Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data

Adam Mikus, Mark Hoogendoorn, Artur Rocha, Joao Gama, Jeroen Ruwaard, Heleen Riper

A R T I C L E   I N F O

Keywords:
- Depression
- Machine learning
- Short term mood
- Prediction

A B S T R A C T

Technology driven interventions provide us with an increasing amount of fine-grained data about the patient. This data includes regular ecological momentary assessments (EMA) but also response times to EMA questions by a user. When observing this data, we see a huge variation between the patterns exhibited by different patients. Some are more stable while others vary a lot over time. This poses a challenging problem for the domain of artificial intelligence and makes one wonder whether it is possible to predict the future mental state of a patient using the data that is available. In the end, these predictions could potentially contribute to interventions that tailor the feedback to the user on a daily basis, for example by warning a user that a fall-back might be expected during the next days, or by applying a strategy to prevent the fall-back from occurring in the first place.

In this work, we focus on short term mood prediction by considering the adherence and usage data as an additional predictor. We apply recurrent neural networks to handle the temporal aspects best and try to explore whether individual, group level, or one single predictive model provides the highest predictive performance (measured using the root mean squared error (RMSE)). We use data collected from patients from five countries who used the ICT4Depression/MoodBuster platform in the context of the EU E-COMPARED project. In total, we used the data from 143 patients (with between 9 and 425 days of EMA data) who were diagnosed with a major depressive disorder according to DSM-IV.

Results show that we can make predictions of short term mood change quite accurate (ranging between 0.065 and 0.11). The past EMA mood ratings proved to be the most influential while adherence and usage data did not improve prediction accuracy. In general, group level predictions proved to be the most promising, however differences were not significant.

Short term mood prediction remains a difficult task, but from this research we can conclude that sophisticated machine learning algorithms/setups can result in accurate performance. For future work, we want to use more data from the mobile phone to improve predictive performance of short term mood.

1. Introduction

Depression is a highly prevalent disorder associated with a huge loss of quality of life, increased mortality rates high levels of service cost. Earlier research has estimated the cost of depression at 177 million euros per year per 1 million inhabitants for major depression on top of 147 million euros per year for minor depression (Cuijpers et al., 2007). Depression is currently the fourth disorder worldwide in terms of disease burden (Üstün et al., 2004). A lot of developments in treatments for depression can be seen in the last decade, where a shift is taking place from the more traditional face-to-face counseling to self-help therapies or blended care settings (see e.g. Kooistra et al., 2016; Riper et al., 2010). These changes have been driven by advancements in technologies in society: Internet and mobile phones are widely available and enable more technologically supported forms of interventions. Also better and more fine-grained ways of measuring the state of patients have resulted, such as EMA (for Ecological Momentary Assessment, see e.g. Shiffman et al., 2008; Asselbergs et al., 2016): measurements that assess the mental state of a patient in context and over time, often via questions posed to the patient on the mobile phone.

EMA questions can be useful for a therapist or researcher to understand how a patient is progressing. The purpose can range from...
gaining a fine-grained insight in the manifestation of depression before or during the start of the treatment towards fluctuations in mood and behaviors during treatment. In addition however, it also makes one wonder whether it is possible to extract patterns that allow one to create forecasts of the mental state for individual patients. Previous research has shown that prediction of mood is difficult due to large individual differences and a limited predictive value of previous measurements for the future developments (see e.g. van Breda et al., 2016b; Becker et al., 2016; van Breda et al., 2016a). There are however more advanced techniques available from the domain of artificial intelligence that might result in models that are better able to predict future developments of a patient. In addition, usage data (from both the EMA/therapeutic app and the web modules) has hardly been exploited to improve predictions.

In this paper, we will focus on predicting short term mood for the EMA data that has been collected within the E-COMPARED project. E-COMPARED stands for European COMPARative Effectiveness research on blended Depression treatment versus treatment-as-usual. The protocol of the trial conducted in the project can be found in Kleiboer et al. (2016). We aim to predict the value reported for the mood of the patient on the next day and will extend previous work van Breda et al. (2016b) by (1) using the response times to EMA requests and usage data of the patient as an additional predictor, and (2) applying more sophisticated machine learning techniques in the form of recurrent neural networks. The EMA dataset we use contains data from 143 patients with at least 9 up to 425 days of EMA data. We evaluate the approaches by comparing the predictions of the model with the reported EMA values.

This paper is organized as follows. First, we will describe how the data has been collected and the setting in which we apply our machine learning algorithms. We will then describe the application of the machine learning algorithms themselves followed by the results. Finally, we will draw conclusions.

2. Dataset description and initial exploration

We use the EMA data collected from patients from five countries who used the MoodBuster platform in the context of the EU E-COMPARED project. In total we used the data from 143 patients (with between 9 and 425 days of EMA data) who were diagnosed with a major depressive disorder according to DSM-IV. Table 1 provides an overview of the EMA questions that were posed.

Next to the EMA data we also exploit the log table of the MoodBuster system which indicates the response times of patients, i.e. the time between the system requesting a mood rating from the user and the moment when the actual input is received. In addition, it stores a lot more information about the behavior of the user, including module completion and the amount of time spent on them, messages being exchanged (when, their frequency and the number of characters), the number of web sessions, and the number of pages passed in the module. Hence, a wealth of data.

Before diving into the details of the method to forecast the EMA rating of the next day, we performed an initial correlation analysis (using the Pearson metric) between the EMA rating of the previous day and the next day. We performed this on a per patient basis as these correlations differ greatly per patient. The results show that from 27% to 45.4% of the patients show relevant (i.e. correlation coefficient above 0.4 or below −0.4) correlations with the specific EMA values of the previous day and today’s mood. For most of the patients, the highest correlation values occurred between the mood of previous day and the mood of today. These correlations are mainly positive, except in the case of worrying - where the category is rated in the opposite direction (the higher the worse).

We have also explored the relationship between the response times and EMA ratings. To be more precise, we computed the correlations between the following values:

1. Today’s response time and today’s mood
2. Yesterday’s response time and today’s mood
3. Today’s mood and tomorrow’s response time

The analysis resulted in 11 patients who showed relevant correlation values. To validate the results of the analysis, simple linear regressions have been performed between the time series of mood and its corresponding response times. The regressive modeling resulted in 6 patients where the p-values were less than 0.05. Now that we gained a bit more insight into the raw data, let us move to our machine learning approach.

3. Machine learning approach

This section addresses the machine learning approach. First, we describe the features used to feed the recurrent neural network and how they are computed. These are based on the raw data we have just explained. Then we explain the machine learning approach and the different settings that we tried.

3.1. Features

As said, we aim at predicting the reported mood value of the next day based on the measurements and response times of the previous day (s). To get most out of our data, we have developed dedicated features that summarize the patients rating and behavior during previous days. These are shown in Table 2. We distinguish between base features and an extended set of features which exploits data about the response times, the web sessions and messages exchanged.

We obtain these values for all patients from the MoodBuster database and take values on a per day basis. The intensity of the EMA questions changes according to the trial protocol (Kleiboer et al., 2016) (more ratings in the start and end weeks of the treatment). Ratings are triggered randomly in the following time intervals: 1) 9 to 10 am; 2) 8 to 9 pm.

We normalized values using a scale between 0 and 1. For category type features (e.g. weekdays) we use a binary encoding, meaning that we create a feature per value and express a 1 when the values holds and a 0 otherwise. We obtain a lot of missing values as some ratings are measured less frequent and in addition, patients do not always provide ratings.

In the case of EMA mood ratings, we interpolate values if we have a gap of at most three days by considering the previous and next value and interpolating in a linear fashion. Otherwise we use the mean value for that feature of the patient over the days we do have values for.

This way, the regular input set does not contain any missing values. In the case of extended input set, due to the difficult nature of missing value imputation, binary indicators have been used as dummy variables to indicate whether the value was missing or not. Missing values have been filled with 0 (by the nature of the collected data 0 is not an option to be given by the patient).
many previous data points we give has been varied, ranging from 7 to
error). The approach is also referred to as a sliding time window. How
values for the next day and learn weights such that the network predicts
points in the past (i.e. the history of the patient) as well as the mood
(2014):
experimenting with different architectures motivated by Sak et al.

## 3.2. Machine learning model

As said, we apply recurrent neural networks. These are neural net-
works that can contain recurrent connections that allows them to ex-
plot trends seen over time and using the complete history of data to
make predictions on future values.

We apply the two most common techniques, namely Long-Short
Term Memory (LSTM (Sepp Hochreiter, 1997)) and Gated Recurrent
Unit (GRU (Cho et al., 2014)) networks. Both are recurrent neural
networks, but have a slightly different structure (GRUs are essentially a
simplification of LSTM’s). Explaining the difference is beyond the scope
of this paper, but it is interesting to see whether we can observe differ-
ence in performance.

We implemented the following three types of configurations for ex-
perimenting with different architectures motivated by Sak et al.
(2014):

1. Single layer deep LSTM or GRU
2. Multi-layer deep LSTM or GRU
3. LSTM or GRU with recurrent projection layer (LSTMP/GRUP)

To train the network, we present the network with a series of data
points in the past (i.e. the history of the patient) as well as the mood
values for the next day and learn weights such that the network predicts
the mood value as close as it can (thereby minimizing the mean squared
error). The approach is also referred to as a sliding time window. How
many previous data points we give has been varied, ranging from 7 to
10 where 7 was found to be best fit. This means that we feed a week of
data through our network and use that to predict the mood value on the
8th day.

## 3.3. Training scenarios

We have trained the machine learning algorithms using three dif-
ferent setups:

1. We create a single predictive model over all patients and train it
with a training set composed of approximately 50% of the data from
all patients (6428 training examples in total) and test it on the re-
mainning (unseen) data of the patients (6497 instances).
2. We create clusters of similar patients and create models per cluster
using a similar training and test set setup as we just described (again
a 50/50 split).
3. We create individual predictive models by just using the data of that
individual, train it on 50% of the data of the patient and test it on
the other 50%.

We have selected these three approaches to study the trade-off be-
tween generalized and individual models, driven by the great diversity
we encounter in the observed patterns exhibited by the patients. The
single predictive model will have a lot of data to train on, but might not
perform well due to the large individual differences, while the in-
dividual models will be tailored but likely suffer from a lack of suffi-
cient data. The clustering could provide a nice middle ground. To create

| Feature | Description |
|---------|-------------|
| Count   | Number days with answers (average mood ratings) |
| Mean    | Mean of the daily average mood ratings in the examined period |
| Standard deviation | Standard deviation of the daily average mood ratings in the examined period |
| Maximum | Maximum rated value of the daily average mood ratings in the examined period |
| Minimum | Minimum rated value of the daily average mood ratings in the examined period |
| Median  | Median of the daily average mood ratings in the examined period |
| Q1, Q3  | First quartile and third quartile of the daily average mood ratings in the examined period |
| Rating ratio | Number of days with mood ratings divided by all days in the examined period |
| Maximum rating time Difference | Largest difference in days between two consecutive rated days in the examined period |
clusters we apply hierarchical clustering on statistical summaries of the full EMA data of the patients. The features used for clustering are listed in Table 3. In addition, we apply a benchmark following prior research, namely to use Support Vector Regression with an RBF kernel for forecasting (cf. van Breda et al., 2016b).

In addition, we vary the features that are used, we try to use only the base features first and then study the difference in performance when further adding the extended features. For the individual models we do not include the extended set of features as we observed too many missing values for which we could not impute the values properly. Each model is trained and tested 5 times (as the recurrent neural networks are of a stochastic nature) and results have been averaged over the 5 iterations. We measure the mean squared error (MSE). Our algorithms and set-ups have been implemented in python with a neural network library called Keras (Chollet et al., 2015).

4. Results

Before we move to the performances of the novel setup we introduce in this paper, we will focus on the clustering results which precedes the application of the other machine learning techniques. Fig. 1 shows the dendrogram of the clustering of the patients. This shows which patients have been grouped together and how close they are. Based on visual inspection of the dendrogram, we select 12 clusters.

The next step is to consider the performance of the benchmark algorithm for our different setups. The results are shown in Table 4. We can see that the performance is quite good already, with the clustering setup being the best performing (i.e. obtaining the lowest RMSE).

| Set-up    | RMSE results (standard deviation) |
|-----------|----------------------------------|
| Regular input |                                   |
| Single      | 0.090 (0.00)                     |
| Clustered    | 0.077 (0.026)                    |
| Individual   | 0.098 (0.051)                    |
| Extended input |                                |
| Single      | 0.100 (0.00)                     |
| Clustered    | 0.100 (0.028)                    |

Let us move on to the results using the recurrent neural networks. Table 5 presents the best results we obtained with the different setups (i.e. generic, cluster and individual models using either the base features or base and extended) and specifies the best recurrent neural network approach for the specific setup. The results are an improvement over the benchmark (though not significant). The results show that all of the implemented recurrent neural network models have comparable results, the differences are low. In the case of clustered data, models fed with regular features shows slightly better performance than models with extended input set which includes the response times. Before diving into the details of the individual scenarios, let us consider a few plots to show how accurate we can make predictions. Figs. 2–5 show example predictions for the unseen test data using the different approaches. The figures show that the general trends exhibited by the patients are quite nicely predicted. Certainly outliers are however notoriously difficult to predict.

More details about the performance are shown in Tables 6–8. We observe that the performances between the different setups do not differ much. The two layer algorithms perform slightly better than LSTM/GRU with only 1 layer. Overall, the best performance was achieved by Gated Recurrent Unit with recurrent projection layer trained with regular inputs in the clustered set-up. While the worst overall performance was created by 1 layer Long-short term memory network in the individual set-up. We do frequently observe the strange notion that the performance on the test set is better than the performance on the training set, very uncommon when faced with machine learning problems. Investigating this notion, we found out that the

---

**Fig. 1.** Dendrogram of the results of hierarchical clustering (the numbers express the IDs of the patients).

**Table 4**

Results of benchmark SVR.

| Set-up    | RMSE results (standard deviation) |
|-----------|----------------------------------|
| Regular input |                                   |
| Single      | 0.090 (0.00)                     |
| Clustered    | 0.077 (0.026)                    |
| Individual   | 0.098 (0.051)                    |
| Extended input |                                |
| Single      | 0.100 (0.00)                     |
| Clustered    | 0.100 (0.028)                    |

**Table 5**

Results of best predictive models (note: mood has been scaled between 0 and 1, and the RMSE should also be interpreted on that scale).

| Set-up    | Best model | RMSE result (standard deviation) |
|-----------|------------|----------------------------------|
| Regular input |            |                                   |
| Single      | GRU 2 layer/GRU | 0.070 (0.00)                     |
| Clustered    | GRU        | 0.066 (0.023)                     |
| Individual   | LSTMP      | 0.086 (0.047)                     |
| Extended input |           |                                   |
| Single      | LSTM 1 layer/ LSTMP | 0.070 (0.00)                     |
| Clustered    | GRU        | 0.075 (0.026)                     |
number of imputed mood ratings in the test data is 20.5% larger than in the training data. The standard deviation of the mood ratings in the training data is 0.177, while it is 0.171 in the test data. Since missing values have been imputed by mean values, where in many case these mean values are repeated for multiple time steps (e.g. 5 days), it could cause our recurrent neural network models to perform better on the test set predictions.

### Table 6
Results single generic model.

| Network      | RMSE train score | RMSE test score | SD train scores | SD test scores |
|--------------|------------------|-----------------|-----------------|----------------|
| Base inputs  |                  |                 |                 |                |
| LSTM 1 layer | 0.091            | 0.073           | 0.001           | 0.001          |
| LSTM 2 layer | 0.090            | 0.074           | 0.002           | 0.003          |
| LSTMP        | 0.089            | 0.073           | 0.000           | 0.001          |
| GRU 1 layer  | 0.092            | 0.076           | 0.005           | 0.006          |
| GRU 2 layer  | 0.090            | 0.070           | 0.000           | 0.000          |
| GRUP         | 0.090            | 0.070           | 0.000           | 0.000          |
| Extended inputs |            |                 |                 |                |
| LSTM 1 layer | 0.090            | 0.070           | 0.000           | 0.000          |
| LSTM 2 layer | 0.090            | 0.074           | 0.000           | 0.006          |
| LSTMP        | 0.090            | 0.070           | 0.000           | 0.000          |
| GRU 1 layer  | 0.090            | 0.074           | 0.000           | 0.006          |
| GRU 2 layer  | 0.092            | 0.076           | 0.000           | 0.006          |
| GRUP         | 0.090            | 0.076           | 0.000           | 0.006          |

### Table 7
Results clustered models.

| Network   | RMSE train score | RMSE test score | SD train scores | SD test scores |
|-----------|------------------|-----------------|-----------------|----------------|
| Base inputs |               |                 |                 |                |
| LSTM 1 layer | 0.082            | 0.076           | 0.054           | 0.054          |
| LSTM 2 layer | 0.073            | 0.069           | 0.037           | 0.025          |
| LSTMP      | 0.074            | 0.069           | 0.038           | 0.026          |
| GRU 1 layer | 0.076            | 0.070           | 0.031           | 0.021          |
| GRU 2 layer | 0.072            | 0.067           | 0.034           | 0.025          |
| GRUP       | 0.073            | 0.066           | 0.035           | 0.023          |
| Extended inputs |          |                 |                 |                |
| LSTM 1 layer | 0.085            | 0.079           | 0.027           | 0.023          |
| LSTM 2 layer | 0.081            | 0.079           | 0.030           | 0.030          |
| LSTMP      | 0.079            | 0.078           | 0.031           | 0.030          |
| GRU 1 layer | 0.083            | 0.078           | 0.027           | 0.022          |
| GRU 2 layer | 0.079            | 0.079           | 0.027           | 0.024          |
| GRUP       | 0.078            | 0.075           | 0.028           | 0.026          |
Table 8
Results individual models.

| Network       | RMSE train score | RMSE test score | SD train score | SD test score |
|---------------|------------------|-----------------|----------------|---------------|
| Base inputs   |                  |                 |                |               |
| LSTM 1 layer  | 0.107            | 0.113           | 0.100          | 0.107         |
| LSTM 2 layer  | 0.085            | 0.088           | 0.044          | 0.050         |
| LSTMP         | 0.083            | 0.086           | 0.041          | 0.047         |
| GRU 1 layer   | 0.087            | 0.092           | 0.038          | 0.047         |
| GRU 2 layer   | 0.080            | 0.090           | 0.038          | 0.047         |
| GRUP          | 0.083            | 0.088           | 0.039          | 0.046         |

5. Conclusions and future work

This research work focused on using sophisticated machine learning algorithms to improve accuracy in predictions for short-term mood in the scope of blended therapy, taking advantage of log data generated from the Internet-based component. Applying recurrent neural networks in the domain of clinical time-series prediction seems promising, where we see that forming clusters from patients input data shows slightly better predictive performance than single and individual set-ups.

In our prediction of short term mood, EMA mood ratings history still remains the most significant input for predictive modeling. Applying response-time and other usage and adherence data has minor significance in predicting patients mood change. However, applying a 7-day time window to predict the mood of the 8th provided the best fit, which suggests some sort of weekly pattern and recent memory influence on people’s mood.

While we explored the prediction problem from an AI perspective, the use of the results in clinical settings, are usually more focused on longer term developments of patients, is something we will explore in the future. For example, to predict whether a treatment will be successful or not. In addition, we want to use different sensor and usage data from the mobile phone including activity data and log data from follow-ups to improve predictive performance. Also, following Hoogendoorn et al. (2016) we will try to exploit free text in the predictions as well.

Acknowledgments

This research has been performed in the context of the EU FP7 project E-COMPARED (project number 603098). We want to thank all trial coordinators in the countries where the MoodBuster platform has been used.

References

Asseleberg, J., Ruwaard, J., Eijds, M., Schrader, N., Sijbrandij, M., Riper, H., 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J. Med. Internet Res. 18 (3).
Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., Ruwaard, J., 2016. How to predict mood? Delving into features of smartphone-based data. In: Americas Conference on Information Systems, AMCIS 2016.
Cho, K., van Merrienboer, B., Güçlüehre, Ç., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR abs/1406.1078.
Chollet, F., et al., 2015. Keras. https://github.com/fchollet/keras.
Cuijpers, P., Smit, F., Oostenbrink, J., De Graaf, R., Ten Have, M., Beekman, A., 2007. Economic costs of minor depression: a population-based study. Acta Psychiatr. Scand. 115, 229–236.
Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., Szolovits, P., 2016. Predicting social anxiety treatment outcome based on therapeutic email conversations. IEEE J. Biomed. Health Inform.
Kleiboer, A., Smit, J., Bosmans, J., Ruwaard, J., Andersson, G., Topoozo, N., Berger, T., Krieger, T., Botella, C., Bairos, R., et al., 2016. European comparative effectiveness research on blended depression treatment versus treatment-as-usual (e-compared): study protocol for a randomized controlled, non-inferiority trial in eight European countries. Trials 17 (1), 387.
Kooistra, L.C., Ruwaard, J., Wiersma, J.E., van Oppen, P., van der Vaart, R., van Gemert-Pijnen, J.E., Riper, H., 2016. Development and initial evaluation of blended cognitive behavioural treatment for major depression in routine specialized mental health care. Internet Interventions 4, 61–71.
Riper, H., Andersson, G., Christensen, H., Cuijpers, P., Lange, A., Eysenbach, G., 2010. Theme issue on e-mental health: a growing field in internet research. J. Med. Internet Res. 12.
Sak, H., Senior, A.W., Beaufays, F., 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp. 338–342.
Sepp Hochreiter, J.S., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Shifman, S., Stone, A.A., Hufford, M.R., 2008. Ecological momentary assessment. Annu. Rev. Clin. Psychol. 4, 1–32.
Usun, T., Ayuso-Mateos, J.L., Chatterji, S., Mathers, C., Murray, C.J., 2004. Global burden of depressive disorders in the year 2000. Br. J. Psychiatry 184, 386–392.
van Breda, W., Hoogendoorn, M., Eiben, A., Anderson, G., Riper, H., Ruwaard, J., Vernmark, K., 2016a. A feature representation learning method for temporal datasets. In: Computational Intelligence (SSCI), 2016 IEEE Symposium Series on, IEEE, pp. 1–8.
van Breda, W., Pastor, J., Hoogendoorn, M., Ruwaard, J., Asselbergs, J., Riper, H., 2016b. Exploring and comparing machine learning approaches for predicting mood over time. In: Innovation in Medicine and Healthcare 2016. Springer, pp. 37–47.

A. Mikus et al.

Internet Interventions 12 (2018) 105–110