Comparative Study on Classification of Depressed and Non-Depressed Subjects using Tree based Ensemble Models

Shivam Kasat, Sonali Agarwal, Sanjay Kumar Sonbhadra, and Narinder Singh Punn

Indian Institute of Information Technology - Prayagraj, Allahabad, India
http://www.iiita.ac.in

Abstract. With increasing technology in the 21st-century use of sensors in our daily life is also increasing, Most of the devices we use in our daily life are equipped with different types of sensors like gyro sensors, motion detection sensors, etc. These sensors provide useful data which can be used not only to count the calories burnt or the number of steps taken but also can be used to measure mental health issues such as changes in mood, personality, inability to cope up with daily problems, stress, etc. In this research paper, we will try to analyze a unique dataset containing sensor data collected from patients suffering from unipolar and bipolar depression. The data contains motor activity measurement of 23 unipolar or bipolar depressed subjects and the control group includes 32 subjects having no sign of depression. We will use synthetic minority oversampling technique (SMOTE) to balance this data-set and will use different tree-based ensemble models (Machine learning) on this dataset to identify if a person is suffering from depression or not. Moreover, we will try to classify the subject into three different classes namely no-depression, mild-depression, and severe-depression based on MADRS (Montgomery-Åsberg Depression Rating Scale scores). This research work can be used as a basis for new applications in smartwatches to warn users regarding their mental health.

Keywords: Depression · Mental Health · Machine learning · Ensemble Model.

1 Introduction

Nowadays carrying a smartphone or wearing a smartwatch is no big deal, and sensors in these smart devices continuously record data which is then fed to complex algorithms or machine learning models to make decisions to improve human life. For example, these smart devices nowadays remind us to drink water based on the number of steps we walked or the calories we burnt. These smart devices can be used for more advanced tasks like in our case warning a user if he is taking too much stress and notifying him to relax or have a deep sleep for some time. Not only this, data collected from such sensors can be put to greater use in medical science like identifying mental health issues, change in usual activity, changes in mood [16].

Mental illness is a health condition involving changes in emotions [or mood. It is a disturbance in the brain resulting in changes in a person's mood, thinking, or behavior [10]. Depression is one of the major mental health issues and is expected to increase in upcoming years [19]. Other common mental health issues include anxiety disorder, mood disorder, schizophrenia disorder. Unipolar and Bipolar depressions are episodic mood disorders and different states of mind like healthy or pathological states can be explained with different stable states separated by abrupt changes between them [3].

Some relation is being observed between motor activity and depression, Increased night time activity and reduced day time activity is observed in the depressive state compared to normal [2]. This can also be observed in case of stress as during stress or anxiety our sleeping pattern tends to change.

Machine learning is being used widely in medical science from Brain tumor detection [18] to detection of COVID-19 infection using blood exams [1]. This research paper will use machine learning to achieve two objectives. Our first objective will aim to classify the subject as depressed or non-depressed using motor-activity data and the second objective will use the same dataset for multi-class classification, classifying the subject into one of the three classes no-depression, mild-depression or severe depression. Prior to performing this classification, we will use the Montgomery-Åsberg Depression Rating Scale (MADRS) [13] values to label them accordingly.

Ensemble model is a type of machine learning model that we are going to use, Ensemble model combines several base models (called weak learners) to produce one optimal predictive model. In this experiment, we will be using tree-based ensemble models for classification tasks including Gradient-based ensemble model (GBM), XGBoost, AdaBoost, CatBoost, and will try to compare their performance also. Evaluation and comparison will be based on Accuracy, Precision, F1-score, and Recall score. Furthermore, we will try to shed some light on the application of this research paper for improving human life.

The rest of the paper is organized as follows: Section 2 describes the background and related work, Section 3 present the dataset details, Section 4 show experiments conducted and outcomes, Section 5 presents the conclusion and future work.

2 Related Work

2.1 Mental Health and Monitoring Systems

Continuous research is being done in the field of Mental Health and Monitoring Systems, We will discuss a few of them here which are related to our work in this research paper.

E. Garcia-Ceja et al. [8]. did some research and survey study on recent work in MHMS using machine learning and classified different works based on: Study duration (short or long term), sensor types (software/external/wearable/social media), or study type (association/detection/forecasting). Association studies help us to understand the relationship between different variables and include methods like linear regression, analysis of variance, t-tests, and correlation analysis. Detection studies deal with recognizing the different states of Mental health using methods like classification models or clustering algorithms. Forecast studies deal with the prediction of continuous variables like MADRS score, epileptic seizures. Different sensors used by researchers include wearable ones like smartwatches, smart-band, smartphones, external sensors like cameras, and microphones continuously monitoring patient's activities like sleeping habits, eating habits, hours of sleep, tone of talk, talking hours, etc. Some researchers used software/social media for collecting data like Twitter or Instagram to collect data as it has been observed that nowadays people share most of their feeling on social media platforms[9].

Studies related to depression and bipolar disorder. One such study by O'Brien, J.T. et al. [15] was an association study about bipolar disorder and depression. twenty-nine controls (healthy subjects) and thirty conditions (subjects with depression) were under study as subjects of the experiment and the goal was to find the relation between depression and physical activity referred to as late-life depression (LLD) in the paper and they concluded that physical activity of subjects suffering from LLD was lower compared to healthy subjects.

Another study of type detection was done by Grunerbl, A. Et al. [9] where the experiment subjects included ten bipolar patients between the age of 18 and 65 years old. They used phone calls and microphone data of patients and achieved an accuracy of 76% with precision and recall over 97% for recognition of bipolar state detection. Along with microphone data they also used GPS and accelerometer data and achieved recognition accuracy of 70% with accelerometer and 80% with GPS.

The same data was used by Maxhuni, A. [11] Et al but along with microphone and GPS data, they also used data obtained through a questionnaire on the participants. They applied various machine learning algorithms and their best average accuracy was 85.57%.

Rochelle C. Mehl conducted observed 438 children, To study the sleeping characteristics of children with a bipolar mood disturbance behavioral profile and found out that Thirteen out of 438 participants fit the pediatric bipolar disorder profile i.e, these children demonstrated significant disturbances with poorer sleep efficiency and longer periods of sleep slow-wave sleep than their matched counterparts [12].

Faurhalt-Jepson, M. Et al. [4] did an association study on 29 bipolar participants regarding actions on their smartphone like daily usage, number of received and sent text messages, number of incoming calls, etc. They found a strong correlation between the recorded information and the mental health of the patients.

Andrew G. Et al. [17] used Instagram photos to study depression and used machine learning on the data. A total of 43,950 photos of 166 participants was used as a dataset. Statistical features extracted from photos using color analysis, metadata, and face detection are used as input to machine learning algorithms and achieved an accuracy of 70%, using social media platforms for depression detection was an interesting study as nowadays social media is used by most of the people to share their feelings.

Mowery, D. et al. [14] did a similar study but used Twitter posts instead of Instagram posts. Their aim was to classify if the Twitter post contains some evidence of depression. Features extracted from tweets included syntax of tweets, usage of emoticons, and sentiment in text. A good accuracy has been achieved in identifying tweets with no evidence of depression but the results were not satisfactory for other cases.

Garcia-Ceja et al. [7] also published a paper where he used machine learning algorithms like Random forests, Decision trees, and Dense neural networks to classify subjects as depressed or non-depressed. The dataset of experiments included twenty-three unipolar and bipolar patients and thirty-two healthy controls and achieved an f1-score of 0.73 using random forests and 0.7 using Dense neural networks.

Frogner et al. [5] used the same dataset and used deep learning to classify subjects into depressed and non-depressed subjects and also did multi-classification of subjects into different categories using 1-D convolutional neural networks. We in this research paper will be working on the same dataset but using tree-based ensemble models. We will try to find out if similar or better results can be achieved using our approach or not.

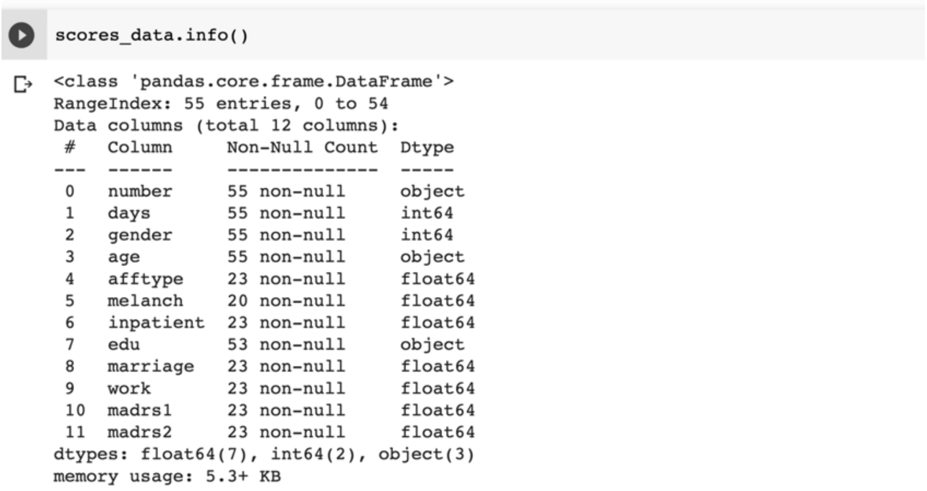
2.2 Dataset

The dataset was collected to study motor activity in patients suffering from schizophrenia and major depression. The dataset consists of motor activity data collected using an actigraph watch which patients were made to wear in their right hand. The actigraph watch measures the daily activity level of patients at a frequency of 32Hz and movement over 0.05g is recorded. Whenever the watch senses a moment, the corresponding voltage level is produced and is stored in the memory unit of the actigraph watch. The counts stored in the watch are proportional to the intensity of the movement. Activity count was continuously recorded at an interval of one minute.

The dataset is named as depressjon dataset [6] and is freely and openly available to everyone. The dataset consists of two folders: one consisting of motor-activity data of each patient suffering from some sort of unipolar or bipolar disorder and the other consisting of motor-activity data of healthy subjects not suffering from any depression disorder. The first folder named “condition” has 23 unipolar and bipolar patients and the other folder named “control” contains the data of 32 healthy controls. For each subject, a different CSV file is provided containing the actigraph data collected over time. In addition, a separate file is provided named as “scores.csv” which contains the MADRS score of each subject along with the following columns:

- number: A unique ID for each subject

- days: Number of days of data collection
- gender: 1 for female 2 for male
- age: Age of the participant
- afftype: affliction type, 1-Bipolar II, 2-Unipolar depressive, 3-Bipolar I
- inpatient: whether patient is inpatient (1) or outpatient(2)
- edu: years of education completed
- marriage: 1-married/cohabiting, 2-single
- work: 1-work/study, 2-unemployed/sick leave/pension
- madrs1: MADRS score before activity measurement started
- madrs2: MADRS score after activity measurement ended



```
scores_data.info()
```

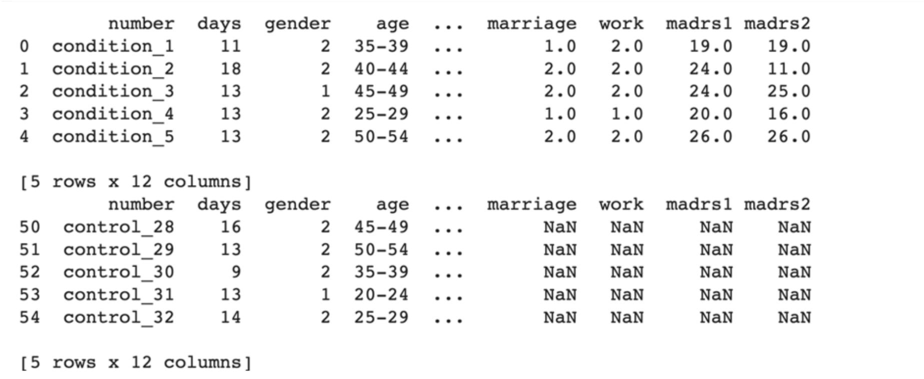
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55 entries, 0 to 54
Data columns (total 12 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   number     55 non-null     object
 1   days       55 non-null     int64
 2   gender     55 non-null     int64
 3   age        55 non-null     object
 4   afftype    23 non-null     float64
 5   melanch    20 non-null     float64
 6   inpatient  23 non-null     float64
 7   edu        53 non-null     object
 8   marriage   23 non-null     float64
 9   work       23 non-null     float64
 10  madrs1     23 non-null     float64
 11  madrs2     23 non-null     float64
dtypes: float64(7), int64(2), object(3)
memory usage: 5.3+ KB
```

Fig. 1. Figure showing dataset's scores.csv info

3 Experiments and Results

We have two objectives to complete and for each objective we will be carrying out a separate experiment. The experiments are stated below:

1. Binary Classification: Create a model to classify subject as depressed or non-depressed when fed with motor activity data of subject.
2. Multi-class Classification: Create a model to classify subject into one of the three classes: No-depression, mild-depression, severe depression when fed with motor activity data of subject.



```
        number  days  gender    age  ...  marriage  work  madrs1 madrs2
0  condition_1    11       2  35-39  ...       1.0   2.0    19.0   19.0
1  condition_2    18       2  40-44  ...       2.0   2.0    24.0   11.0
2  condition_3    13       1  45-49  ...       2.0   2.0    24.0   25.0
3  condition_4    13       2  25-29  ...       1.0   1.0    20.0   16.0
4  condition_5    13       2  50-54  ...       2.0   2.0    26.0   26.0

[5 rows x 12 columns]
        number  days  gender    age  ...  marriage  work  madrs1 madrs2
50  control_28    16       2  45-49  ...       NaN   NaN     NaN    NaN
51  control_29    13       2  50-54  ...       NaN   NaN     NaN    NaN
52  control_30     9       2  35-39  ...       NaN   NaN     NaN    NaN
53  control_31    13       1  20-24  ...       NaN   NaN     NaN    NaN
54  control_32    14       2  25-29  ...       NaN   NaN     NaN    NaN

[5 rows x 12 columns]
```

Fig. 2. Figure showing first and last five rows of scores.csv

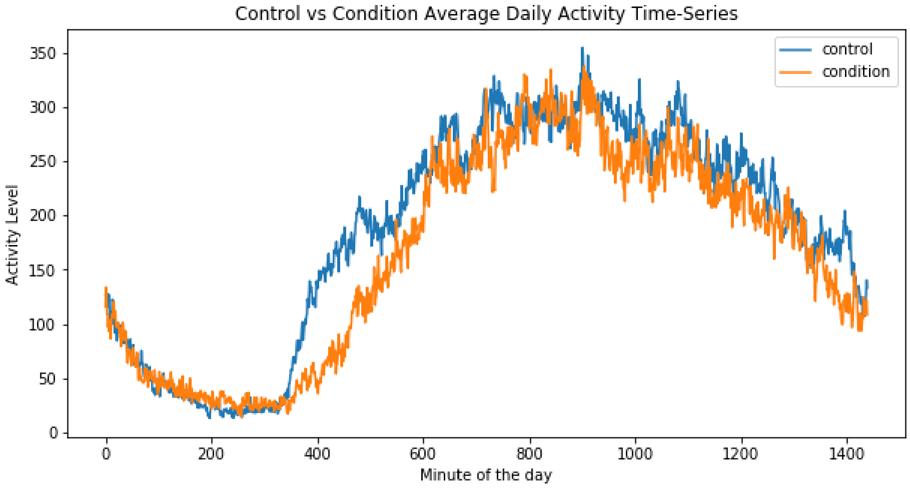


Fig. 3. Figure showing average daily activity of condition and control subject

Since the input data for each subject varies in length, as for some subjects the observation duration is seven days and for some others, it was around ten days. So we need to make the input data size constant for our models to work. While experimenting we found out the segment length of 24 hours is giving in better results compared to lesser length segments. So we modeled the data as an input motor activity segment of 1440 mins and labeled each segment corresponding to the experiment and subject. For experiment 1 labels are either 0 or 1 for non-depressed and depressed respectively and for experiment 2 the labels are 0, 1 & 2 for no-depression, mild-depression & severe depression respectively.

3.1 Experiment-1: Binary Classification

Dataset Preparation: As stated earlier the dataset is a time-series data having different input length size for each data point which is not suitable for machine learning models. So we modeled the dataset such that the input size is a constant segment of length 1440 mins (1 Day) and labeled 0 for non-depressed subjects and 1 for depressed correspondingly. After modeling we found out that we have a total of 1029 data points out of which 359 belonged to the condition group and 670 belonged to the control group. Now we had data to be fed to machine learning models but this data was unbalanced so we needed to balance it first. So we first split the data into training and testing set in a ratio of 8:2 respectively resulting in a training data size of 823 data points which was unbalanced data. To balance the training set we performed oversampling on it using SMOTE oversampling technique resulting in 1100 training samples. After modeling the data and balancing the training set fed this data to different ensemble models.

Training and Testing: We fed the data prepared to AdaBoost, XGBoost, CatBoost, GBM.

Table 1. Table showing performance matrices for AdaBoost, XGBoost, CatBoost, GBM for binary classification

Ensemble Model	Accuracy	F1-Score	Precision Score	Recall Score
AdaBoost	0.70	0.70	0.70	0.71
XGBoost	0.72	0.72	0.72	0.73
CatBoost	0.78	0.78	0.78	0.78
GBM	0.76	0.76	0.67	0.76

3.2 Experiment-2: Multiclass Classification

Dataset Preparation: Prior to performing this experiment too, we need to prepare our dataset and associate the data points with appropriate labels based on the MADRS score values of each subject. The labels associated with each data point is based on the following rules:

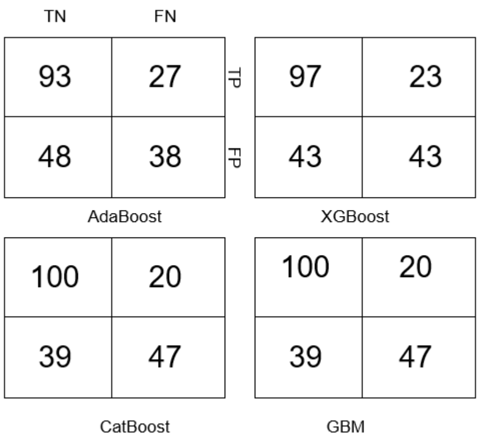


Fig. 4. Figure showing confusion matrix for AdaBoost, XGBoost, CatBoost & GBM for binary classification

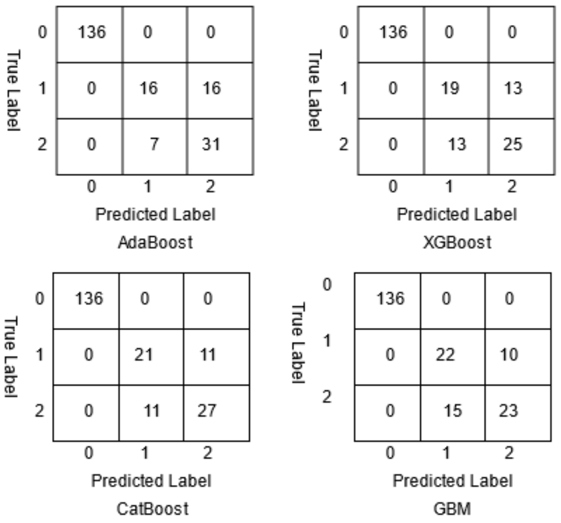
1. Calculate average MADRS using madrs1 and madrs2 values in scores.csv file of dataset.
2. Average MADRS ≥ 0 and Average MADRS ≤ 10 implies No-depression.
3. Average MADRS ≥ 11 and Average MADRS ≤ 19 implies mild-depression.
4. Average MADRS ≥ 20 and Average MADRS ≤ 36 severe depression.

Since this data for each subject is of different length we converted the data into a segment length of 1440 mins (1 Day) and the obtained dataset has 1029 samples we split this dataset into training and testing set in ratio 8:2. The obtained training set contained unbalanced data i,e. the number of data points from the no-depression class (labeled 0) consisted of 546 samples, mild-depression class (labeled 1) consisted of 119 samples whereas severe depression class (labeled 2) contained 158 samples. We used SMOTE oversampling technique to balance the dataset so that each of the classes has an equal number of data samples which resulted in 546 samples in each class.

Training and Testing: We fed the obtained data to AdaBoost, XGBoost, CatBoost and GBM. The result on testing set are shown below:

Table 2. Table showing performance matrices for AdaBoost, XGBoost, CatBoost, GBM for multiclass classification

Ensemble Model	Accuracy	F1-Score	Precision Score	Recall Score
AdaBoost	0.89	0.89	0.89	0.89
XGBoost	0.87	0.87	0.87	0.87
CatBoost	0.89	0.89	0.89	0.89
GBM	0.88	0.88	0.88	0.88



True Label	0	1	2
0	136	0	0
1	0	16	16
2	0	7	31

Predicted Label

AdaBoost

True Label	0	1	2
0	136	0	0
1	0	19	13
2	0	13	25

Predicted Label

XGBoost

True Label	0	1	2
0	136	0	0
1	0	21	11
2	0	11	27

Predicted Label

CatBoost

True Label	0	1	2
0	136	0	0
1	0	22	10
2	0	15	23

Predicted Label

GBM

Fig. 5. Figure showing confusion matrix for AdaBoost, XGBoost, CatBoost & GBM for multiclass classification

4 Discussion

For binary classification, we got the best accuracy of 0.78 and F1-score of 0.78 on the test set using CatBoost ensemble model with SMOTE on the training set which is slightly better than the F1-scores obtained using other ensemble models for the same training and testing set.

For multiclass classification, we got the best accuracy of 0.89 and F1-score of 0.89 on the test set using CatBoost ensemble model with SMOTE on the training set which is slightly better than F1-scores obtained using other ensemble models for the same training and testing set.

Also, our f1-score is better compared to Garcia-Ceja et al. [17] which is around 0.73 using SMOTE technique and Random Forest.

Gracia-Ceja et al. also proposed classification based on MADRS score as future work in his research paper which we tried to achieve. Although the overall performance was not exceptional we were able to classify subjects into different classes which good accuracy and f1-score.

5 Conclusion

In this research paper, we presented a comparative study of the performance of different ensemble models on the depressjon dataset. We carried out two different experiments, one to classify the subjects into depressed and non-depressed and the other to classify subjects based on average MADRS value into three different classes namely no-depression, mild depression, and severe depression. We found out that ensemble model tends to perform well to detect depression in both the experiment but due to small size of data the performance of these model were not exceptional. If more data can be collected then the performance of these models can further be improved and these models can be used in smartwatches and smartphones to continuously monitor the motor activity of the user and warn them if there is some sign of depression.

References

1. Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., Cabitza, F.: Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems **44**(8), 1–12 (2020)
2. Burton, C., McKinstry, B., Tătar, A.S., Serrano-Blanco, A., Pagliari, C., Wolters, M.: Activity monitoring in patients with depression: a systematic review. Journal of affective disorders **145**(1), 21–28 (2013)
3. Fasmer, O.B., Akiskal, H.S., Kelsoe, J.R., Oedegaard, K.J.: Clinical and pathophysiological relations between migraine and mood disorders. Current Psychiatry Reviews **5**(2), 93–109 (2009)
4. Faurholt-Jepsen, M., Vinberg, M., Frost, M., Debel, S., Margrethe Christensen, E., Bardram, J.E., Kessing, L.V.: Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. International journal of methods in psychiatric research **25**(4), 309–323 (2016)
5. Frogner, J.I., Noori, F.M., Halvorsen, P., Hicks, S.A., Garcia-Ceja, E., Torresen, J., Riegler, M.A.: One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In: Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care. pp. 9–15 (2019)
6. Garcia-Ceja, E., Riegler, M., Jakobsen, P., Tørresen, J., Nordgreen, T., Oedegaard, K.J., Fasmer, O.B.: Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In: Proceedings of the 9th ACM multimedia systems conference. pp. 472–477 (2018)
7. Garcia-Ceja, E., Riegler, M., Jakobsen, P., Torresen, J., Nordgreen, T., Oedegaard, K.J., Fasmer, O.B.: Motor activity based classification of depression in unipolar and bipolar patients. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). pp. 316–321. IEEE (2018)
8. Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K.J., Tørresen, J.: Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing **51**, 1–26 (2018)
9. Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Oehler, S., Tröster, G., Mayora, O., Haring, C., Lukowicz, P.: Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics **19**(1), 140–148 (2014)
10. Manderscheid, R.W., Ryff, C.D., Freeman, E.J., McKnight-Eily, L.R., Dhingra, S., Strine, T.W.: Peer reviewed: evolving definitions of mental illness and wellness. Preventing chronic disease **7**(1) (2010)
11. Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., Morales, E.F.: Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive and Mobile Computing **31**, 50–66 (2016)

12. Mehl, R.C., O'Brien, L.M., Jones, J.H., Dreisbach, J.K., Mervis, C.B., Gozal, D.: Correlates of sleep and pediatric bipolar disorder. Sleep **29**(2), 193–197 (2006)
13. Montgomery, S.A., Åsberg, M.: A new depression scale designed to be sensitive to change. The British journal of psychiatry **134**(4), 382–389 (1979)
14. Mowery, D.L., Park, Y.A., Bryan, C., Conway, M.: Towards automatically classifying depressive symptoms from twitter data for population health. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES). pp. 182–191 (2016)
15. O'Brien, J., Gallagher, P., Stow, D., Hammerla, N., Ploetz, T., Firbank, M., Ladha, C., Ladha, K., Jackson, D., McNaney, R., et al.: A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. Psychological medicine **47**(1), 93–102 (2017)
16. Penedo, F., Dahn, J.R.: Exercise and well-being: a review of mental and physical health benefits associated with physical activity. Current Opinion in Psychiatry **18**, 189–193 (2005)
17. Reece, A.G., Danforth, C.M.: Instagram photos reveal predictive markers of depression. EPJ Data Science **6**, 1–12 (2017)
18. Sharma, K., Kaur, A., Gujral, S.: Brain tumor detection based on machine learning algorithms. International Journal of Computer Applications **103**(1), 7–11 (2014)
19. Twenge, J.M.: Time period and birth cohort differences in depressive symptoms in the us, 1982–2013. Social Indicators Research **121**(2), 437–454 (2015)