German Medical Data Sciences 2022 - Future Medicine
R. Röhrig et al. (Eds.)

© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220800

Machine Learning Based Classification of Depression Using Motor Activity Data and Autoregressive Model

Alexander SCHULTE[a] and Tim BREIKSCH[a] and Jonas BROCKMANN[a] and Nadja BAUER[a1]
[a] *University of Applied Sciences and Arts Dortmund*

Abstract. Machine learning based disease classification have already achieved amazing results in medicine: for example, models can find a tumor in computer tomography images at least as accurately as experts in the field. Since the development and widespread use of actigraphy watches, activity data has been used as a basis for diagnosing various diseases such as depression or Alzheimer's disease. In this study, we use a dataset with activity measurements of mentally ill and healthy people, calculate various features and achieve a classification accuracy of over 78%. The paper describes and motivates the used features, discusses differences between healthy, bipolar 2 and unipolar participants and compares several well-known machine learning classifiers on different classification tasks and with different feature sets.

Keywords. Machine learning, depression classification, actometer data, actigraphy watch, depresjon dataset, autoregressive model

1. Introduction

Depression is a serious problem and the most common mental illness in the population. The Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) 2019 rank depressive disorders under the top 25 burdens worldwide in 2019 [6]. [3] show that the COVID-19 pandemic noticeably impacts the mental health of the population as the infection rates are associated with increased prevalence of major depressive disorder. Even though there are often no clear physical signs, the illness can have a disruptive effect on the life of an affected person over an indefinite period of time. A distinction can be made between different types and intensities, which can result in quite different progressions of the disease. Bipolar depression is characterized not only by the classic phases of lack of motivation - as common for unipolar depression - but also by opposite, impulsive phases [4]. A correct classification can thus help to assess the possible extent of the depression more precisely and to improve possible treatment.

Richter et al. [10] summarize in their overview study different machine learning (ML) based behavioral diagnostics tools for depression. The authors distinguish between neuroimaging data (such as brain network patterns) and behavioral data, the latter being divided into social media usage and movement sensor data. The behavioral data are

[1] Corresponding Author, Dr. Nadja Bauer, Lecturer at the Faculty of Computer Science, University of Applied Sciences and Arts Dortmund, E-mail: nadja.bauer@fh-dortmund.de.

particularly interesting for diagnostic purposes because they are generally easier to obtain. In this paper, we deal with movement sensor data and give a short overview of related works in this subfield.

Berle et al. [1] provide one of the first studies for distinguishing the movement patterns of healthy and mentally ill people. They collect actigraphy watch data over a two-week period from both groups. The used features are activity averages (full-time and nights), as well as interdaily and intradaily stability (measurement of the strength of circadian rhythmicity). The main conclusion is that mentally ill patients tend to show a lower motor activity as well as a more structured behavior than the control group. However, ML techniques for automated disease classification were not applied. Parts of this dataset have been published by [5] for research purposes and form the basis for the analysis in this paper.

Garcia-Ceja et al. [5] have not only published the so-called "depresjon" dataset, but have already applied different ML methods and compared the results. However, the underlying features are not mentioned. Linear SVM was shown to be the best method with an accuracy of 72.7%. The accuracy of the so called zeroR-classifier (assignment of majority class) is 58%. Authors emphasize the need for sophisticated feature engineering.

Currently, many studies are being conducted worldwide in this direction. [8] compare activity data of older (mostly female) single people with (n=18) and without (n=29) depression. Used features include activity averages, light condition and various sleep parameters, showing low levels of daytime activity for depressed individuals in particular. Logistic regression showed by far the best classification accuracy with 91%, random forest reaches 67% and zeroR 61%.

Minaeva et al. [9] analyze two (not freely available) datasets (development and validation) with activity data of depressed patients, looking at a variety of features, including sleep behavior and some parameters of the fitted circadian curve. The development dataset includes 43 depressive and 82 non-depressive people, with a validation set of 27 people for each group. Backward stepwise logistic regression is used as a ML-method. The classification accuracy on the development dataset amounts to 71.8 % (zeroR accuracy: 65.6 %). For the validation dataset, instead of an accuracy only an AUC value is reported, being 0.65. After backward selection two activity data driven features were left: average of daily gross motor activity and acrophase (time of maximum activity levels across 24-hour periods).

Rodríguez-Ruiz et al. [11] also use the "depresjon" dataset and report an almost perfect accuracy of 99%. However, a closer look reveals inaccuracies. The original time series are divided into segments of 60 minutes and 24 features (some based on Fourier transformation) are calculated, resulting in a dataset of 11945 observations. Although the data were divided into training and validation sets, it was not excluded that the data of the same individuals occur in both subsets leading to information leakage and hence almost perfect accuracy.

Zanella-Calzada et al. [13] also use the "depresjon" dataset for detecting depressive episodes, achieving an accuracy of 89% for the classification of depressed vs. healthy subjects. However, their results are not comparable with ours in many respects - for one, they claim to have 5895 subjects (2112 cases / 3783 controls), although the linked dataset contains only 23 unipolar and bipolar depressed patients and 32 healthy participants. Thus, the patient data is probably divided into different sections risking information leakage leading to a too optimistic classification accuracy. On the other hand, they use an out-of-bag estimate as validation strategy, which only resembles cross-validated error

rates after many repetitions. Our approach is much broader, since we additionally want to distinguish between the two depression types and extract a much larger and heterogeneous set of features.

Sing et at. [12] follow a very similar procedure as [13]: they divide the original data into a total of 13844 segments with the goal of learning a classification model on these segments. The same features as in [13] are extracted and a random forest model is used as well. The problem of such an approach is that although individual data segments can be well classified into depressed vs. non-depressed ones, it is not possible to make a diagnosis for a person. Indeed, it is quite possible that parts of a person's data segments will be classified as unipolar, another as bipolar 1, and still another as healthy. Therefore, our goal is to classify individuals and not isolated episodes.

2. Data

We use the "depresjon" dataset of [5] mentioned in the introduction. It consists of two groups: the so-called conditional group of 23 patients with a major depressive disorder and the control group with 32 non-depressive contributors. Each study participant wore an actometer for about two weeks. The wrist-mounted actometers recorded any motion above 0.05g with a sampling frequency of 32Hz leading to data entries in minute intervals. 15 patients of the conditional group have a unipolar and 8 a bipolar disorder. However, for bipolar patients there is a distinction between bipolar 1 and bipolar 2 disorders, while there is only one person with a bipolar 1 disorder. Therefore, we excluded the corresponding data, leading to a dataset with 54 participants.

Some datasets contain long episodes of zero data (probably caused by taking off the actigraphy watch) which could negatively influence the results of the machine learning analysis. Hence, if a period of zero data exceeds a certain value, it is removed from the dataset. Starting with smaller values, we concluded that it suffices to focus this filter on the highest occurring periods, using a value of 5760 minutes (corresponding to 4 days), without changing the result in a remarkable way, impacting only four participants from the control group in total (with id numbers 1, 3, 31 and 32).

Although our work relies only on the existing annotated data and compares different ML approaches to classification, it is necessary to question the quality of the data for the generalizability of the results. Only little information is known about the conducted study (see [1]), such as that the control group was composed of hospital employees (n=23) and students (n=5). There are considerably more women in the control group than in the conditional group. Other confounders that could affect the internal and external validity of the study are not discussed. Consequently, the results of papers based on these data should be interpreted with great caution. At the same time, this highlights the need for further data from similar studies that meet a high standard of clinical research to be made freely available to the ML community.

3. Feature Extraction

Since the data are in the format of a time series with over 20 000 samples per person, it is not possible to take each data entry as a feature. Hence, the time series have to be compressed into a small set of meaningful features which then influence the classification accuracy. Some of these features are motivated through state-of-the-art

works and others are proposed by the authors. All calculations were conducted in the Java programming language. We will also mention some used package names.

The first batch of features consists of 11 self-explanatory ones that do not require formal definitions: the highest occurring value (*maximum*), as well as the lowest one (*minimum*), the average (*average*), the median (*median*), the mode (*mode*), the standard deviation (*stdDev*), the variance (*variance*), the coefficient of variation (*varCoeff*), the kurtosis (*kurtosis*) over all samples and the number of occurrences of the value 0 (*nullCount*). An additional standard deviation value is calculated based on the average values of each 24-hour interval to check the heterogeneity between days (*dailyDev*). As almost all studies in this area distinguish between day and night activity, we compute the heterogeneity between the averaged night activities (*nightlyDev*), analogue to *dailyDev*, while defining the night as the time from 10 p.m. to 7 a.m. A further sleep quality feature is defined as: *sleepQuality = (1 – (nightlyDev / stdDev))*. The idea is, that individuals with low night activity fluctuations in relation to overall fluctuations will have *sleepQuality* close to 1, while individuals with very similar fluctuation levels overall and during nights will have *sleepQuality* close to 0.

As smoothing is one of the standard tools for time series analysis, the second batch of 5 features is extracted from the smoothed time series after a moving average with a window size of 11 was applied. The following features are then calculated: maximum (*maMax*), minimum (*maMin*), average (*maAverage*), standard deviation (*maStdDev*) and a relative maxima difference defined as: *maxDiffFactor = (max – maMax) / max*. The larger this characteristic is, the more noticeable is the strongest outlier in the time series.

The third batch of 5 features is applied on the Fourier transformed data as this approach was proposed in [11]: maximum (*fftMax*), average (*fftAverage*), standard deviation (*fftStdDev*), variance (*fftVar*), coefficient of variation (*fftVarCoeff*) and kurtosis (*fftKurtosis*). The window size for the Fourier transformation is the length of the respective time series. Calculations were done using the "FFT4J" java package.

Finally, the fourth feature batch is based on an autoregressive (AR) model. The main idea was to investigate whether the behavior of the participants can be predicted well by an ARIMA model as we expect that the "predictability" of the activity patterns could be a good feature to distinguish between control and conditional group (see [1]). An ARIMA model has several parameters: lag order p, degree of differencing d and the order of the moving average q [2]. We tested in first pre-experiments the impact of different parameter settings and have found that the best results can be achieved for p larger than 20, $d=0$ and $q=0$. So, the ARIMA model was reduced to an AR model with $p=25$.

For feature calculations, we first train the AR model on the first two thirds of each person's activity data and then predict the next activity value t_n as well as the lower and upper confidence bounds for the prediction. Then the absolute difference between the true and the predicted value for t_n is calculated. Furthermore, the normalized variance for the prediction (prediction variance divided by variance of the fitted model) and the root mean squared error for the trained model are noted (as a measurement for model uncertainty and goodness of model fit, respectively). In the next successive steps, the training data is extended by t_n and the model is trained once again for predicting t_{n+1}. Finally, the computed features for all prediction steps are averaged resulting in: *arPredError*, *arLowerConf*, *arUpperConf*, *arNormVar*, *arRootMeanSqrtError*. The Java package "timeseries-forecast" was used for these calculations.

4. Explorative Data Analysis

In this section explorative comparisons of features depending on the group of participants will be provided. Here we divide the conditional group into bipolar and unipolar patients. Fig. 1 compares the activity data (after applying the moving average as described above) of two participants: one from the control group (A) und one from the conditional group (B) for a time period of two weeks. First, clear day-night cycles are visible. The comparison shows a clear difference in average activity: in general, the healthy person seems to be more active than the other one. This behavior is measurable across almost all cases as can be seen in Fig. 2 (a) for the *average* feature. Similar to findings in [1], [8] or [9], higher activity of the control group can also be stated when comparing the features *maximum*, *stdDev*, *variance*, *median*, *dailyDev* and *mode*.

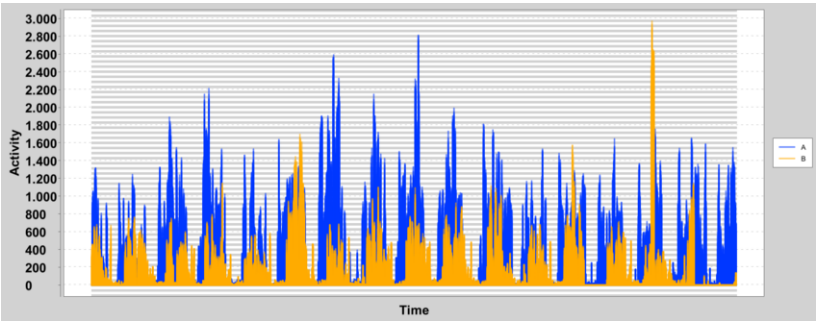

Figure 1. Activity data (after applying moving average) of a healthy participant (A) and of a participant with unipolar depression (B).

The *kurtosis* and *varCoeff* features show no meaningful difference between healthy, bipolar 2 or unipolar individuals. The *minimum* activity of each dataset is 0 and therefore this feature is irrelevant for further classification. The conditional group shows an increased *nullCount* value compared to the control group, which can again be explained by lower activity. The *sleepQuality* of bipolar 2 patients seems to be lower than for other participants (see Fig. 2 (b)). This can be explained with the higher *nightlyDev* of this group.

The moving average batch of features yields similar results: higher values for the *maMax*, *maAverage* and *maStdDev* features for the control group. Unipolar patients show a higher *maxDiffFactor* value compared to healthy and bipolar 2 participants, which could be an indicator for higher "extreme values" in their activity.

For the frequency domain-based attributes from the third feature batch the following can be stated: the attributes *fftVar*, *fftStdDev* and *fftMax* have higher values for the healthy control group, while for *fftAverage* and *fftVarCoeff* no clear differences between the groups can be observed. The value of *fftKurtosis* for the conditional group (especially for bipolar 2 patients) is lower compared to the healthy group (see Fig. 2 (d)). Spectral kurtosis is an indicator for randomly occurring fluctuation in the activity profile, so depressive patients seem to me more "predictable" in their activity patterns.

The last batch with AR based features also shows differences between bipolar 2, unipolar and healthy participants. The value for *arPredError* is substantially lower for depressed patients of both kinds compared to healthy participants, although bipolar 2 patients retain a higher value than unipolar patients (see Fig. 2 (d)). Hence, participants with a unipolar disorder seem to have more structured activity patterns. *arRootMeanSquareError* values are in general lower for depressed patients meaning a better model fit for this group. However, the condition group shows higher *arNormVar* values compared to the healthy control group, which might be caused by smaller variances of the fitted models. According to the distribution of the *arLowerConf* and *arUpperConf* features, the bipolar 2 patients show smaller prediction intervals than other participants.

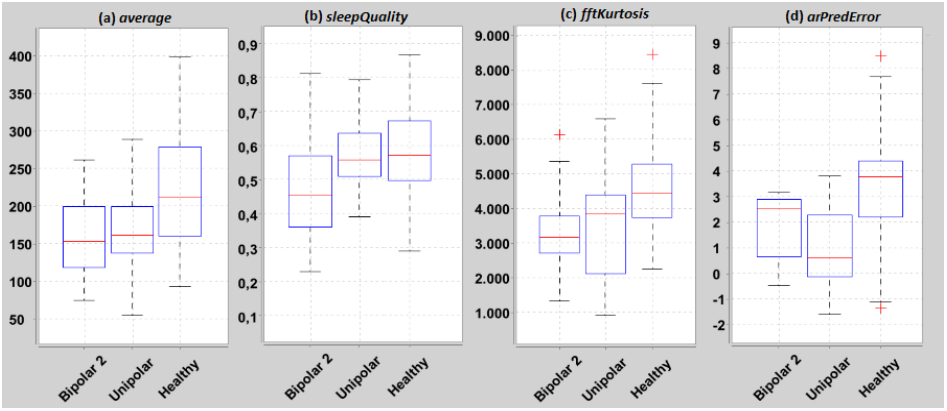

Figure 2. Distribution comparison of some selected features for three groups: patients with bipolar 2 disorder, patients with unipolar disorder and healthy participants (control group).

5. Machine Learning based Classification

The theoretical aspects of machine learning techniques applied in this paper (like classifiers, validation, performance measures) can be found in [7].

The following classification tasks will be considered: 1) unipolar vs. non-unipolar, 2) bipolar 2 vs. non-bipolar 2, 3) healthy vs. non-healthy (bipolar and unipolar) participants. Especially for the second task the problem of unbalanced classes occurs as there are only seven bipolar 2 patients against 47 non-bipolar 2 participants. For this reason, we will compare the classification results with the zeroR-classifier (classification to the majority class). Furthermore, in order to measure the impact of AR based features three sets of features are compared: 1) "No AR" - all features except for the AR batch, 2) "AR only" - only AR based features and 3) "Combined" - all features. We compare RandomForest, AdaBoost and LogReg (logistic regression) classifiers, as these approaches count to the top data mining algorithms and in order to verify the remarkably good performance of logistic regression in [8]. Machine learning was realized using the Java library "Smile".

To avoid overfitting, 50 times replicated 10-fold cross validation is applied and measured by the classification accuracy. As a reminder, the dataset consists of 54 observations, 26 features and a target variable with three labels (unipolar / bipolar 2 / healthy). The results are presented in Table 1, while the best result for each classification

task is highlighted in bold. The main finding is that the AR features seem to lead to a relevant improvement of the classification accuracy for all three tasks, while for bipolar 2 vs. non-bipolar 2 and for healthy vs. non-healthy classification "AR only" shows the best results. Logistic regression achieves better results on extremely unbalanced classification tests (bipolar 2 vs non-bipolar 2), slightly beating the zeroR-classifier. AdaBoost performs better on both other tasks with accuracies clearly better than the zeroR-classification.

The achieved accuracies (especially for healthy vs. non-healthy task) seem to outperform the state-of-the-art results. In order to verify the feature importance, we applied RandomForest and AdaBoost models on the whole dataset for the healthy vs. non-healthy classification task and computed variable importance values. The top tree features for RandomForest are: *arPredError*, *fftMax* and *mode*, and for AdaBoost: *average* (by far), *nullCount* and *arPredError*.

Table 1. Results of the classification benchmark. Validation criterion: accuracy measure, validation strategy: 50 times replicated 10-fold cross validation

Classification task	Classifier	Feature set		
		No AR	AR only	Combined
unipolar vs. non-unipolar	RandomForest	74.07 %	74.27 %	76.93 %
	AdaBoost	66.53 %	72.27 %	**79.33 %**
	LogReg	73.85 %	78.71 %	71.90 %
	zeroR	72.22 %	72.22 %	72.22 %
Bipolar 2 vs. non-bipolar 2	RandomForest	87,16 %	87.33 %	87.20 %
	AdaBoost	81,47 %	80.00 %	84.60 %
	LogReg	87.33 %	87.28 %	**87.39 %**
	zeroR	87.04 %	87.04 %	87.04 %
Healthy vs. non-healthy	RandomForest	65.59 %	76.87 %	70.60 %
	AdaBoost	68.30 %	**78.40 %**	75.00 %
	LogReg	59.30 %	76.05 %	59.20 %
	zeroR	59.26 %	59.26 %	59.26 %

6. Discussion

In this paper we compared several state-of-the-art features and models for the ML based detection of unipolar and bipolar 2 depression disorder - the most common psychiatric disorders worldwide. The features based on the prediction ability of an AR model seem to contribute relevant improvements to the classification accuracy leading to better results than in related works (78 % accuracy for heathy vs. non-healthy classification). However, the shortcoming of our study is the small dataset. Our requests regarding further data on several researchers remained unanswered. A larger amount of high-quality data could eventually make it possible to create reliable ML-models with as few features as possible to assist physicians based on a patient's activity patterns alone. Of course, it should be noted that this can only be used in practice once a model is proven by further prospective performance evaluation testing.

7. Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: AS, TB, JB: conception of the work, conduction of experiments and first draft of the manuscript; NB: mentoring, substantial revision of the manuscript.

References

[1] Berle JO, Hauge ER, Oedegaard KJ, Holsten F, Fasmer OB. Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression. BMC Res Notes. 2010 May 27;3:149.
[2] Box E P, Jenkins, GM. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1970.
[3] COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. 2021 Nov 6;398(10312):1700-1712.
[4] Cuellar AK, Johnson SL, Winters R. Distinctions between bipolar and unipolar depression. Clin Psychol Rev. 2005 May;25(3):307-39.
[5] Garcia-Ceja E, Riegler M, Jakobsen P, Tørresen J, Nordgreen T, Oedegaard KJ, Fasmer OB. Depresjon: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients, In MMSys'18 Proceedings of the 9th ACM on Multimedia Systems Conference, Amsterdam, The Netherlands, 2018 June; 12-15.
[6] GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020 Oct 17;396(10258):1204-1222.
[7] Hastie, T, Tibshirani R, Friedman, J. The elements of statistical learning: data mining, inference and prediction. Springer; 2009.
[8] Kim H, Lee S, Lee S, Hong S, Kang H, Kim N. Depression Prediction by Using Ecological Momentary Assessment, Actiwatch Data, and Machine Learning: Observational Study on Older Adults Living AloneJMIR Mhealth Uhealth 2019;7(10).
[9] Minaeva O, Riese H, Lamers F, Antypa N, Wichers M, Booij SH. Screening for Depression in Daily Life: Development and External Validation of a Prediction Model Based on Actigraphy and Experience Sampling Method. J Med Internet Res. 2020 Dec 1;22(12).
[10] Richter T, Fishbain B, Richter-Levin G, Okon-Singer H. Machine Learning-Based Behavioral Diagnostic Tools for Depression: Advances, Challenges, and Future Directions. J Pers Med. 2021 Sep 26;11(10):957.
[11] Rodríguez-Ruiz JG, Galván-Tejada CE, Vázquez-Reyes S, Galván-Tejada JI, Gamboa-Rosales H. Classification of Depressive Episodes Using Nighttime Data; a Multivariate and Univariate Analysis. Programming and Computer Software; Dec 2020; 46(8): 689-698.
[12] Singh PM, Sathidevi PS. (2022). Design and Implementation of a Machine Learning-Based Technique to Detect Unipolar and Bipolar Depression Using Motor Activity Data. In Smart Trends in Computing and Communications. 2022; 99-107. Springer Singapore.
[13] Zanella-Calzada LA, Galván-Tejada CE, Chávez-Lamas NM, Gracia-Cortés M, Magallanes-Quintanar R, Celaya-Padilla JM, Galván-Tejada JI, Gamboa-Rosales H. Feature extraction in motor activity signal: Towards a depression episodes detection in unipolar and bipolar patients. Diagnostics. 2019; 9(1): 8.