Mixed Effects Random Forests for Personalised Predictions of Clinical Depression Severity

Rob Lewis [1] Asma Ghandeharioun [§1] Szymon Fedor [§1] Paola Pedrelli [2] Rosalind W. Picard [1] David Mischoulon [2]

[§]Equal contribution [1] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA [2]The Depression Clinical and Research Program, Massachusetts General Hospital, Boston, MA, USA

Abstract

This work demonstrates how mixed effects random forests enable accurate predictions of depression severity using multimodal physiological and digital activity data.

- **Data**: physiological and digital activity features collected continuously and passively from 31 patients with major depressive disorder (MDD) in an 8-week study
- **Approach**: a mixed effects random forest (MERF) is compared to standard machine learning models (including random forests) and personal baselines when predicting clinical Hamilton Depression Rating Scale scores ($HDRS_{17}$). Three data scenarios are considered: *Random*, *Time Split*, and *User Split*
- **Results**: compared to the personal baselines, accuracy is significantly improved for each patient by an average of 0.199-0.276 in terms of mean absolute error ($p \ll 0.05$). Performance also far exceeds the standard random forest (and other machine learning baselines)
- **Significance**: this is noteworthy as these simple personal baselines frequently outperform machine learning methods in mental health prediction tasks [1]

However, we find that these improvements pertain exclusively to scenarios where labelled patient data are available to the model at training time. Investigating methods that improve accuracy when generalising to new patients is left as important future work.

Mixed Effects Random Forests

In this work we empirically assess the accuracy of a mixed effects random forest (MERF) on repeated measures $HDRS_{17}$ scores. We explicitly acknowledge the authors of the MERF theory [2] and of the opensource Python implementation [3] whose work we build upon in this paper.

- The method is referred to as *mixed effects* as it contains both *fixed effect* parameters – i.e., those that are shared by all patients in the dataset – and *random effect* parameters – i.e., those that are unique for each patient
- Beyond the random effect parameters, we are interested in the random forest component of this method (the *fixed effect*), given the random forest's ability to maintain performance when there are many more features than observations (i.e., in the $p \gg n$ context)

Model definition

- i=1, ..., m are *clusters* (i.e., patients) with n_i observations each (j=1, ..., n_i)
- $\boldsymbol{Y_i}$ is the regression target variable (n_i×1). $\boldsymbol{X_i}$ is a design matrix of input features (n_i×p) and $f(\boldsymbol{X_i})$ is the *fixed effect* random forest estimator
- $\boldsymbol{Z_i}$ is also a design matrix (n_i×q), that usually contains a subset of features from $\boldsymbol{X_i}$. $\boldsymbol{b_i}$ are *random effect* parameters (q×1) for each i, and $\boldsymbol{Z_i b_i}$ is assumed to be linear
- $\boldsymbol{\epsilon_i}$ is the measurement error for each i; and, $\boldsymbol{D}$, $\boldsymbol{R_i}$ and $\boldsymbol{V_i}$ are covariance matrices

The mixed effects random forest (MERF) is thus defined by:

$$\boldsymbol{Y_i} = f(\boldsymbol{X_i}) + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i} \quad (1)$$
$$\boldsymbol{b_i} \sim N(0, \boldsymbol{D}) \quad (2)$$
$$\boldsymbol{\epsilon_i} \sim N(0, \boldsymbol{R_i}) \quad (3)$$
$$\boldsymbol{V_i} = \mathrm{Cov}(\boldsymbol{Y_i}) = \boldsymbol{Z_i D Z_i^T} + \boldsymbol{R_i} \quad (4)$$
$$\boldsymbol{R_i} = \boldsymbol{\sigma^2 I_{n_i}} \quad (5)$$

The model parameters are fit using an expectation maximisation (EM) procedure with convergence monitored by a generalised log likelihood objective function.

Model expectation values for intercept-only mixed effects random forest

While $\boldsymbol{Z_i}$ may include many features, only a random intercept is used in the experiments of this paper. Thus, $\boldsymbol{Z_i}$ becomes a (n_i×1) vector of ones, and so (1-5) can be expressed for each observation ij as:

$$Y_{ij} = f(X_{ij}) + b_i + \epsilon_{ij} \quad (6)$$
$$E(Y_{ij}|b_i) = f(X_{ij}) + b_i \quad (7)$$
$$E(Y_{ij}) = f(X_{ij}) \quad (8)$$

Study Protocol, Data Collected and Experimental Settings

An 8-week observational study including 31 patients with Major Depressive Disorder

- 1,643 days of data were collected from 31 patients with major depressive disorder (MDD)
- Clinical assessments were performed by clinicians during 6 visits (once during screening followed by 5 bi-weekly visits during the 8-week monitoring period)
- These clinical scores include the $HDRS_{17}$, which is commonly used to measure depressive symptom severity in clinical trials. This is used as the **target variable** in the machine learning experiments. It is summed to create a total score of 0 to 52
- Multimodal data is also collected continuously and passively from study participants using mobile phones and physiological-sensor wristbands (Empatica E4)

The data collection and feature engineering were informed by prior work identifying biomarkers and correlates of depressive symptomatology

- **Electrodermal Activity (EDA)**: skin conductance level (SCL) and skin conductance response (SCR) are measured on the left and right wrists. Various statistics are derived from the raw values, both for each wrist individually and the difference between wrists
- **Heart Rate Variability (HRV)**: is measured on the left and right wrists. Various HRV metrics are calculated in the time and frequency domains. Heart Rate (HR) is also measured
- **Sleep**: sleep time is calculated (over 24 hours and during the night). Other sleep characteristics are calculated using actigraphy, such as sleep onset time, wakeups, maximal night uninterrupted sleep, and a sleep regularity index
- **Motion / Physical Activity**: motion frequency (i.e., fraction of time in motion within a period) and magnitude (i.e., the intensity of the motion) were calculated at various temporal aggregations using accelerometer data collected from the left and right wrists
- **Digital Activity**: an app (MovisensXS) was used to collect smartphone activity data, including streams for location, call and messaging (sms) activity, and app usage and screen on / off time
- **Environment / Weather**: location data were used to obtain historical weather information using the DarkSky API. Features include temperature, precipitation, humidity, and UV index

Experimental settings

- **Data Filtering**: the dataset contains 2,820 features, and its rows are filtered to only include data captured on days with clinical scores, resulting in 149 observations in total
- **Evaluation Scenarios**: 3 scenarios are considered i) *Random Split*: with train:test split of 70:30, ii) *Time Split*: initial 3 observations per patient are for training and the remaining for test, and iii) *User Split*: observations for one patient are held out from the model as a testing set, while observations from remaining patients are used to train (repeated 31 times)

Experimental Results: Comparison of Mixed Effects Random Forest by Group-level MAE to Personal Baselines and Standard ML Models

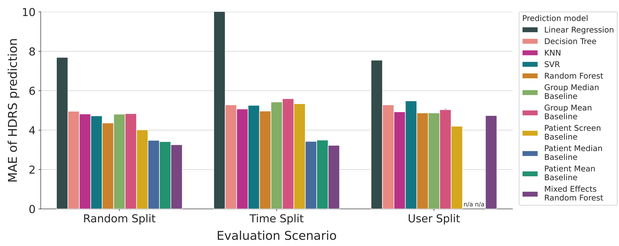

Figure 1. Group-level mean absolute error (MAE) of testing set $HDRS_{17}$ prediction by evaluation scenario and model type.

- In two scenarios – the *Random Split* and *Time Split* – the mixed effects random forest shows an improvement over both the standard random forest (as well as other ML baselines) and, more importantly, the patient median and mean baselines
- However, it is also noteworthy that the mixed effects approach provides less of an improvement in the *User Split* scenario, and indeed using the patient $HDRS_{17}$ score at screening is a far better predictor in this setting

Experimental Results: Participant-level User lift of Mixed Effects Random Forest over Personal Baselines

- The MAE is also calculated at the *participant-level* and used to derive a *user lift* metric, which represents the improvement of the MERF model over the baseline (e.g., if the baseline MAE is 4 and the MERF MAE is 3, then the *user lift* is 1)
- A corollary of this *participant-level* approach is that one can formally test if, on average, the *user lift* is significantly greater than zero
- To do so a one-sample one-tailed nonparametric permutation test is performed, from which the p-values are reported and considered significant if $p \leq 0.05$

Scenario	N. Seeds	Avg. PBL Err.	Avg. MERF Err.	Avg. User Lift	User Lift p-value
Random Split	10	3.349	3.165	0.199 (↑)	**0.000**
Time Split	10	3.450	3.174	0.276 (↑)	**0.004**
User Split	10	4.198	4.739	-0.541 (↓)	n/a

Table 1. *Participant-level* errors by scenario. PBL is: the *Patient Mean* if *Random* scenario; the *Patient Median* if *Time Split* scenario; else, the *Patient Screen* score if *User Split* scenario. **NB**: the permutation test is not calculated in the *User Split* scenario as the *user lift* is clearly less than zero.

Improvement over the personal baselines is reflected at the participant-level in the *Random Split* and *Time Split* scenarios, with permutation tests suggesting the lift is significant ($p \leq 0.05$).

Discussion and Future Work

These results suggest that a mixed effects approach allows random forests to significantly outperform baselines in $HDRS_{17}$ predictions when patients in the testing set have also contributed some data to the training set.

- The large improvement versus the standard random forest likely stems from the ability of the mixed effects model to fit a random effect intercept parameter for each patient ($\boldsymbol{b_i}$), which ensures predictions are adjusted by the average observed scores for each patient, cf. (7)
- That said, given the significant lift of MERF over personal average baselines, it is clear that the model learns more than just a *participant-level* intercept
- Indeed, it is probable that the additional lift is due to the random forest fixed effect terms, $f(\boldsymbol{X_i})$, that – when estimated using the EM procedure – learn relations between the features and $HDRS_{17}$ scores that can be shared across patients

Future work will seek to improve accuracy and understand generalisability to new tasks

- Low accuracy in the *User Split* scenario is a clear limitation, and may be improved by using patient characteristics to compute initial random effect parameter values for new patients
- Introducing additional random effect parameters may further improve accuracy in the *Random Split* and *Time Split* scenarios
- The model only uses data from days with clinical scores; thus, incorporating the additional days via time-lagged approaches should be considered
- Finally, we intend to repeat these analyses with alternative target variables and datasets (e.g., including individuals without a current MDD diagnosis) to further understand how this method generalises to related mental health prediction tasks

References

[1] O. Demasi, K. Körding, and B. Recht.
Meaningless comparisons lead to false optimism in medical machine learning.
PLoS ONE, 12, 2017.

[2] Ahlem Hajjem, François Bellavance, and Denis Larocque.
Mixed-effects random forest for clustered data.
Journal of Statistical Computation and Simulation, 84(6):1313–1328, 2014.

[3] MERF: https://github.com/manifoldai/merf.
Mixed effects random forest (Python), Manifold AI.
(Accessed: May 24, 2021).

[4] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F. Ionescu, Darian Bhathena, Lauren B. Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, Lisa Sangermano, David Mischoulon, Johnathan E. Alpert, and Rosalind W. Picard.
Monitoring changes in depression severity using wearable and mobile sensors.
Frontiers in Psychiatry, 11:1413, 2020.