Motor Activity Based Classification of Depression in Unipolar and Bipolar Patients

Enrique Garcia-Ceja
University of Oslo
enriqug@ifi.uio.no

Michael Riegler
University of Oslo and Simula@OsloMet
michael@simula.no

Petter Jakobsen
University of Bergen
petter.jakobsen@helse-bergen.no

Jim Torresen
University of Oslo
jimtoer@ifi.uio.no

Tine Nordgreen
University of Bergen
tine.nordgreen@helse-bergen.no

Ketil J. Oedegaard
University of Bergen
ketil.joachim.odegaard@helse-bergen.no

Ole Bernt Fasmer
University of Bergen
Ole.Fasmer@uib.no

***Abstract*—Wearable sensors measuring different parts of people's activity are a common technology nowadays. Data created using these devices holds a lot of potential besides measuring the quantity of daily steps or calories burned, since continuous recordings of heart rate and activity levels usually are collected. Furthermore, there is an increasing awareness in the field of psychiatry on how these activity data relates to various mental health issues such as changes in mood, personality, inability to cope with daily problems or stress and withdrawal from friends and activities. In this paper we present the analysis of a unique dataset containing sensor data collected from patients suffering from depression. The dataset contains motor activity recordings of 23 unipolar and bipolar depressed patients and 32 healthy controls. We apply machine learning to classify patients into depressed and nondepressed. For evaluation of the algorithms, leave one patient out validation is performed. The best results achieved are an F1 score of 0.73 and a MCC of 0.44. The overall findings show that sensor data contains information that can be used to determine the depression status of a person.**

***Index Terms*—Depression, Bipolar Disorder, Depressive Disorder, Major, Motor Activity, Machine Learning, Artificial Intelligence**

I. Introduction

The use of on body sensors to monitor personal health has become quite normal these days. Modern people are collecting vast amounts of data every day, for purposes such as increasing quality of life, supervise their fitness levels, or even to change bad habits. New advances in ubiquitous computing and wearable sensors have attracted new potential applications in the field of psychiatry. New devices like smartphones and smartwatches could allow the continuous monitoring of patients in an unobtrusive manner and providing timely interventions when necessary. Since 2010 mental health related problems are the main cause for years lived with disability worldwide. Depression is number one of the most frequent disorders and the current trend is indicating that the prevalence will increase even more in the coming years [1]–[3]. Dealing with depression can be demanding since it can create physically, economically and emotionally problems often leading to problems with work and sick leaves [4]. Mental health problems are related to disturbance in internal biological systems [5]. These are complex systems and, because relations between the sensor data and the mood are not well understood yet, changes within these systems are difficult to detect. Research indicates that early warning signals occurs in critical transition periods preceding abrupt noticeable changes of state, and is usually indicated by a phenomena called critical slowing down [6]. Critical slowing down indicates that the system becomes slower and slower in recovering from small disturbances, i.e., a reduced ability to restore itself to its original condition [7]. Unipolar depression and bipolar disorder are episodic disorders, where the pathologic state and the healthy state might be understood as representing different stable states separated by sudden changes [8]. It is important to point out that most depressions people experience are not of this kind. In context to this; the state of biological systems are somehow measurable through recordings of motor-activity. Evidence indicates that a depressive state is associated with reduced daytime motor-activity, as well as increased nighttime activity when comparing to healthy controls [9]. Reduced motor-activity is likewise reported in bipolar depressions, besides increased variability in activity levels compared to others [10]. Activity and movement measurements have become an emerging topic in the field of mental health. Several studies use sensors to measure patients movements over time and connect them to diagnosis or self reports [11], [12]. Usually, in these studies the data is analyzed using standard linear and nonlinear statistical methods. Reported findings include increased autocorrelations and variances as indicators of a critical slowing down [6], and increased skewness is also observed [13]. As one can easily see, such data also holds potential for machine learning applications which is used more and more in the context of psychiatry and psychology [14]–[17]. In this paper we use a unique dataset containing motor activity of depressed and nondepressed bipolar patients to perform depression classification using machine learning. Two different machine learning approaches are compared to classify, namely a Random Forest classifier and a Deep Neural Network (DNN). Given the class imbalance nature of the dataset, we also evaluated two different data balancing techniques (random oversampling and SMOTE). The performance of the classification is evaluated using leave one participant

out validation. The reported performance with an F1 score of 0.73 and a MCC of 0.44 indicates that motor activity can give information about the depression state of a patient which can be important and useful for patient monitoring and follow up.

The main contributions are, (i) Experiments on an open dataset containing sensor data of patients with depression and control participants; (ii) Detailed analysis of motor activity in depressed patients; (iii) Detailed evaluation of machine learning algorithms for depressed patient classification; (iv) Detailed failure analysis of the applied algorithms.

II. Depression background

Depression is a severe mental disorder with characteristic symptoms like sadness, the feeling of emptiness, anxiety and sleep disturbance, as well as general loss of initiative and interest in activities [18]. Additionally, features like the feeling of guilt or worthlessness, reduced energy, concentration problems, suicidality and psychotic symptoms might be present. The severity of a depression is determined by the quantity of symptoms, their seriousness and duration, as well as the consequences on social and occupational function [19]. Depressions are also common in Bipolar disorder, another severe psychiatric disorder. The main difference between unipolar depression and bipolar disorder is the periodic occurrence of mania in the latter, a state associated with inated self-esteem, impulsivity, increased activity, reduced sleep and goal-directed actions [20]. Both diseases are genetic disorders, and might be understood as a genetic vulnerability to the environment disturbing the internal biological state and potentially trigger mood episodes [21]. Depression is associated with disrupted biological rhythms caused by environmental disturbance like seasonal change in daylight, alteration of social rhythms due to for instance shiftwork or longitude traveling; besides linked to lifestyles associated with diurnal rhythms inconsistent with the natural daylight cycle [22], [23]. The appearance of depressive symptoms relates furthermore to physical health issues, medical side effects, life events and social factors, besides alcohol and substance abuse [18], and such factors might also potentially cause symptoms of depression in all humans. The global lifetime prevalence of depression is roughly 15% [19], but the incidences of episodes with a severity level not meeting the requirements for a depressive diagnosis are far more prevalent [24]. Actigraph recordings of motor activity are considered an objective method for observing depression, although this topic is far from exhaustive studied within psychiatric research [10].

Depression diagnosis. The Montgomery-Åsberg Depression Rating Scale (MADRS) is used to grade the current severity of an ongoing depression [25]. Clinicians rate ten items relevant for depression based on observation and conversation with the patient, and the sum score (0-60) states the severity of the depression. Scores below 10 are classified as absence of depressive symptoms [26], and scores above 30 indicate a severe depressive state [27].

III. Related Work

In the last years, sensor devices have experienced remarkable improvements in terms of size reduction and energy consumption. These advances have propelled the creation of new sets of devices like smart-watches and smartphones which have huge sensing capabilities. Recently, researchers have been looking into new ways to use those devices to monitor users in a continuous and unobtrusive manner. These technologies have also allowed the collection of objective data through prolonged periods of time. By using machine learning methods to analyze that data, it is possible to understand users' context and behaviors, thus, allowing the implementation of practical systems such as fitness tracking, indoor location, activity recognition and so on. These ubiquitous technologies have a huge potential in the mental healthcare field. Specifically, there have been some preliminary works that make use of sensing devices to automatically monitor depression in patients. For example O'brien *et al.* [28] monitored adults with late-life depression using a wrist-worn activity measurement unit and they found that their physical activity was reduced compared to healthy controls. In another study of Faurholt-Jepsen *et al.* [29] they monitored patients with bipolar disorder using smartphone sensors and found that the more severe the depressive symptoms, the less answered incoming calls and fewer outgoing calls were registered. They also found that depressed patients moved less based on cell tower IDs. Our work differs from the previous two in that we use machine learning on statistical features computed from the sensor data to automatically classify *depressed* v.s. *nondepressed* persons. Systematic reviews on the use of actigraphy in studies on depression indicates that a depressive state is associated with reduced daytime motor-activity, as well as increased nighttime activity when comparing to healthy controls [9]. Similarly, reduced motor-activity is likewise reported in bipolar depressions, besides increased variability in activity levels compared to others [10]. Some studies have also used machine learning to detect depressive states. For example, Grünerbl *et al.* [30] used smartphone data such as acceleration, sound, location, etc. to classify depressed and manic states in bipolar patients achieving an accuracy of 76% with a Naive Bayes classifier. The difference with our work, is that we aim to detect if a person is a depressed patient or not which is more diagnosis focused. It has also been shown that social media has potential to identify depression, for example, by analyzing uploaded photos to Instagram [31]. Table I shows a summary of related works about depression monitoring using sensors.

IV. Data Collection

The dataset collected used in this paper is public available and can be found at **http://datasets.simula.no/depresjon/** or directly downloaded from **https://doi.org/10.5281/zenodo.1219550**. The dataset was originally collected for the study of motor activity in schizophrenia and major depression [11]. Motor activity was monitored with an actigraph watch worn at the right wrist (Actiwatch, Cambridge Neurotechnology Ltd, England, model AW4). The actigraph watch measures activity

TABLE I
RELATED WORKS ON SENSOR BASED MONITORING OF DEPRESSION.

Reference	Sensing Device	Description
O'brien, J.T. *et al.* [28]	wrist-worn actigraph	They found that physical activity was reduced in adults with Late-life depression compared to healthy controls and showed slower fine motor movements.
Faurholt-Jepsen, M. *et al.* [29]	smartphone	They found that in patients with bipolar disorder the more severe the depressive symptoms, the less answered incoming calls, fewer outgoing calls and patients moved less.
Grünerbl, A. *et al.* [30]	smartphone	Used smartphone data to classify depressed and manic states in bipolar patients achieving a 76% recognition accuracy.
Maxhuni, A *et al.* [32]	smartphone	Used audio, motor activity and questionnaires to classify mood in bipolar patients with an accuracy of 85%.
Reece, A. G. *et al.* [31]	Instagram	Correctly identified 70% of all depressed cases with Random Forest from uploaded photos to Instagram.
Burton *et al.* [9]	wrist-worn actigraph	This systematic review identified that a depressive state is associated with reduced daytime motor-activity, as well as increased nighttime activity when comparing to healthy controls.
Scott *et al.* [10]	wrist-worn actigraph	This systematic review identified that reduced motor-activity is associated with bipolar depressions, besides increased variability in activity levels compared to healthy controls.

levels. The sampling frequency is 32Hz and movements over 0.05 g are recorded. A corresponding voltage is produced and is stored as an activity count in the memory unit of the actigraph watch. The number of counts is proportional to the intensity of the movement. Total activity counts were continuously recorded in one minute intervals.

This dataset consists of actigraphy data collected from 23 unipolar and bipolar depressed patients (condition group): 8 bipolar depressed and 15 unipolar depressed. Five subjects were hospitalized during their data collection period, and 18 were outpatients. The severity level of the ongoing depression was rated by a clinician on the Montgomery-Asberg Depression Rating Scale [25] at the beginning and conclusion of the motor-activity recordings. In addition, the dataset contains actigraphy data from 32 non-depressed contributors (control group), consisting of 23 hospital employees, 5 students and 4 former patients without current psychiatric symptoms. Furthermore, the sleep and wake cycles for all patients are available but they were not considered within the analysis of this paper.

The actigraph devices were used by the study participants for an average of 12.6 days in the control and condition groups. The total number of collected days was 693 comprising 402 days in the control group and 291 in the condition group (Figure 1). Note that the actigraph files might contain more days but only the first n days were considered in our analysis. Where n is the number of days reported in the *days* column from the *scores.csv* file.

Feature Extraction. In order to train machine learning classifiers, each participants' day was characterized by a feature vector which was computed by extracting a set of features on a per day basis from the activity level. The extracted features were the mean activity level, the corresponding standard deviation and the percentage of events with no activity i.e, $activitylevel = 0$. The features were then normalized between 0 and 1.

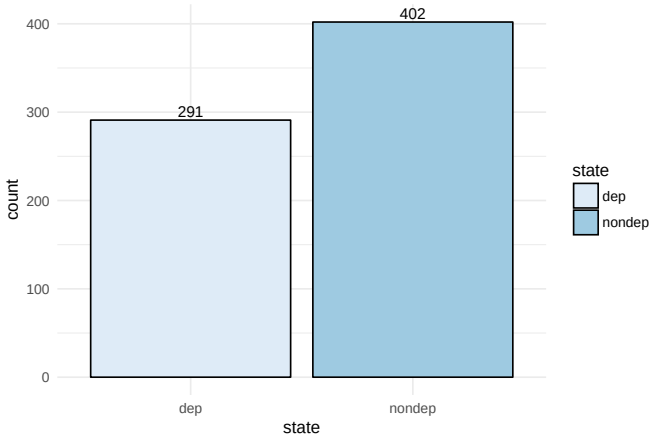

Fig. 1. Distribution of days by state.

V. EXPERIMENTS AND RESULTS

Since the main objective is to classify a person as depressed or nondepressed, we proposed the following approach to accomplish this. Each participant collected data for d_i consecutive days where d_i represents the number of days collected by participant i. Then, we extracted statistical features (see Feature Extraction section) from each day resulting in d_i feature vectors per participant. To avoid overfitting, have a better generalization estimate and stability of findings, we adopted a Leave-One-User-Out validation strategy, i.e., for each user i use all the data from all other users $\neq i$ to train the classifier and test them using the data from user i. In order to obtain the final classification (depressed or nondepressed) for a particular person, a vector of predictions p is first obtained from the trained classifier. Each entry of p corresponds to the prediction of a particular day. The final label is obtained by majority voting, i.e., output the most frequent prediction from p.

Given the imbalanced nature of the data, we conducted our experiments using different class balancing techniques [33]. Specifically, we used two oversampling techniques which consist of augmenting the minority class data. Firstly, we used *random oversampling* which consists of duplicating data points selected at random. Secondly, we used SMOTE [34] which consists of creating new synthetic samples that are generated at random from similar neighboring points. Furthermore, we tested two different machine learning classifiers namely: Random Forest [35] and a Deep Neural Network (DNN). For comparison purposes, we also trained a *Baseline classifier* which outputs a random class based just on their prior probabilities.

For the here presented experiments we use different metrics described in Table II. Well chosen metrics depicting different aspects of the performance are the basis for a qualitative good evaluation. For a medical dataset it is also recommended to apply weighting of the metrics by the number of samples in the respective classes and report the weighted average which can help with the problem of imbalanced classes many medical datasets are suffering from.

The results of the respective algorithms are presented in Table III for the Random Forest and in Table IV for the DNN. The Figures 2-3 show the resulting confusion matrices for both methods.

Looking at the results there are several observations and findings to be made. Regarding the overall performance, Random Forest outperforms DNN in terms of classification performance. This can best be seen in the MCC value which is 0.44 for the Random Forest with SMOTE compared to 0.39 for the DNN with random oversampling (a MCC of 0 would indicate a random decision).

Looking at the more specific metrics reveals that Random Forest is better in detecting depression with a recall of 0.69 compared to 0.56 of DNN but the DNN performs better in detecting *nondepressed* (0.78 to 0.75) which is also depicted in the sensitivity and specificity metrics. Concluding, it can be said that both methods have their strength and weak points and for future work it might be interesting to combine the two to achieve better results. These findings also show how important it is to look at different metrics at the same time.

In terms of class balancing techniques it is nicely shown how the performance improves applying different methods. The baseline is outperformed by all techniques (even without oversampling) but oversampling overall increases significantly looking at the MCC metrics (MCC of 0.31 for the best no oversampling performance compared to a MCC of 0.44 for the best SMOTE run). For Random Forest, the best performance increase is achieved with SMOTE outperforming all other techniques. Random oversampling does not improve the results significantly and even performs worse than no oversampling. In the DNN experiments, Random oversampling performed the best.

All in all the classification performance is promising with weighted F1 score of 0.73 and a MCC of 0.44. Nevertheless, there is still improvement potential and it is also important to understand the missclassifications of the methods which will be discussed in the following.

A. Understanding of Missclassifications

As one can see, the classification results are promising and way above random or majority class baselines. Nevertheless, there is improvement potential and for a clinical application also improvement is needed. To better understand why missclassification happened and how it could be improved in future work we conducted a more detailed analysis of the results. As basis for the analysis we looked at the results from all performed experiments to get a better understanding. For the control group it was observed that each of the controls was

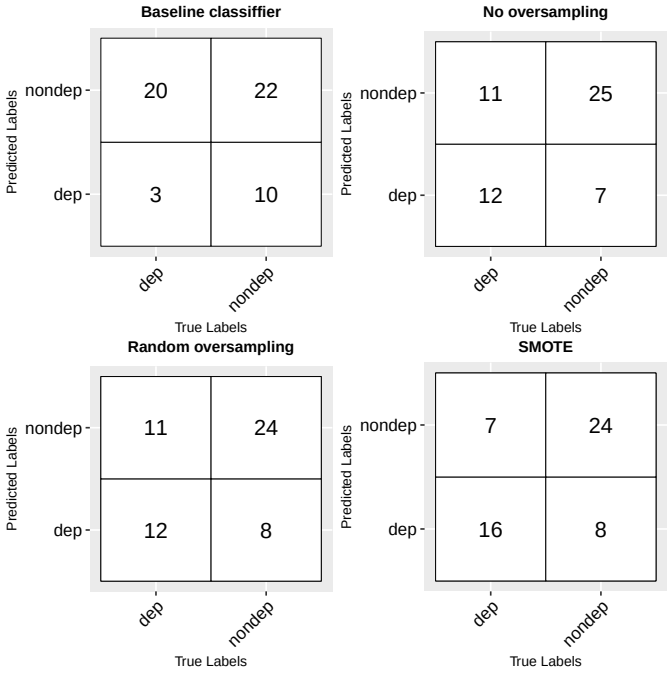

Fig. 2. Random Forest confusion matrices. SMOTE performs best.

correctly classified in at least one run. This is an indication that this is mainly related to the classification method and could most probably be improved with applying more sophisticated classification algorithms or oversampling. What would improve the performance could be for example applying time series analysis using recurrent neural networks or Hidden Markov Models. The most interesting finding from the failure analysis was that out of the seven missclassified condition

TABLE II
USED METRICS FOR THE HERE PRESENTED EXPERIMENTS. TRUE POSITIVE (TP) NUMBER OF CORRECT CLASSIFIED POSITIVE SAMPLES, TRUE NEGATIVE (TN) NUMBER OF CORRECT CLASSIFIED NEGATIVE SAMPLES, FALSE POSITIVE (FP) NUMBER OF NEGATIVE SAMPLES WRONGLY CLASSIFIED AS POSITIVE, FALSE NEGATIVE (FN) NUMBER OF POSITIVE SAMPLES INCORRECTLY CLASSIFIED AS NEGATIVE.

Metric	Description
Precision (PREC)	Depicts the fraction of true positives among those classified as positives. It is also called positive predictive value.
Recall/Sensitivity (REC/SEN)	This metric is the ratio of correctly classified relevant samples among all relevant samples in the dataset.
Accuracy (ACC)	Represents the percentage of correctly classified positive and negative samples. Can be misleading for imbalanced datasets and should be interpreted in combination with other metrics.
Specificity (SPEC)	SPEC, also called true negative rate depicts the classifiers performance in terms of correctly classified negative samples.
Matthews correlation coefficient (MCC)	MCC is a balanced measure which takes into account TP, FP, TN and FN. It can show the classifiers performance even if the classes are imbalanced.
F1-score (F1)	Harmonic mean of precision and recall.

TABLE III
RANDOM FOREST CLASSIFICATION RESULTS.

	PREC			REC/SEN			SPEC			F1			MCC	ACC
	+	-	w.a	+	-	w.a	+	-	w.a	+	-	w.a	overall	overall
Baseline	0.23	0.52	0.40	0.13	0.69	0.45	0.69	0.13	0.36	0.17	0.59	0.41	-0.21	0.45
No oversampling	0.63	0.69	0.67	0.52	0.78	0.67	0.78	0.52	0.63	0.57	0.73	0.67	0.31	0.67
Random oversampling	0.60	0.68	0.65	0.52	0.75	0.65	0.75	0.52	0.62	0.56	0.72	0.65	0.28	0.65
SMOTE	0.67	0.77	0.73	0.69	0.75	0.73	0.75	0.69	0.72	0.68	0.76	0.73	0.44	0.73

+:depressed, **-**:nondepressed, **w.a**:weighted average.

TABLE IV
DEEP NEURAL NETWORK CLASSIFICATION RESULTS.

	PREC			REC/SEN			SPEC			F1			MCC	ACC
	+	-	w.a	+	-	w.a	+	-	w.a	+	-	w.a	overall	overall
Baseline	0.09	0.5	0.33	0.04	0.69	0.42	0.69	0.04	0.31	0.06	0.58	0.36	-0.33	0.42
No oversampling	0.67	0.67	0.67	0.43	0.84	0.67	0.84	0.43	0.60	0.53	0.75	0.66	0.30	0.67
Random oversampling	0.68	0.72	0.71	0.56	0.81	0.71	0.81	0.56	0.67	0.62	0.76	0.70	0.39	0.71
SMOTE	0.65	0.71	0.69	0.56	0.78	0.69	0.78	0.56	0.65	0.60	0.75	0.69	0.35	0.69

+:depressed, **-**:nondepressed, **w.a**:weighted average.

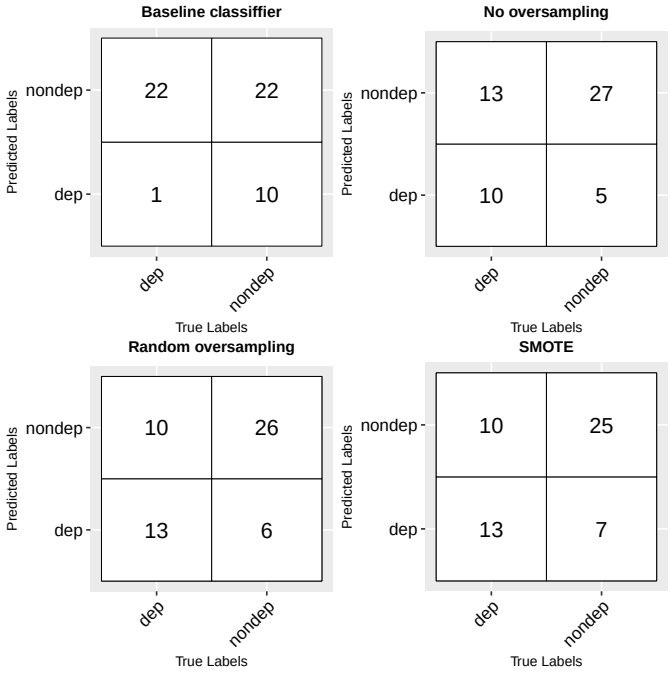

Fig. 3. DNN confusion matrices. SMOTE performs best.

patients of the best run (seen in the bottom right confusion matrix of Figure 2) five were never correctly classified in any of the runs (condition 3, 4, 7, 8 and 13). As a first step to understand this better we looked into the static data (age, gender, work, marriage, education, MADRS scores and days in care). In these variables no connection between the different patients could be observed. Further investigation included calculating several statistics (kurtosis, median absolute deviation (MAD), mean, number of zeros, skewness, standard deviation (std)) to gain a better understanding. Using these statistics we could observe that the patients seemed to be very active compared to the average of the other condition patients (low number of no activity periods) ranging somewhere between controls and conditions. This was also confirmed trough a higher std compared to the other patients. Looking at the MADRS scores of the patients also revealed that these patients had in general a large drop in the MADRS scores from admission day to release. This could be related to better working medication or therapy that was not captured in the dataset. For future investigations it might be interesting to categorize patients in more than just depressed and nondepressed classes based on the MADRS scale since it appears that the five always missclassified patients have something in common within them but not with the controls or the other conidition patients. Finally, for a future dataset we would also recommend to collect MADRS scores for the controls since this could also contain some interesting information for a better understanding of the classification performance.

VI. CONCLUSION AND FUTURE WORK

In this paper we presented approaches to classify depressed and nondepressed patients using motor activity collected via actigraphy devices. The experiments were performed on a dataset that we also share publicly to enable reproducibility and comparability of results. The presented methods included a Random Forest and a DNN approach including the comparison of several data balancing techniques. The best results achieved are an F1 score of 0.73 and a MCC of 0.44. In addition to the classification experiments we also conducted an analysis of the missclassified cases concluding that the depression classification problem might be too complex for simple binary classes and adding more classes based on the MADRS scale could be a solution. Overall the results are very promising and indicate that hospitals could explore the use of sensors to monitor patients during their daily life. Furthermore, it shows that sensor data collected by patients outside of the health care environment holds potential and should be explored further. For future work, we would like to investigate more sophisticated classification algorithms such as recurrent neural networks and hidden Markov models. Furthermore, we would also want to take into account sleep-wake cycles as additional

information when training the predictive models. Finally, it is important to point out for future work that the main utility of actigraphy in this population (patients at risk of entering a depression) could be more measuring state fluctuations within an individual, rather than case-control trait comparisons. One reason is that even patients with recurrent major depressive episodes do not spend most of their time in a depression, but detecting a shift into depression earlier than current practice could enable many improved therapeutics or prevention strategies.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] G. V. Polanczyk, G. A. Salum, L. S. Sugaya, A. Caye, and L. A. Rohde, "Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents," *Journal of Child Psychology and Psychiatry*, vol. 56, no. 3, pp. 345–365, 2015.

[2] J. M. Twenge, "Time period and birth cohort differences in depressive symptoms in the us, 1982–2013," *Social Indicators Research*, vol. 121, no. 2, pp. 437–454, 2015.

[3] M. Olfson, B. G. Druss, and S. C. Marcus, "Trends in mental health care among children and adolescents," *New England Journal of Medicine*, vol. 372, no. 21, pp. 2029–2038, 2015.

[4] M. A. Vammen, S. Mikkelsen, Å. M. Hansen, J. P. Bonde, M. B. Grynderup, H. Kolstad, L. Kærlev, O. Mors, R. Rugulies, and J. F. Thomsen, "Emotional demands at work and the risk of clinical depression: A longitudinal study in the danish public sector," *Journal of occupational and environmental medicine*, vol. 58, no. 10, pp. 994–1001, 2016.

[5] E. M. Marco, E. Velarde, R. Llorente, and G. Laviola, "Disrupted circadian rhythm as a common player in developmental models of neuropsychiatric disorders," in *Neurotoxin Modeling of Brain Disorders — Life-long Outcomes in Behavioral Teratology*. Springer, 2016, pp. 155–181.

[6] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, p. 53, 2009.

[7] A. Bayani, F. Hadaeghi, S. Jafari, and 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," *Chronobiology international*, vol. 34, no. 2, pp. 235–245, 2017.

[8] O. B. Fasmer, H. S. Akiskal, J. R. Kelsoe, and K. J. Oedegaard, "Clinical and pathophysiological relations between migraine and mood disorders," *Current Psychiatry Reviews*, vol. 5, no. 2, pp. 93–109, 2009.

[9] C. Burton, B. McKinstry, A. S. Tătar, A. Serrano-Blanco, C. Pagliari, and M. Wolters, "Activity monitoring in patients with depression: a systematic review," *Journal of affective disorders*, vol. 145, no. 1, 2013.

[10] J. Scott, G. Murray, C. Henry, G. Morken, E. Scott, J. Angst, K. R. Merikangas, and I. B. Hickie, "Activation in bipolar disorders: a systematic review," *JAMA psychiatry*, vol. 74, no. 2, pp. 189–196, 2017.

[11] J. O. Berle, E. R. Hauge, K. J. Oedegaard, F. Holsten, and O. B. Fasmer, "Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression," *BMC research notes*, vol. 3, no. 1, p. 149, 2010.

[12] N. Razavi, H. Horn, P. Koschorke, S. Hügli, O. Höfle, T. Müller, W. Strik, and S. Walther, "Measuring motor activity in major depression: the association between the hamilton depression rating scale and actigraphy," *Psychiatry research*, vol. 190, no. 2, pp. 212–216, 2011.

[13] V. Guttal and C. Jayaprakash, "Changing skewness: an early warning signal of regime shifts in ecosystems," *Ecology letters*, vol. 11, no. 5, pp. 450–460, 2008.

[14] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *International Journal of Neural Systems*, vol. 27, no. 02, p. 1650041, 2017.

[15] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from smartphones' accelerometer data: a first step," *IEEE journal of biomedical and health informatics*, vol. 20, no. 4, pp. 1053–1060, 2016.

[16] N. Keshan, P. Parimi, and 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.

[17] L. B. Leng, L. B. Giin, and W. Y. Chung, "Wearable driver drowsiness detection system based on biomedical and motion sensors," in *2015 IEEE SENSORS*, Nov 2015, pp. 1–4.

[18] NICE, "National institute for health and clinical excellence. depression in adults: recognition and management. NICE guideline CG90," https://www.nice.org.uk/guidance/cg90, 2009, [last visited, February 14, 2018].

[19] S. Pilling, I. Anderson, D. Goldberg, N. Meader, C. Taylor, T. G. D. Groups *et al.*, "Depression in adults, including those with a chronic physical health problem: summary of nice guidance," *BMJ*, vol. 339, no. 10.1136, 2009.

[20] R. Hirschfeld, "Differential diagnosis of bipolar disorder and major depressive disorder," *Journal of affective disorders*, vol. 169, pp. S12–S16, 2014.

[21] D. Landgraf, M. J. McCarthy, and D. K. Welsh, "The role of the circadian clock in animal models of mood disorders." *Behavioral neuroscience*, vol. 128, no. 3, p. 344, 2014.

[22] L. B. Alloy, T. H. Ng, M. K. Titone, and E. M. Boland, "Circadian rhythm dysregulation in bipolar spectrum disorders," *Current psychiatry reports*, vol. 19, no. 4, p. 21, 2017.

[23] W. Bechtel, "Circadian rhythms and mood disorders: are the phenomena and mechanisms causally related?" *Frontiers in psychiatry*, vol. 6, p. 118, 2015.

[24] P. M. Lewinsohn, A. Solomon, J. R. Seeley, and A. Zeiss, "Clinical implications of "subthreshold" depressive symptoms." *Journal of abnormal psychology*, vol. 109, no. 2, p. 345, 2000.

[25] S. A. Montgomery and M. Asberg, "A new depression scale designed to be sensitive to change." *The British journal of psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

[26] C. Hawley, T. Gale, and T. Sivakumaran, "Defining remission by cut off score on the madrs: selecting the optimal value," *Journal of affective disorders*, vol. 72, no. 2, pp. 177–184, 2002.

[27] M. J. Müller, H. Himmerich, B. Kienzle, and A. Szegedi, "Differentiating moderate and severe depression using the montgomery–åsberg depression rating scale (madrs)," *Journal of affective disorders*, vol. 77, no. 3, pp. 255–260, 2003.

[28] 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," *Psychological medicine*, vol. 47, no. 1, pp. 93–102, 2017.

[29] M. Faurholt-Jepsen, M. Vinberg, M. Frost, S. Debel, E. Margrethe Christensen, J. E. Bardram, and L. V. Kessing, "Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder," *International journal of methods in psychiatric research*, vol. 25, no. 4, pp. 309–323, 2016.

[30] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2015.

[31] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *CoRR*, vol. abs/1608.03282, 2016. [Online]. Available: http://arxiv.org/abs/1608.03282

[32] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–66, 2016.

[33] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.