stored in the internal device memory. For smaller devices with limited non-volatile memory like smart watches or wrist bands, the data can be sent periodically to a smartphone via Bluetooth. When the data contains sensitive information about the participant it is usually encrypted. For research analysis, the final storage destination is usually a server. This servers should provide highly secure environments. For example, there are services that provide integrated solutions for managing sensitive data (e.g., health data) including storage [118].
- **Transmission.** This is how the data is sent from the wearable devices to the computer where the analysis will take place. The data can be copied from the wearable device to a local computer through a wired connection (e.g., USB, serial port, etc.) or wireless connection (Bluetooth). The sensor data can also be transmitted to a server by an Internet connection. The advantage of doing so, is that some preliminary data analysis can be made as soon as there is available data instead of waiting until the end of the data collection period; being especially important for long term duration studies. Another advantage of periodically sending the data to a remote server is that it is backed up in case of sensor failures or losing the device. For example, in [116] they periodically send activity data from bipolar patients collected from smartphones to a central server.

- **Energy consumption.** This aspect needs to be considered since energy consumption will have an impact on usability and data quality and quantity. Collecting data from more sensors and at higher sampling rates might provide more information for the analyses but will cause devices to decrease their battery life and perhaps discouraging users to participate. At the predictive phase, some or all preprocessing can be performed locally on the devices or send the data to a central server. In [119], they evaluated energy consumption in terms of amount of information exchange between mobile devices to find a trade off between energy and accuracy but this will be highly dependent on each application.
- **Data labeling.** In order to learn and find patterns, machine learning algorithms require training data. They rely on the amount and quality of data to generate good predictive models. The data labeling phase consists of tagging the sensor data with their corresponding ground truth state. The data labeling process will have an impact on training the final predictive models, thus, it needs to be planned carefully. There exist several approaches for tagging the data. One of them is by videotaping the sessions and then tagging the data by visual inspection [120–122]. This approach is convenient for on-site experiments but not for off-site experiments. A more flexible solution used by Khan et al. [123] to label activities for elderly monitoring was based on a Bluetooth headset combined with speech recognition software to perform the annotations. Another labeling technique is shadowing, i.e., an observer gathers notes from the participants while keeping his/her presence unknown [124]. For off-site mental state monitoring it is common to use self-report questionnaires and/or expert evaluations. The self-report questionnaires are usually filled in using a mobile application which presents the questions in periodic intervals (e.g., once, twice a day, etc.) [87,125]. For mental health-care applications, periodic clinical assessments are conducted in person (at the hospital, clinic, etc.) or by phone [108]. In order to gather more reliable data some works have used a combination of both: in person and phone assessments [55]. Aside from which method to use for labeling the data, the periodicity also needs to be defined. For instance, longer periods of time between self-reports, will lead to less labeled training data.

In addition, generating ground truth labels in MHMS is challenging due to the subjective nature of the phenomenon. Depending on the use case, different techniques have been used:

- **Specialist assessment.** In this case a specialist/doctor assesses the condition of interest of the subject. This can be done on-site [40] or by phone [108]. Some works have used a combination of both to increase the amount of labeled data [55]. The advantage of these methods is that the data is more reliable as compared to self-reports. On the other hand, the process is time consuming and slower since the assessments are conducted every 1–4 weeks. In the case of bipolar disorder, the specialist uses questionnaires such as the Young Mania Rating Scale (YMRS) [126] to assess mania symptoms and the Hamilton Rating Scale (HDRS) [127] to assess depressive symptoms. In the case of social phobia, the Social Phobia Inventory (SPIN) is commonly used [128].
- **Self-report.** Under this scheme the participant is in charge of the labeling process. This is usually conducted by using questionnaires displayed in a mobile application. The questionnaires can be presented at different times during the day and the user answers them according to their current state [45,87]. Some studies also include retrospective questionnaires. For example, the authors in [49] asked the participants before going to sleep to score how stressful their day was. One of the advantages of self-reports is that they can be applied more often since they individual do not need to go to a particular place (hospital, clinic, etc.). A disadvantage of this method is the recall bias and also the participants may skip several of the questionnaires.
- **Event based.** In this case the labels are assigned according to some event. For example, the authors in [129] labeled the data as *stress* the days during exams period and *non-stress* the weeks after. With this method, the participants are relieved from the labeling task. Lu et al. [51] asked the participants to perform several activities with different degrees of difficulty like having a job interview and recruiting participants. Then, the stress levels were labeled according to the activity that took place within the corresponding time. One of the disadvantages of this method is that it requires good synchronization between the different events and the sensor data.

Finally it can be said that the method of data acquisition and ground truth creation is heavily depending on the use case and study type and should be chosen accordingly to fulfill the requirements of the study. In any case data acquisition and labeling should be carefully planed and conducted to avoid problems (missing data, missing labels, label-data synchronization, etc.) later on in the analysis and evaluation part since it is often hard or impossible to redo collection and labeling work.

4.5. Exploratory data analysis and preprocessing

The next step after collecting the data is the Exploratory Data Analysis (EDA) and preprocessing. EDA is an approach that helps to better understand the data. Several visualization methods can be used for this purpose like histograms (Fig. 3), scatter plots, heat maps, box plots, etc. For example, the authors in [114] used a heat map to visualize the activity levels between depressed and non-depressed participants. An EDA can also be useful to detect outliers and identify missing values due to sensor malfunctioning. The preprocessing step consists of applying filters and transformations to the raw data in order to make it suitable for further analysis. Filtering methods can be applied to reduce noise and remove outliers. Examples of transformations are scaling, quantization, binarization, and so on. Another common preprocessing technique is

dimensionality reduction which is used to visualize high dimensional data and to reduce the number of variables to make the model training process more computationally efficient. Two common dimensionality reduction techniques are Principal Component Analysis (PCA) [130] and Multidimensional scaling (MDS) [131]. Many machine learning algorithms require compact representations of the data instead of sensor raw signals. These representations are often in the form of *feature vectors* which are numerical n-dimensional vectors that represent an object. The process of building feature vectors from the original data is called *feature extraction* and is one of the most important steps for mental states prediction. From the literature [45,54,114], some of the common extracted features for mental states detection are: arithmetic mean, standard deviation, min, max, skewness, kurtosis, root mean square, power spectrum density, energy, correlation coefficient, etc. By using this approach, each mental state sample can be represented by its corresponding feature vector. Not all features from the feature vector may be relevant or add a significant value to the prediction. In order to determine the importance of each feature, some *feature selection* algorithms [132] can be applied, thus, reducing the dimensionality of the data. Some feature selection methods are applied before the model training, whereas others, are applied as part of the model training.

4.6. Machine learning model training

With the advent of information technologies and communications, the amount of data that is generated everyday is growing at a fast pace. As of 2017, 2.5 quintillion bytes of data are created every day and with new sensors and devices this growth rate will accelerate even more [133]. Given that the computational power of machines has been increasing in the last years, it is now possible to use that processing power to analyze those huge amounts of data to extract knowledge. Machine learning can be thought of (but not limited to), as a set of computational algorithms that automatically find interesting patterns and relationships from those vast amounts of data. Kononenko, I. and Kukar, M. [134] defined learning as: *"any modification of the system that improves its performance in some problem solving task."*. The result of learning is knowledge that the system can use to solve new problems. An algorithm infers the properties of a given set of data and that information allows it to make predictions about other data that it might see in the future. This is possible because almost all nonrandom data contains patterns which allows a machine to generalize [135] (pp. 3). A prediction model $f(\mathbf{x}, \boldsymbol{\phi}) : y$ is a function that maps a set of input variables $\mathbf{x}$ into a response variable y with a set of parameters $\boldsymbol{\phi}$. Training a machine learning model means finding the parameters' values that optimize some criteria such as: maximizing the prediction accuracy, minimizing model size, minimizing time complexity of prediction, maximize final model comprehensibility, etc. Training a machine learning model requires the data to be in predefined numerical formats (e.g., feature vectors) which is accomplished during the preprocessing step (Section 4.5).

4.6.1. Types of models

There are five main types of machine learning algorithms: supervised learning, unsupervised learning, semi-supervised learning, transfer learning and reinforcement learning.

- **Supervised learning.** In supervised learning, the algorithms are presented with a set of training examples (also known as *instances*). Each instance is a pair of input and output values. The input is often represented as an array that characterizes the instance in a numerical format. This array is typically called the *feature vector*. The output is the value we want to predict, e.g., a category or a real number. For classification problems, the output is usually called the class. The algorithms will try to learn a mapping from the input to the output values (the training phase). Once the algorithm is trained, it will be able to predict unknown output values when provided with new unseen input values. The algorithm performance can be tested by 'hiding' the output values of the instances and letting the algorithm predict them. Then, the predictions can be compared with the real output which is also called the *ground truth*. When the variable to be predicted is categorical it is called *classification* and when the output prediction is quantitative it is called *regression* [58] (pp. 10). For example, consider a set of images each containing a single flag. If we wanted to predict to which country each flag belongs to, this can be stated as a classification problem because the output is categorical, i.e., the values are from a finite set and there is no explicit ordering between them. On the other hand, the problem of predicting house prices based on attributes such as size, location, number of rooms, etc. is a regression one because the price is a quantitative measurement. For instance, if we have a set of signals from physiological sensor readings with their corresponding mental state labels (e.g., stressed and not stressed) those can be used as training instances to train a model. Since the output is categorical (stressed and not stressed), a classification model also called a classifier can be used. Classifiers are the most often used type of machine learning model for mental states detection. This is because detecting the presence/absence of a certain mental state can be formulated as a classification problem by letting the output categorical variable take one of the desired mental states of interest as values. Some classifiers instead of outputting the final category or class, produce as output numerical values such as probabilities from which the resulting category can be inferred (e.g., by selecting the one with highest probability). Some of the supervised learning classifiers that have been used for mental states detection are (see Table 2): Decision trees, AdaBoost, Support Vector Machines (SVM), Naive Bayes, Markov Models, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Artificial Neural Networks, Linear Discriminant Analysis, Hidden Markov Models, etc. Decision trees are very common due to its simplicity and interpretability. Carneiro et al. [48] used a J48 tree which is a type of decision tree to classify stress states achieving accuracies of 78%. Naive Bayes models have also been used in MHMS. Garcia-Ceja et al. [45] used them for stress detection whereas Grünerbl et al. [53] used them for bipolar states prediction.

- **Unsupervised learning.** Unlike supervised learning, the output variable y in unsupervised learning, is not known at training time, thus, the task is to find the different types of categories that may arise naturally from similarities in the *input data*. For example, finding groups of similar users. Clustering algorithms are a common type of unsupervised learning and are used to find groups or hierarchies within the data. Sometimes clustering is used as a preprocessing step for supervised learning. Common clustering algorithms are k-means, k-medoids, DBSCAN, hierarchical clustering, etc. [136]. For a detailed introduction about supervised and unsupervised learning, please refer to [63,137]. Unsupervised learning can also be used as a preprocessing step before using supervised methods. Regarding MHMS, Garcia-Ceja et al. [45] and Xu et al. [138] used k-means clustering to find groups of similar users to build better stress detection models.
- **Semi-supervised learning.** Semi-supervised learning refers to the setting when there are a lot of training data but the output variable (label) is known just for a small percent of samples. Semi-supervised algorithms train models by using both the labeled and unlabeled instances [139]. Semi-supervised learning is relevant for mental states detection since it is often difficult to tag the data with their corresponding ground truth type (class). For example, tagging daily mood states is usually done via questionnaires, but sometimes participants forget to answer them, and in consequence, several days end up untagged. For bipolar disorder, tagging the data involves determining the current patient state (depressed, manic, euthymic, etc.) which requires an expert's evaluation, often at the hospital. This type of issues will limit the amount of available tagged data. Semi-supervised learning seems a promising way to deal with this type of problem, however, it has been under-explored for mental states detection but still, some works have used it before like in [54] for bipolar episodes prediction and in [140] for stress detection.
- **Transfer learning.** In transfer learning, knowledge from similar or different domains where labeled data exists is used to learn something new in a different domain where no or very few labeled training data exists [141]. For example, knowledge gained while learning to recognize horses from images could apply when trying to recognize cows. A more specific example is a system that learned to detect different image categories based on very basic levels like cats, dogs, cars, etc. and then is used to learn how to distinguish if an image contains a certain disease or not [142]. One big advantage of this method is that it can help to overcome the problem of not having enough labeled data. Especially for image related use cases it seems to work quite well [143]. A disadvantage is that if the domains are too different the results are not very good. Furthermore, it is not well researched in other contexts than images and videos. Transfer learning has a great potential to be used for MHMS since it can reduce the amount of labeled data needed which is one of the main challenges in health applications. Maxhuni et al. [140] used transfer learning based on decision trees for stress detection tasks when there is scarce labeled data. Hosseini et al. [144] used deep transfer learning with a convolutional network to perform early Alzheimer's disease diagnostics from brain images.
- **Reinforcement learning.** In reinforcement learning, an agent learns by trial-and error through interactions with its environment [145]. The agent's goal is to maximize its rewards and decide the best action based on the current state. Unlike in supervised learning where the learning algorithm is presented with the inputs and target outputs, in reinforcement learning, an agent explores different actions and selects the one that maximizes the reward. When an agent performs the right/wrong action, it is rewarded or punished accordingly. Over time, it tries to learn what action lead to the best reward. One of the advantages of reinforcement learning is that it does not require a human expert with knowledge about the problem application domain. Reinforcement learning has been previously applied in health care applications. For example, in [146] the authors used reinforcement learning to minimize the frequency and duration of epilepsy seizures.

In practice, different types of algorithms are often combined to get the final predictive models. For example, *unsupervised learning* methods are commonly used as a preprocessing step before building *supervised learning* models. There also exists the distinction between different training schemes: *user-dependent* and *user-independent* (also called general) models. The former, are trained just with data from the specific user under consideration. The latter are trained with data from all other users excluding the target user (the one that is intended to use the system). The advantage of *user-dependent* models is that they capture the specific behavior of each user and often, yield better results but they require a lot of training data for that particular user. On the other hand, *user-independent* models do not require any data for the target user but they might not perform well for 'atypical' users. A representative example of this is from the work of Lu et al. [51] where they compared *user-independent* and *user-dependent models* for stress detection. Their results showed that *user-dependent* models performed much better. Some works have proposed hybrid models that combine the characteristics of *user-independent* and *user-dependent* models to have the best of both [45,51,138].

4.6.2. Machine learning software tools

There are many different machine learning software tools/libraries for training and evaluating models. Some of them are independent software programs while others are extension libraries for a particular programming language. Table 4 shows a list of common machine learning tools. Usually, authors do not report what software and libraries were used but, for example, Weka was used for bipolar detection in [53,55] and scikit-learn was used for anxiety recognition in [47].

Once a machine learning model is trained, we want to estimate how good it will perform in real-life situations, i.e., its performance when making predictions on unseen samples. And that is the topic of the next section.

Table 4
Machine learning software tools/libraries.

Name	Description
Weka [147]	Collection of machine learning algorithms for data mining tasks. It has a graphical user interface but can also be used as an external library for Java programs.
Spark MLlib [148]	This is a scalable machine learning library for massive datasets.
R [149]	Even though R is a programming language, it has a substantial repository with libraries implementing many machine learning algorithms.
scikit-learn [150]	This is a library for the Python programming language with various algorithms for classification, clustering, dimensionality reduction, preprocessing, etc.
Matlab machine learning toolbox [151]	Matlab is a numerical environment which has several toolboxes including one for machine learning.
Keras [152]	A high level deep learning library for python.

4.7. Machine learning model evaluation

The purpose of model evaluation is to assess how well the trained model will be able to generalize, i.e., its estimated performance on unseen samples. One method to estimate the generalization capability of a model is to divide the dataset into two subsets: a training and a testing set. With this scheme, the model is first trained with the training set and its performance is assessed using the testing set and this is called *holdout validation*. The samples are often assigned randomly to one of the training and testing subsets. Machine learning models tend to memorize patterns (overfitting) from the data they were trained with, thus, they will usually perform very well when evaluated with the same data but they may not be able to generalize well to new data. By using holdout validation, we make sure that the trained model does not contain information about the testing set samples and in consequence, this allow us to have better generalization estimates. Some models require some parameter tuning. If these parameters are constantly tuned based on the performance on the testing set, there is the risk of overfitting the model since information from the testing set is being injected into the training process. To avoid this, the entire data set can be divided into three subsets: training, validation, and testing sets. The training set is used to build the model and the validation set is used to tune its parameters. Then, the testing set is used to assess the generalization performance. The split percentage for the three sets depends on each application but it is common to divide the data in 60/20/20, i.e., 60% of the data goes to the training set, 20% to the validation set and the remaining 20% to the testing set. Holdout validation is suitable when we have vast amounts of data. When the amount of data is limited, *k-fold cross validation* is preferred. This method consists of dividing randomly the data into k subsets of approximately equal size. Then k iterations are performed. In each iteration, one of the subsets is used to test the model and the remaining to train the model. The average performance across all iterations is reported. The advantage of this method is that the estimate variance is reduced as k increases but the computational demands also increase. The most typical value for k is 10. When k is equal to the total number of samples in the data set it is called Leave-one-out cross validation (LOOCV).

In MHMS, a common method to evaluate a model's performance of new users is to use as the training set, the data from all other users, and use as testing set the data from the new user. This is also known as a *user-independent* or *general* model and it does not require training data for the target user [153]. For example, Karam *et tal.* [108] used a user-independent approach for bipolar states prediction based on phone calls and Miranda et al. used the same approach for anxiety detection [47]. Another practice is to perform hold-out validation or k-fold cross validation independently for each user. In this case, the training and testing data belong to the same user. These are known as *user-dependent* models [153]. This type of validation was used by Grünerbl et al. [55] for the diagnosis of depressive and manic episodes based on smartphone mobility traces. They divided each patients' data into training (66%) and testing (33%) data. Some MHMS works have made comparisons between user-independent and user-dependent models such as in [45,51] for stress detection. In general, user-dependent models perform better but require more training data.

So far, we have talked about model performance but, how is this performance quantified? There exist several metrics depending on the type of problem (classification, regression, etc.). For classification problems common performance metrics are:

- *Accuracy:* Proportion of correctly classified instances.
- *Sensitivity:* Also called *recall* and is the true positive rate, i.e., the proportion of positives that are correctly classified as such.
- *Specificity:* This is the true negative rate, i.e., the proportion of negatives that are correctly classified as such.
- *Precision:* Also called the positive predictive value and represents the fraction of true positives among those classified as positives:

$$precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (1)$$

- F_1 *score:* This measure is a weighted average of the precision and recall.

It is important to point out that a single metric to evaluate performance should be avoided (often only accuracy is used). Using several different metrics at the same time gives a better overview of the real performance and robustness of a classifier or prediction approach. Furthermore, when the classes are imbalanced, accuracy may be misleading, so its a good practice to also report the other metrics. Imbalances occur when just a small percentage of samples belong to a specific class. For example, if a healthy individual reports the mood state as *normal* or *depressed* during a one year period (365 days), the person may end up with just a few samples of type *depressed* (e.g., 5 days). If a trivial model that *always* classifies every instance as *normal* is built, its accuracy on this example data set will be 0.98 which seems very good but it will fail in detecting *depressed* states. If the state of interest is *depressed*, the specificity will be 0.0 for this classifier. Having imbalanced data is common in mental health monitoring because the events of interest are often rare. There are some machine learning methods to deal with this problem such as under/over sampling, cost sensitive classification [154], SMOTE (a type of oversampling) [155], and so on.

For regression problems, typical performance metrics are the mean squared error, root mean squared error, mean absolute error, correlation coefficients, etc. In mental state detection, the output class is often *ordinal*. This means that it has a natural ordering. For example, depression levels can be modeled with an ordinal variable taking the values {*low, medium, high*}. In this case $low < medium < high$. Typical performance metrics consider all errors as equal: confusing *low* with *medium* has the same error weight as confusing *low* with *high* but clearly, the latter error should be more severely penalized. There are some metrics that can be used for ordinal variables such as the mean squared error, mean absolute error, linear correlation, accuracy within *n*, etc. In [156], several performance measures for ordinal variables were evaluated. From the revised literature, the majority of works for mental state detection do not take into account the ordering of the class, neither for training or evaluating the models.

4.8. Clinical evaluation

When evaluating the clinical applicability of studies on health monitoring systems, the validity of the studies need to be addressed [157]. Imperfections in the design, method or implementation of a study might introduce the risk of bias (systematic errors), consequently reducing the trustworthiness of the results, as the outcome might be dubious due to imprecision. Typical questions for addressing the validity of MHMS studies are: 1) were the ground truth data of high quality?; 2) were the labels (only) self-reported or (always/sometimes) acquired by medical experts?; and 3) were the chosen sensors appropriate to detect the targeted medical condition?. The patient population also needs to be representative for whom the health monitor system is intended for, both, in terms of diagnostic profile and healthcare environment. The inclusion procedure to a study should in addition ensure a representative sample, meaning a studied population with a broad spectrum of disease severities which reflects the clinical reality, and not a carefully selected population for result optimization. Therefore, the inclusion and exclusion criteria for a study need to be critically examined, as well as the characteristics of the included patients. The Cochrane Collaboration, a non-profit society for organizing medical research findings, has developed checklists for assessment of methodological quality and clinical applicability for classification studies [158].

4.9. Deployment

The deployment step consists of making the system as a whole ready for use. Several considerations need to be taken into account to have a robust operating monitoring system. The hardware and software infrastructure plays an important role in deployment. They need to be scalable, reliable, secure, robust, among other properties. After validation, machine learning models are deployed for production use. Sometimes they are translated into other programming languages and/or distributed among many computing units to make them faster and scalable at training and prediction time. Some, or all of the preprocessing can be done locally (e.g., inside a smartphone) or leverage some of the work to a server. For example, consider a monitoring system with a smartphone and a smart watch working together. The smart watch could collect physiological signals like heart rate and galvanic skin response and send them to the smartphone via Bluetooth for further processing. The smartphone would then, be in charge of aggregating the data and preprocessing it for feature extraction. Then, the feature vectors could be sent to a remote server to get the final prediction from a classifier or the classifier could be implemented directly inside the smartphone. These decisions will depend on the specific application and should consider processing time, battery, data transfer, responsiveness, model complexity, etc. Recently, some methods to deal with this decision problem in an automatic manner have been proposed [159]. As pointed out by Giurgiu et al. [160] running applications entirely on a mobile device is limited by the computational resources while running them remotely is limited by the network and latency. They proposed a method to partition application operations on the fly between a mobile device and cloud processing based on the device's CPU load, network conditions or user inputs and achieved significant reductions in power consumption.

The deployed system should also be upgradeable and allow the integration of incremental improvements, e.g., updating the prediction model. According to Coupaye and Estublier [161] when deploying or applying an update, one must take care of consistency constraints, i.e., compatibility between different application versions are to be observed. Thus, temporal constraints need to be considered, i.e, deployment must be done following scheduling and synchronization rules.

5. Research challenges and opportunities

Although there have been great advances in the automatic monitoring of mental health, there are still challenges to be solved to have fully functional and automated systems. These challenges pose several research opportunities to improve aspects within the different phases involved in building automatic mental state monitoring systems. Next, we present a list with some of the research challenges and opportunities for MHMS.

5.1. Data labeling

One of the biggest challenges for developing mental monitoring systems is data labeling, i.e., associating a set of sensor data to the corresponding true mental state at that time span. This is also known as the ground truth data and is required to train the machine learning models. The final performance of the models will depend on the quality of the ground truth data. Obtaining this data requires a lot of effort and is time consuming, for example, for bipolar disorder patients, an expert evaluation is required to identify the current state which requires a visit to the hospital. One method that could be used when there is not enough labeled data is *transfer learning* (Section 4.6.1). In the work of Maxhuni et al. [140], transfer learning was used for stress detection tasks when there is scarce labeled data. Another issue is the subjectivity of the data for self-reported assessments. For off-site studies, participants may be required to report their current state through a questionnaire which depends on self-perception and is prone to bias and capture errors (e.g., mislabeling a state). This type of errors will have a high impact on the predictive models. Whilst there are methods to deal with label noise [162], they have rarely been explored for mental state monitoring. A question that often arises during data labeling is how much data before and after the label time stamp should be considered to belong to that state. For instance, if a participant reports at 9AM his current stress level as high, what start and end time data should we consider as belonging to this stress level? This is a difficult question since we do not know exactly when the high stress level began or ended. A possible partial solution would be to ask the participant to report his/her state within some time span, although, this would lead to some recall bias. Given this challenges, we may end up with scarce labeled data which may not be sufficient to train good classifiers. To alleviate this problem, one possible approach is the use of semi-supervised learning methods (as discussed in Section 4.6.1) which we believe, have good potential for mental health monitoring.

5.2. Inter-user variance

Since each person is unique, physiological and behavioral patterns will vary between users. While this is beneficial for some applications that require user identification/authentication like in behavioral biometrics [163,164], this imposes a problem for mental monitoring systems because *one size, does not fit all*. This means that *user-independent* models tend to perform worse than *user-dependent models* (see Section 4.6.1). Performance differences between *user-independent* and *user-dependent* models can be appreciated from the work for stress detection of Muaremi et al. [49] in which they achieved a 53% and 61% accuracy for the *user-independent* and *user-dependent* models, respectively. Another example is from the work of Zenonos et al. [87] for mood recognition in which they obtained an average accuracy of 62% for the *user-independent* model and 70% for the *user-dependent* model. Previous works in mental state detection have proposed different types of hybrid models that combine the strengths of both types of models [45,51,138]. Still, there is room for potential research to improve this type of hybrid models that adapt to each user's specific characteristics.

5.3. Intra-user variance

This refers to the variability of physiological and behavioral patterns for the same user, specially over time. This may be due to changes in daily routines such as starting in a new job, moving to another city, enrolling in new activities, etc. Intra-user variance can also be triggered by biological changes such as metabolic, change in voice during puberty, illnesses, and in females the intra-user variance has been shown to fluctuate with the menstrual cycle phase in several studies on detecting stress in voice and hearth-rate [77]. Short-term behavioral changes can also be observed within participants. For example, the authors in [165] observed different location change patterns between weekdays and weekends based on GPS, cell towers and WiFi data. Even a user-dependent model that performs well right now, can start to decrease its performance over time due to the intra-user variance. For this type of situations, time adaptive models are highly desirable. The problem when the relation between the input data and the output variable changes over time is formally known as *concept drift* and several adaptation methods have been proposed in the context of supervised learning [166]. To the best of our knowledge, the intra-user variance problem has not yet been addressed in any of the previous mental state monitoring works.

5.4. Sensor data fusion

As we have seen, there are many types of data sources from which physiological and behavioral data can be collected to make inferences (see Table 3). These data sources can be from software or hardware. Each data source has its own format, measuring units, sampling rate, etc., hence, requiring different preprocessing steps. Predictive models require their input data to be in a predefined format. Usually, this is accomplished by *aggregation* (also known as early fusion), i.e., generating

feature vectors which contain the aggregated data from all sensing modalities. These aggregated feature vectors are then used to train the models and make predictions. One of the problems with this approach is that the dimension of feature vectors can increase rapidly, thus, making the training phase slower. This may not represent a problem when models are trained on server side but can take considerable time and battery if models are trained locally in a wearable device (smartphone). Another problem with the *aggregation* approach is that each sensing modality has its own statistical properties and number of features, thus, sensors with lower number of features will be underrepresented. A possible solution is to use a *late fusion* approach which consists of training classifiers for each sensor and then combining their results to get the final classification. In the work of Soleymani et al. [19] about image search intent recognition, they obtained better results by using *late fusion* to combine different sensors types (heart rate, facial expression, user interaction with the system, skin conductance, eye movement and content features of the images).

In a study of Garcia-Ceja et al. [18], the authors used a multi-view stacking approach which is a type of *late fusion* method for activity recognition from different types of sensors and also obtained better results than using *aggregation*. As feature vectors' size increases, there are more chances of having correlated features which can be a problem for some learning algorithms. Some machine learning models have difficulties dealing with missing data or may not work at all. As more data sources are used, the probability of feature vectors with missing values increases. Missing values can be due to sensor failures or a user may decide to turn off some of them to preserve privacy or save battery. All of these represent challenges not just for mental state monitoring systems but for the entire machine learning field. Recently, there have been research advances in sensor data fusion methods [167] to deal with these types of problems that arise in multimodal sensing scenarios. Given the heterogeneous nature of sensors for mental health monitoring, this opens up opportunities to devise new methods for sensor data fusion.

5.5. Integration with other systems

The challenges we have presented so far, mainly focus on machine learning aspects, however, these types of systems are expected to work alongside many other applications and within a larger ecosystem. This involves user databases, system interfaces and administrative tools for the caregivers and physicians, health care intervention and support systems (whether web based [168], smartphone based [30], or others). The adoption and development of communication protocols between different modules and systems are key aspects for having robust and scalable solutions.

5.6. Clinical validation

Clinical validation (the data represents what is it supposed to measure and that it is clinical meaningful) is important as one aim of EMA's is to offer just-in-time interventions based on the collected sensor data. Environmental noise, inaccurate data collection and low adherence may affect the clinical validity of the data. In order to evaluate the clinical validity a clinical expert, representing the gold standard, can label the sensor data as "normal" or "abnormal". However, in the field of mental health there is a continuum, and the threshold for "abnormal" varies across contexts and individuals. Consequently, to label all sensor data by clinical experts is not viable. An alternative, is to detect changes from a base line instead of detecting explicit states as suggested by Grünerbl et al. [53]. According to this, it may be the variation of a mental and behavioral states, and not the current state, that indicates a debut or relapse of a mental health disorder. This is, for example, supported by Chow et al. [169] who reported that variance in GPS data (staying at home) across a day and between days was associated with variance in self-reported mood. In such an approach, the first goal is to understand patterns of sensor-data and to test the clinical validity of these data against a gold standard assessment procedure.

The principal approach within medical research is the evidence based approach; where knowledge is founded on the best available clinical relevant evidence from systematic research [170]. Systematic reviews on the use of machine learning within psychiatric [171] and neuroscientific [172] research emphasizes the need for a theory-driven machine learning approach, that is building on existing knowledge and hypothesis, to achieve relevant results and a deeper understanding. In other words, this is the method to get high quality ground truth data. Furthermore, the innovative explanation algorithm presented by Ribeiro et al. [173] has the potential to accomplish further scientific understanding, as it gives both; insight into, and reasons for the predictions of the classifiers.

6. Limitations of this survey

The primary objective of this paper was to survey general related work that illustrates the current use of multimodal sensing approaches and machine learning for automatic mental health monitoring. Different use cases were included such as anxiety disorder, bipolar disorder, depression, migraine, etc. This paper does not include an exhaustive review for each of the specific cases, but presents a subset of relevant works for several of the cases in the context of mental health monitoring applications. The main results for the different works presented in Table 2, are for illustrative purposes only and should not be thought of as a formal comparison or conclusive. In Section 4, we presented a general overview of the most common used phases in MHMS, however, some works might have used additional or fewer steps, though, we tried to extract the most common ones for clarity and generalization.

7. Conclusions

In this paper, we surveyed state-of-the-art research works on mental state monitoring with a primary focus on those which use sensors to gather behavioral data and machine learning to analyze these data. We identified key characteristics among the reviewed literature and proposed a classification taxonomy (Section 3) that we believe, will help new researches in this field to understand the overall structure of such systems. We also identified the key phases of mental state monitoring systems starting with the experiment design to deployment (Section 4). These include key aspects and considerations for the data collection process, data analysis, and machine learning model training and evaluation. We presented some of the research challenges of MHMS and future opportunities to advance the field. Based on the surveyed literature, the application of multimodal sensing technologies along with machine learning methods represents a great opportunity in the advancement of providing mental health care technology tools for treatment.

Acknowledgment

This publication is part of the INTROducing Mental health through Adaptive Technology (INTROMAT) project, funded by the Norwegian Research Council (259293/o70).

References

[1] H.A. Whiteford, L. Degenhardt, J. Rehm, A.J. Baxter, A.J. Ferrari, H.E. Erskine, F.J. Charlson, R.E. Norman, A.D. Flaxman, N. Johns, et al., Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010, Lancet 382 (9904) (2013) 1575–1586.
[2] G.V. Polanczyk, G.A. Salum, L.S. Sugaya, A. Caye, L.A. Rohde, Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents, J. Child Psychol. Psychiatry 56 (3) (2015) 345–365.
[3] J.M. Twenge, Time period and birth cohort differences in depressive symptoms in the us, 1982–2013, Social Indic. Res. 121 (2) (2015) 437–454.
[4] M. Olfson, B.G. Druss, S.C. Marcus, Trends in mental health care among children and adolescents, New England J. Med. 372 (21) (2015) 2029–2038.
[5] R.C. Kessler, R.G. Frank, The impact of psychiatric disorders on work loss days, Psychol. Med. 27 (4) (1997) 861873.
[6] M.W. DeVries, B. Wilkerson, Stress, work and mental health: a global perspective, Acta Neuropsychiatr. 15 (1) (2003) 44–53.
[7] K. Demyttenaere, R. Bruffaerts, J. Posada-Villa, I. Gasquet, V. Kovess, J.P. Lepine, M.C. Angermeyer, S. Bernert, G. de Girolamo, P. Morosini, G. Polidori, T. Kikkawa, N. Kawakami, Y. Ono, T. Takeshima, H. Uda, E.G. Karam, J.A. Fayyad, A.N. Karam, Z.N. Mneimneh, M.E. Medina-Mora, G. Borges, C. Lara, R. de Graaf, J. Ormel, O. Gureje, Y. Shen, Y. Huang, M. Zhang, J. Alonso, J.M. Haro, G. Vilagut, E.J. Bromet, S. Gluzman, C. Webb, R.C. Kessler, K.R. Merikangas, J.C. Anthony, M.R. Von Korff, P.S. Wang, T.S. Brugha, S. Aguilar-Gaxiola, S. Lee, S. Heeringa, B.-E. Pennell, A.M. Zaslavsky, T.B. Ustun, S. Chatterji, WHO World Mental Health Survey Consortium, Prevalence, severity, and unmet need for treatment of mental disorders in the world health organization world mental health surveys, JAMA 291 (21) (2004) 2581–2590, http://dx.doi.org/10.1001/jama.291.21.2581.
[8] S. Shiffman, A.A. Stone, M.R. Hufford, Ecological momentary assessment, Annu. Rev. Clin. Psychol. 4 (2008) 1–32.
[9] R.C. Moore, C.A. Depp, J.L. Wetherell, E.J. Lenze, Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults, J. Psychiatr. Res. 75 (Suppl. C) (2016) 116–123, http://dx.doi.org/10.1016/j.jpsychires.2016.01.011.
[10] J. Firth, J. Torous, J. Nicholas, R. Carney, S. Rosenbaum, J. Sarris, Can smartphone mental health interventions reduce symptoms of anxiety? a meta-analysis of randomized controlled trials, J. Affect. Disord. 218 (2017) 15–22.
[11] J. Torous, R. Friedman, M. Keshavan, Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions, JMIR mHealth uHealth 2 (1) (2014) e2.
[12] T. Donker, K. Petrie, J. Proudfoot, J. Clarke, M.-R. Birch, H. Christensen, Smartphones for smarter delivery of mental health programs: a systematic review, J. Med. Internet Res. 15 (11) (2013) e247.
[13] D. De, P. Bharti, S.K. Das, S. Chellappan, Multimodal wearable sensing for fine-grained activity recognition in healthcare, IEEE Internet Comput. 19 (5) (2015) 26–35.
[14] O. Lara, M. Labrador, A survey on human activity recognition using wearable sensors, 15 (3) (2013) 1192–1209, http://dx.doi.org/10.1109/SURV.2012.110112.00192.
[15] R.F. Brena, J.P. García-Vázquez, C.E. Galván-Tejada, D. Muñoz-Rodriguez, C. Vargas-Rosales, J. Fangmeyer, Evolution of indoor positioning technologies: A survey, J. Sensors 2017 (2017).
[16] R. LiKamWa, Y. Liu, N.D. Lane, L. Zhong, Can your smartphone infer your mood, in: PhoneSense workshop, 2011, pp. 1–5.
[17] N. Eagle, A.S. Pentland, Reality mining: sensing complex social systems, Personal Ubiquit. Comput. 10 (4) (2006) 255–268.
[18] E. Garcia-Ceja, C.E. Galván-Tejada, R. Brena, Multi-view stacking for activity recognition with sound and accelerometer data, Inf. Fusion 40 (2018) 45–56, http://dx.doi.org/10.1016/j.inffus.2017.06.004, http://www.sciencedirect.com/science/article/pii/S1566253516301932.
[19] M. Soleymani, M. Riegler, P. Halvorsen, Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR '17, ACM, 2017, pp. 251–259, http://dx.doi.org/10.1145/3078971.3078995.
[20] M. Sarchiapone, C. Gramaglia, M. Iosue, V. Carli, L. Mandelli, A. Serretti, D. Marangon, P. Zeppegno, The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis, BMC Psychiatry 18 (1) (2018) 22.
[21] H.A. Chang, C.C. Chang, N.S. Tzeng, T.B. Kuo, R.B. Lu, S.Y. Huang, Heart rate variability in unmedicated patients with bipolar disorder in the manic phase, Psychiatry Clin. Neurosci. 68 (9) (2014) 674–682.
[22] J.O. Berle, E.R. Hauge, K.J. Oedegaard, F. Holsten, O.B. Fasmer, Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression, BMC Res. Notes 3 (1) (2010) 149.
[23] F. Gravenhorst, A. Muaremi, J. Bardram, A. Grünerbl, O. Mayora, G. Wurzer, M. Frost, V. Osmani, B. Arnrich, P. Lukowicz, G. Tröster, Mobile phones as medical devices in mental disorder treatment: An overview, Personal Ubiquit. Comput. 19 (2) (2015) 335–353, http://dx.doi.org/10.1007/s00779-014-0829-5.
[24] J. Nicholas, M.E. Larsen, J. Proudfoot, H. Christensen, Mobile apps for bipolar disorder: a systematic review of features and content quality, J. Med. Internet Res. 17 (8) (2015) e198.
[25] A. Pantelopoulos, N.G. Bourbakis, A survey on wearable sensor-based systems for health monitoring and prognosis, IEEE Trans. Syst. Man Cybern. C 40 (1) (2010) 1–12.
[26] L. Bayındır, A survey of people-centric sensing studies utilizing mobile phone sensors, J. Ambient Intell. Smart Environ. 9 (4) (2017) 421–448.

[27] D.C. Mohr, M. Zhang, S.M. Schueller, Personal sensing: Understanding mental health using ubiquitous sensors and machine learning, Annu. Rev. Clin. Psychol. 13 (2017) 23–47.
[28] J. Stephens, J. Allen, Mobile phone interventions to increase physical activity and reduce weight: a systematic review, J. Cardiovasc. Nurs. 28 (4) (2013) 320.
[29] S.C. Guntuku, D.B. Yaden, M.L. Kern, L.H. Ungar, J.C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Curr. Opin. Behav. Sci. 18 (2017) 43–49.
[30] J. Wang, Y. Wang, C. Wei, N. Yao, A. Yuan, Y. Shan, C. Yuan, Smartphone interventions for long-term health management of chronic diseases: an integrative review, Telemed. e-Health 20 (6) (2014) 570–583.
[31] A. Huguet, S. Rao, P.J. McGrath, L. Wozney, M. Wheaton, J. Conrod, S. Rozario, A systematic review of cognitive behavioral therapy and behavioral activation apps for depression, PLoS One 11 (5) (2016) e0154248.
[32] T.N. Alotaiby, S.A. Alshebeili, T. Alshawi, I. Ahmad, F.E.A. El-Samie, EEG seizure detection and prediction algorithms: a survey, EURASIP J. Adv. Signal Process. 2014 (1) (2014) 183.
[33] A. Mosenia, S. Sur-Kolay, A. Raghunathan, N. Jha, Wearable medical sensor-based system design: A survey, IEEE Trans. Multi-Scale Comput. Syst. 3 (2) (2017) 124–138.
[34] I.L. Kampmann, P.M. Emmelkamp, N. Morina, Meta-analysis of technology-assisted interventions for social anxiety disorder, J. Anxiety Disord. 42 (2016) 71–84.
[35] K. Sampei, M. Ogawa, C.C.C. Torres, M. Sato, N. Miki, Mental fatigue monitoring using a wearable transparent eye detection system, Micromachines 7 (2) (2016) 20.
[36] J. O'brien, P. Gallagher, D. Stow, N. Hammerla, T. Ploetz, M. Firbank, C. Ladha, K. Ladha, D. Jackson, R. McNaney, et al., A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression, Psychol. Med. 47 (1) (2017) 93–102.
[37] M. Slater, D.-P. Pertaub, A. Steed, Public speaking in virtual reality: Facing an audience of avatars, IEEE Comput. Graph. Appl. 19 (2) (1999) 6–9.
[38] H. Grillon, F. Riquier, D. Thalmann, Eye-tracking as diagnosis and assessment tool for social phobia, in: 2007 Virtual Rehabilitation, 2007, pp. 138–145, http://dx.doi.org/10.1109/ICVR.2007.4362154.
[39] D. Miranda, M. Calderón, J. Favela, Anxiety detection using wearable monitoring, in: Proceedings of the 5th Mexican Conference on Human-Computer Interaction, ACM, 2014, p. 34.
[40] M. Faurholt-Jepsen, M. Vinberg, M. Frost, S. Debel, E. Margrethe Christensen, J.E. Bardram, L.V. Kessing, Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder, Int. J. Methods Psychiatr. Res. 25 (4) (2016) 309–323.
[41] C. Holmgard, G.N. Yannakakis, K.-I. Karstoft, H.S. Andersen, Stress detection for ptsd via the startlemart game, in: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE, 2013, pp. 523–528.
[42] M. Price, N. Mehta, E.B. Tone, P.L. Anderson, Does engagement with exposure yield better outcomes? components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia, J. Anxiety Disord. 25 (6) (2011) 763–770.
[43] O.M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, J.M. Ferrandez, Stress detection using wearable physiological and sociometric sensors, Int. J. Neural Syst. 27 (02) (2017) 1650041.
[44] L.B. Leng, L.B. Giin, W.Y. Chung, Wearable driver drowsiness detection system based on biomedical and motion sensors, in: 2015 IEEE SENSORS, 2015, pp. 1–4, http://dx.doi.org/10.1109/ICSENS.2015.7370355.
[45] E. Garcia-Ceja, V. Osmani, O. Mayora, Automatic stress detection in working environments from smartphones' accelerometer data: a first step, IEEE J. Biomed. Health Inf. 20 (4) (2016) 1053–1060.
[46] N. Keshan, P. Parimi, I. Bichindaritz, Machine learning for stress detection from ECG signals in automobile drivers, in: Big Data (Big Data), 2015 IEEE International Conference on, IEEE, 2015, pp. 2661–2669.
[47] D. Miranda, J. Favela, B. Arnrich, et al., Detecting anxiety states when caring for people with dementia, Methods Inf. Med. 56 (1) (2017) 55–62.
[48] D. Carneiro, J.C. Castillo, P. Novais, A. Fernández-Caballero, J. Neves, Multimodal behavioral analysis for non-invasive stress detection, Expert Syst. Appl. 39 (18) (2012) 13376–13389.
[49] A. Muaremi, B. Arnrich, G. Tröster, Towards measuring stress with smartphones and wearable devices during workday and sleep, BioNanoScience 3 (2) (2013) 172–183.
[50] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, G. Riva, Using activity-related behavioural features towards more effective automatic stress detection, PLoS One 7 (9) (2012) e43571.
[51] H. Lu, D. Frauendorfer, M. Rabbi, M.S. Mast, G.T. Chittaranjan, A.T. Campbell, D. Gatica-Perez, T. Choudhury, Stresssense: detecting stress in unconstrained acoustic environments using smartphones, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, 2012, pp. 351–360.
[52] A. Sano, R.W. Picard, Stress recognition using wearable sensors and mobile phones, in: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE, 2013, pp. 671–676.
[53] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, P. Lukowicz, Smartphone-based recognition of states and state changes in bipolar disorder patients, IEEE J. Biomed. Health Inf. 19 (1) (2015) 140–148.
[54] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, E.F. Morales, Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive Mob. Comput. 31 (2016) 50–66.
[55] A. Gruenerbl, V. Osmani, G. Bahle, J.C. Carrasco, S. Oehler, O. Mayora, C. Haring, P. Lukowicz, Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients, in: Proceedings of the 5th Augmented Human International Conference, ACM, 2014, p. 38.
[56] J. Pagán, D. Orbe, M. Irene, A. Gago, M. Sobrado, J.L. Risco-Martín, J.V. Mora, J.M. Moya, J.L. Ayala, Robust and accurate modeling approaches for migraine per-patient prediction from ambulatory data, Sensors 15 (7) (2015) 15419–15442.
[57] A.G. Reece, C.M. Danforth, Instagram photos reveal predictive markers of depression, 2016, CoRR, abs/1608.03282, http://arxiv.org/abs/1608.03282.
[58] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer series in statistics New York, 2009.
[59] R. Ferdous, V. Osmani, J.B. Márquez, O. Mayora, Investigating correlation between verbal interactions and perceived stress, in: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, 2015, pp. 1612–1615.
[60] G.A. Seber, A.J. Lee, Linear Regression Analysis, vol. 329, John Wiley & Sons, 2012.
[61] J. Cohen, P. Cohen, S.G. West, L.S. Aiken, Applied Multiple Regression/correlation Analysis for the Behavioral Sciences, Routledge, 2013.
[62] J. Quinlan, C4. 5: Programs for Machine Learning,
[63] I. Witten, E. Frank, M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2011.
[64] P. Van Overschee, B. De Moor, N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, Automatica 30 (1) (1994) 75–93.
[65] P. Siirtola, H. Koskimäki, H. Mönttinen, J. Röning, Using sleep time data from wearable sensors for early detection of migraine attacks, Sensors 18 (5) (2018).
[66] F. Mormann, T. Kreuz, C. Rieke, R.G. Andrzejak, A. Kraskov, P. David, C.E. Elger, K. Lehnertz, On the predictability of epileptic seizures, Clin. Neurophysiol. 116 (3) (2005) 569–587.

[67] Epilepsy Foundation, Types of seizures https://www.epilepsycolorado.org/wp-content/uploads/2016/01/2-Types-of-Seizures.pdf [online]. (Accessed 17 November 2017).
[68] T. Nakamura, J. Kim, T. Sasaki, Y. Yamamoto, K. Takei, S. Taneichi, Intermittent locomotor dynamics and its transitions in bipolar disorder, in: Noise and Fluctuations (ICNF), 2013 22nd International Conference on, IEEE, 2013, pp. 1–4.
[69] M. Scheffer, J. Bascompte, W.A. Brock, V. Brovkin, S.R. Carpenter, V. Dakos, H. Held, E.H. Van Nes, M. Rietkerk, G. Sugihara, Early-warning signals for critical transitions, Nature 461 (7260) (2009) 53.
[70] A. Bayani, F. Hadaeghi, S. Jafari, G. Murray, Critical slowing down as an early warning of transitions in episodes of bipolar disorder: A simulation study based on a computational model of circadian activity rhythms, Chronobiol. Int. 34 (2) (2017) 235–245.
[71] J. Scott, G. Murray, C. Henry, G. Morken, E. Scott, J. Angst, K.R. Merikangas, I.B. Hickie, Activation in bipolar disorders: a systematic review, JAMA Psychiatry 74 (2) (2017) 189–196.
[72] T. Abreu, M. Bragança, The bipolarity of light and dark: a review on bipolar disorder and circadian cycles, J. Affect. Disord. 185 (2015) 219–229.
[73] L.B. Alloy, T.H. Ng, M.K. Titone, E.M. Boland, Circadian rhythm dysregulation in bipolar spectrum disorders, Curr. Psychiatry Rep. 19 (4) (2017) 21.
[74] S. Panda, J.B. Hogenesch, S.A. Kay, Circadian rhythms from flies to human, Nature 417 (6886) (2002) 329–335.
[75] C. Dibner, U. Schibler, U. Albrecht, The mammalian circadian timing system: organization and coordination of central and peripheral clocks, Annu. Rev. Physiol. 72 (2010) 517–549.
[76] H. Cohen, Z. Kaplan, M. Kotler, I. Mittelman, Y. Osher, Y. Bersudsky, Impaired heart rate variability in euthymic bipolar patients, Bipolar Disord. 5 (2) (2003) 138–143.
[77] C.L. Giddens, K.W. Barron, J. Byrd-Craven, K.F. Clark, A.S. Winter, Vocal indices of stress: a review, J. Voice 27 (3) (2013) 390–e21.
[78] T. Van Kasteren, A. Noulas, G. Englebienne, B. Kröse, Accurate activity recognition in a home setting, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, 2008, pp. 1–9.
[79] T. Van Kasteren, G. Englebienne, B.J. Kröse, An activity monitoring system for elderly care using generative and discriminative models, Personal Ubiquit. Comput. 14 (6) (2010) 489–498.
[80] Z. Li, Z. Wei, L. Huang, S. Zhang, J. Nie, Hierarchical activity recognition using smart watches and RGB-depth cameras, Sensors 16 (10) (2016).
[81] K.V. Laerhoven, O. Cakmakci, What shall we teach our pants? in: Proceedings of the 4th IEEE International Symposium on Wearable Computers, ISWC '00, IEEE Computer Society, Washington, DC, USA, 2000, pp. 77–83, http://dl.acm.org/citation.cfm?id=851037.856531.
[82] S.-W. Lee, K. Mase, Activity and location recognition using wearable sensors, IEEE Pervasive Comput. 1 (3) (2002) 24–32, http://dx.doi.org/10.1109/MPRV.2002.1037719.
[83] G.M. Harari, N.D. Lane, R. Wang, B.S. Crosier, A.T. Campbell, S.D. Gosling, Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges, Perspect. Psychol. Sci. 11 (6) (2016) 838–854.
[84] Software Development Kit definition, 2010, http://www.webopedia.com/TERM/S/SDK.html [online]. (Accessed 17 November 2017).
[85] Android OS, 2017, http://www.android.com [online]. (Accessed 17 November 2017).
[86] iOS https://support.apple.com/ios. (Accessed 17 November 2017).
[87] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, M. Sooriyabandara, HealthyOffice: Mood recognition at work using smartphones and wearable sensors, in: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 1–6, http://dx.doi.org/10.1109/PERCOMW.2016.7457166.
[88] A. Sadeh, C. Acebo, The role of actigraphy in sleep medicine, Sleep Med. Rev. 6 (2) (2002) 113–124.
[89] A. Sadeh, P.J. Hauri, D.F. Kripke, P. Lavie, The role of actigraphy in the evaluation of sleep disorders, Sleep 18 (4) (1995) 288–302.
[90] S.H. Jones, D.J. Hare, K. Evershed, Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder, Bipolar Disord. 7 (2) (2005) 176–186.
[91] G.L. Faedda, K. Ohashi, M. Hernandez, C.E. McGreenery, M.C. Grant, A. Baroni, A. Polcari, M.H. Teicher, Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls, J. Child Psychol. Psychiatry 57 (6) (2016) 706–716.
[92] M. Stoppa, A. Chiolerio, Wearable electronics and smart textiles: a critical review, Sensors 14 (7) (2014) 11957–11992.
[93] L.M. Castano, A.B. Flatau, Smart fabric sensors and e-textile technologies: a review, Smart Mater. Struct. 23 (5) (2014) 053001.
[94] E.R. Post, M. Orth, Smart fabric, or" wearable clothing", in: Wearable Computers, 1997. Digest of Papers., First International Symposium on, IEEE, 1997, pp. 167–168.
[95] F. Axisa, P.M. Schmitt, C. Gehin, G. Delhomme, E. McAdams, A. Dittmar, Flexible technologies and smart clothing for citizen medicine, home healthcare, and disease prevention, IEEE Trans. Inf. Technol. Biomed. 9 (3) (2005) 325–336, http://dx.doi.org/10.1109/TITB.2005.854505.
[96] G. López, V. Custodio, J.I. Moreno, LOBIN: E-textile and wireless-sensor-network-based platform for healthcare monitoring in future hospital environments, IEEE Trans. Inf. Technol. Biomed. 14 (6) (2010) 1446–1458.
[97] Y. t. Zhang, C.C.Y. Poon, C. h. Chan, M.W.W. Tsang, K. f. Wu, A health-shirt using e-textile materials for the continuous and cuffless monitoring of arterial blood pressure, in: 2006 3rd IEEE/EMBS International Summer School on Medical Devices and Biosensors, pp. 86–89, 2006, http://dx.doi.org/10.1109/ISSMDBS.2006.360104.
[98] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, E.P. Scilingo, Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis, IEEE J. Biomed. Health Inf. 18 (5) (2014) 1625–1635.
[99] C. Pang, J.H. Koo, A. Nguyen, J.M. Caves, M.-G. Kim, A. Chortos, K. Kim, P.J. Wang, J.B.-H. Tok, Z. Bao, Highly skin-conformal microhairy sensor for pulse signal amplification, Adv. Mater. 27 (4) (2015) 634–640, http://dx.doi.org/10.1002/adma.201403807.
[100] G. Schwartz, B.C. Tee, J. Mei, A.L. Appleton, D.H. Kim, H. Wang, Z. Bao, Flexible polymer transistors with high pressure sensitivity for application in electronic skin and health monitoring, Nature Commun. 4 (2013) 1859.
[101] L.Y. Chen, B.C.-K. Tee, A.L. Chortos, G. Schwartz, V. Tse, D.J. Lipomi, H.-S.P. Wong, M.V. McConnell, Z. Bao, Continuous wireless pressure monitoring and mapping with ultra-small passive sensors for health monitoring and critical care, Nature Commun. 5 (2014) 5028.
[102] K. Vega, N. Jiang, A. Yetisen, V. Kan, X. Liu, N. Barry, P. Maes, A. Khademhosseini, S. Yun, J. Paradiso, The dermal abyss: Interfacing with the skin by tattooing biosensors, in: Proceedings of the 2017 ACM International Symposium on Wearable Computers, ACM, 2017.
[103] S. Yoon, J.K. Sim, Y.-H. Cho, A flexible and wearable human stress monitoring patch, Sci. Rep. 6 (2016) 23468.
[104] R. Ferdous, V. Osmani, O. Mayora, Smartphone app usage as a predictor of perceived stress levels at workplace, in: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015 9th International Conference on, IEEE, 2015, pp. 225–228.
[105] R. LiKamWa, Y. Liu, N.D. Lane, L. Zhong, MoodScope: Building a mood sensor from smartphone usage patterns, in: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '13, ACM, New York, NY, USA, 2013, pp. 389–402, http://dx.doi.org/10.1145/2462456.2464449.
[106] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: ICWSM, vol. 13, 2013, pp. 128–137.
[107] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, L. Feng, User-level psychological stress detection from social media using deep neural network, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 507–516.
[108] Z.N. Karam, E.M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, M.G. Mcinnis, Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, 2014, pp. 4858–4862.

[109] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, E. Kalogerakis, RisQ: Recognizing smoking gestures with inertial sensors on a wristband, in: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '14, ACM, New York, NY, USA, 2014, pp. 149–161, http://dx.doi.org/10.1145/2594368.2594379.
[110] H. Lu, J. Yang, Z. Liu, N.D. Lane, T. Choudhury, A.T. Campbell, The jigsaw continuous sensing engine for mobile phone applications, in: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, SenSys '10, ACM, New York, NY, USA, 2010, pp. 71–84, http://dx.doi.org/10.1145/1869983.1869992.
[111] A. McMurtry, C. Clarkin, F. Bangou, E. Duplàa, C. MacDonald, N. Ng-A-Fook, D. Trumpower, Making interdisciplinary collaboration work: Key ideas, a case study and lessons learned, Alberta J. Educ. Res. 58 (3) (2012) 461–473.
[112] M. Riegler, M. Lux, C. Griwodz, C. Spampinato, T. de Lange, S.L. Eskeland, K. Pogorelov, W. Tavanapong, P.T. Schmidt, C. Gurrin, D. Johansen, H. Johansen, P. Halvorsen, Multimedia and medicine: Teammates for better disease detection and survival, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 968–977.
[113] H. Gao, C.H. Liu, W. Wang, J. Zhao, Z. Song, X. Su, J. Crowcroft, K.K. Leung, A survey of incentive mechanisms for participatory sensing, IEEE Commun. Surv. Tutor. 17 (2) (2015) 918–943, http://dx.doi.org/10.1109/COMST.2014.2387836.
[114] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J.T. rresen, T. Nordgreen, K.J. Oedegaard, O.B. Fasmer, Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients, in: Proceedings of the 9th ACM on Multimedia Systems Conference, MMSys'18, ACM, New York, NY, USA, 2018, pp. 472–477, http://dx.doi.org/10.1145/3204949.3208125.
[115] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Torresen, T. Nordgreen, K.J. Oedegaard, O.B. Fasmer, Motor activity based classification of depression in unipolar and bipolar patients, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018, pp. 316–321, http://dx.doi.org/10.1109/CBMS.2018.00062.
[116] V. Osmani, A. Maxhuni, A. Grünerbl, P. Lukowicz, C. Haring, O. Mayora, Monitoring activity of patients with bipolar disorder using smart phones, in: Proceedings of International Conference on Advances in Mobile Computing & Multimedia, MoMM '13, ACM, New York, NY, USA, 2013, pp. 85:85–85:92, http://dx.doi.org/10.1145/2536853.2536882.
[117] G.A. of the World Medical Association, et al., World medical association declaration of Helsinki: ethical principles for medical research involving human subjects, J. Am. College Dent. 81 (3) (2014) 14.
[118] University of Oslo, Services for sensitive data (TSD), http://www.uio.no/english/services/it/research/sensitive-data/. (Accessed 21 May 2018).
[119] D. Gordon, J.H. Hanne, M. Berchtold, A. Shirehjini, M. Beigl, Towards collaborative group activity recognition using mobile devices, 18 (3) 326–340, http://dx.doi.org/10.1007/s11036-012-0415-x.
[120] D. Miranda, J. Favela, C. Ibarra, N. Cruz, Naturalistic enactment to elicit and recognize caregiver state anxiety, J. Med. Syst. 40 (9) (2016) 192, http://dx.doi.org/10.1007/s10916-016-0551-0.
[121] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine, in: International Workshop on Ambient Assisted Living, Springer, 2012, pp. 216–223.
[122] P.V.K. Borges, N. Conci, A. Cavallaro, Video-Based human behavior understanding: A survey, IEEE Trans. Circuits Syst. Video Technol. 23 (11) (2013) 1993–2008, http://dx.doi.org/10.1109/TCSVT.2013.2270402.
[123] A.M. Khan, Y.-K. Lee, S. Lee, T.-S. Kim, Accelerometers position independent physical activity recognition system for long-term activity monitoring in the elderly, Med. Biol. Eng. Comput. 48 (12) (2010) 1271–1279.
[124] S. Sen, D. Chakraborty, V. Subbaraju, D. Banerjee, A. Misra, N. Banerjee, S. Mittal, Accommodating user diversity for in-store shopping behavior recognition, in: Proceedings of the 2014 ACM International Symposium on Wearable Computers, ISWC '14, ACM, New York, NY, USA, 2014, pp. 11–14, http://dx.doi.org/10.1145/2634317.2634338.
[125] M. Sysoev, A. Kos, M. Poganik, Noninvasive stress recognition considering the current activity, 19 (7) (2015) 1045–1052, http://dx.doi.org/10.1007/s00779-015-0885-5.
[126] R. Young, J. Biggs, V. Ziegler, D. Meyer, A rating scale for mania: reliability, validity and sensitivity, Br. J. Psychiatry 133 (5) (1978) 429–435.
[127] M. Hamilton, Development of a rating scale for primary depressive illness, Br. J. Soc. Clin. Psychol. 6 (4) (1967) 278–296.
[128] K.M. Connor, J.R. Davidson, L.E. Churchill, A. Sherwood, R.H. Weisler, E. Foa, Psychometric properties of the Social Phobia Inventory (SPIN): New self-rating scale, Br. J. Psychiatry 176 (4) (2000) 379–386.
[129] G. Bauer, P. Lukowicz, Can smartphones detect stress-related changes in the behaviour of individuals?, in: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, 2012, pp. 423–426, http://dx.doi.org/10.1109/PerComW.2012.6197525.
[130] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Phil. Mag. J. Sci. 2 (11) (1901) 559–572.
[131] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (3–4) (1966) 325–338.
[132] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[133] I.B.M. Marketing Cloud, 10 Key Marketing Trends for 2017 https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN. (Accessed 17 November 2017).
[134] I. Kononenko, M. Kukar, Machine Learning and Data Mining: Introduction to Principles and Algorithms, Horwood Publishing, 2007.
[135] T. Segaran, Programming Collective Intelligence: Building Smart Web 2.0 Applications, "O'Reilly Media, Inc.", 2007.
[136] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[137] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[138] Q. Xu, T.L. Nwe, C. Guan, Cluster-based analysis for personalized stress evaluation using physiological signals, 19(1) (2015) 275–281, http://dx.doi.org/10.1109/JBHI.2014.2311044.
[139] O. Chapelle, B. Schölkopf, A. Zien, et al., Semi-Supervised Learning, MIT press Cambridge, 2006.
[140] A. Maxhuni, P. Hernandez-Leal, L.E. Sucar, V. Osmani, E.F. Morales, O. Mayora, Stress modelling and prediction in presence of scarce data, Journal of biomedical informatics 63 (2016) 344–356.
[141] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[142] K. Pogorelov, M. Riegler, S.L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P.T. Schmidt, P. Halvorsen, Efficient disease detection in gastrointestinal videos–global features versus neural networks, Multimedia Tools Appl. 76 (21) (2017) 22493–22525.
[143] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[144] E. Hosseini-Asl, M. Ghazal, A. Mahmoud, A. Aslantas, A. Shalaby, M. Casanova, G. Barnes, G. Gimel'farb, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by a 3d deeply supervised adaptable convolutional network, Front. Biosci. (Landmark edition) 23 (2018) 584.
[145] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: A survey, J. Artificial Intelligence Res. 4 (1996) 237–285.
[146] J. Pineau, A. Guez, R. Vincent, G. Panuccio, M. Avoli, Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach, Int. J. Neural Syst. 19 (04) (2009) 227–240.
[147] University of Waikato, Weka: Data mining software in Java, 2017, http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 17 November 2017).
[148] Spark MLlib, 2017, http://spark.apache.org/docs/1.2.0/mllib-guide.html. (Accessed 17 November 2017).
[149] R programming language, 2017, https://cran.r-project.org/. (Accessed 17 November 2017).
[150] scikit-learn, 2017, http://scikit-learn.org. (Accessed 17 November 2017).
[151] MathWorks, Matlab machine learning toolbox, 2017, https://www.mathworks.com/solutions/machine-learning.html. (Accessed 17 November 2017).

[152] Keras: the python deep learning library https://keras.io/, (Accessed 17 November 2017).
[153] E. Garcia-Ceja, R. Brena, Building personalized activity recognition models with scarce labeled data based on class similarities, in: J.M. Garcıa-Chamizo, G. Fortino, S.F. Ochoa (Eds.), Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information, in: Lecture Notes in Computer Science, vol. 9454, Springer International Publishing, 2015, pp. 265–276, http://dx.doi.org/10.1007/978-3-319-26401-1_25.
[154] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al., Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (1) (2006) 25–36.
[155] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.
[156] L. Gaudette, N. Japkowicz, Evaluation methods for ordinal classification, in: Y. Gao, N. Japkowicz (Eds.), Advances in Artificial Intelligence, in: Lecture Notes in Computer Science, vol. 5549, Springer Berlin Heidelberg, 2009, pp. 207–210, http://dx.doi.org/10.1007/978-3-642-01818-3_25.
[157] J. Reitsma, A. Rutjes, P. Whiting, V. Vlassov, M. Leeflang, J. Deeks, et al., Assessing methodological quality, in: Cochrane handbook for systematic reviews of diagnostic test accuracy version, 2009, 1.
[158] P.F. Whiting, M.E. Weswood, A.W. Rutjes, J.B. Reitsma, P.N. Bossuyt, J. Kleijnen, Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies, BMC Med. Res. Methodol. 6 (1) (2006) 9.
[159] J. Liu, E. Ahmed, M. Shiraz, A. Gani, R. Buyya, A. Qureshi, Application partitioning algorithms in mobile cloud computing: taxonomy, review and future directions, J. Netw. Comput. Appl. 48 (2015) 99–117.
[160] I. Giurgiu, O. Riva, G. Alonso, Dynamic software deployment from clouds to mobile devices, in: ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing, Springer, 2012, pp. 394–414.
[161] T. Coupaye, J. Estublier, Foundations of enterprise software deployment, in: Software Maintenance and Reengineering, 2000. Proceedings of the Fourth European, IEEE, 2000, pp. 65–73.
[162] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey 25 (5) (2014) 845–869.
[163] C. Bo, L. Zhang, X.Y. Li, Q. Huang, Y. Wang, Silentsense: Silent user identification via touch and movement behavioral biometrics, in: Proceedings of the 19th Annual International Conference on Mobile Computing & Networking, MobiCom '13, ACM, New York, NY, USA, 2013, pp. 187–190, http://dx.doi.org/10.1145/2500423.2504572.
[164] K.O. Bailey, J.S. Okolica, G.L. Peterson, User identification and authentication using multi-modal behavioral biometrics, Comput. Secur. 43 (2014) 77–89.
[165] S. Servia-Rodríguez, K.K. Rachuri, C. Mascolo, P.J. Rentfrow, N. Lathia, G.M. Sandstrom, Mobile sensing at the service of mental well-being: a large-scale longitudinal study, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 103–112, http://dx.doi.org/10.1145/3038912.3052618.
[166] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 44.
[167] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44.
[168] M.L. Ybarra, W.W. Eaton, Internet-based mental health interventions, Mental Health Serv. Res. 7 (2) (2005) 75–87.
[169] P.I. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, L.E. Barnes, B.A. Teachman, Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students, J. Med. Internet Res. 19 (3) (2017) e62.
[170] D.L. Sackett, Evidence-based medicine, in: Seminars in Perinatology, vol. 21, Elsevier, 1997, pp. 3–5.
[171] Q.J. Huys, T.V. Maia, M.J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications, Nature Neurosci. 19 (3) (2016) 404.
[172] J.W. Krakauer, A.A. Ghazanfar, A. Gomez-Marin, M.A. MacIver, D. Poeppel, Neuroscience needs behavior: correcting a reductionist Bias, Neuron 93 (3) (2017) 480–490.
[173] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.