Predicting Psychosis Using the Experience Sampling Method with Mobile Apps

Andrea Katrinecz*, Daniel Stamate*, Wajdi Alghamdi
Data Science & Soft Computing Lab,
and Department of Computing
Goldsmiths, University of London
London, United Kingdom
Email d.stamate@gold.ac.uk
*Joint first-authors

Daniel Stahl
Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology & Neuroscience,
King's College London
London, United Kingdom

ESM-MERGE Investigators

Philippe Delespaul, Jim van Os, Sinan Guloksuz
Department of Psychiatry and Psychology
Maastricht University Medical Centre
Maastricht, the Netherlands

***Abstract*—Smart phones have become ubiquitous in the recent years, which opened up a new opportunity for rediscovering the Experience Sampling Method (ESM) in a new efficient form using mobile apps, and provides great prospects to become a low cost and high impact mHealth tool for psychiatry practice. The method is used to collect longitudinal data of participants' daily life experiences, and is ideal to capture fluctuations in emotions (momentary mental states) as an early indicator for later mental health disorder. In this study ESM data of patients with psychosis and controls were used to examine emotion changes and identify patterns. This paper attempts to determine whether aggregated ESM data, in which statistical measures represent the distribution and dynamics of the original data, are able to distinguish patients from controls. Variable importance, recursive feature elimination and ReliefF methods were used for feature selection. Model training and tuning, and testing were performed in nested cross-validation, and were based on algorithms such as Random Forests, Support Vector Machines, Gaussian Processes, Logistic Regression and Neural Networks. ROC analysis was used to post-process these models. Stability of model performances was studied using Monte Carlo simulations. The results provide evidence that pattern in mood changes can be captured with the combination of techniques used. The best results were achieved by SVM with radial kernel, where the best model performed with 82% accuracy and 82% sensitivity.**

Keywords— Predicting psychosis, ESM, mHealth, SVM, Gaussian Process, Random Forests, Neural Networks, feature selection, ROC analysis, Monte Carlo

I. Introduction

Csikszentmihalyi and colleagues (1977) [1] developed a novel technique called Experience Sampling Method (ESM) or Ecological Momentary Assessment (EMA) that uses a structured diary approach to capture momentary mental states—*in other words, emotions (e.g. worried)*—in the context of daily life (*e.g., alone vs with company*) by asking participants to respond to randomly-repeated brief assessments (usually around 10 times/day) for a period of time (usually around a week) [2]. The main advantage of the technique over earlier self-report questionnaires is that the experiences are recorded in real time: right when and where they are experienced (there is no recall bias). The method yields rich longitudinal data, which allow for investigating dynamic flow of mental states. The early use of this method involved paper-and-pencil questionnaires and pagers, which were radio frequency devices that prompted the participants by a signal to complete a form. Although the method became more reliable, as it involved extensive manual processes, it was still cumbersome.

The evolvement of ubiquitous smartphones is a breakthrough in the development of the methodology: numerous ESM mobile apps were introduced in the recent years. As mobile phones have become a natural part of life, respondents are less likely to miss any beeps due to signal limitations or not having the prompting device with them. The process of answering the questions has become quicker and simpler. The latest developments make possible context information to be captured automatically by sensors such as heart rate or GPS location, enabling the study of interaction between mental and physiological processes in daily life. Due to the availability of the device and development of computational data processing, the ESM method has become extremely cost-effective. All these advantages encourage better response rates and more accurate results with less bias, and make it possible to obtain much larger sets of samples than earlier. The exploration of the new technique is still at its early stage, but the results are promising. A recent paper discusses potentials of the technique to be used in several clinical applications involving patients in the process of diagnosis and treatment, and shows how it can become a regular, low cost and high impact tool of clinical practice [2].

Challenges in Psychiatry include the difficulty in classifying mental health disorders, where typically no clear boundaries exist between classes, and the same symptoms can indicate different disorders. Furthermore, assessment of patients is complex, based on evaluation and measurement of psychological, biological and social factors gathered from interviews, examinations and medical history. Recent research shows an increased interest in developing novel statistical and machine learning approaches to predicting psychosis [3]–[5]. Psychosis is a mental health problem often involving hallucinations or delusions, causing those affected to perceive or interpret things differently from people around them. In this study, we aimed to build a predictive modelling approach to differentiating patients with psychosis from controls. In particular, we intended to see whether it is possible to capture the patterns and the dynamics in emotion changes applying combined machine learning algorithms and statistical methods to ESM data. This approach was explored for the first time in psychiatric researches.

II. Methods

A. Samples

Data were derived from the pooled ESM-MERGE dataset, which consisted of 510 variables and 98,480 observations, collected by 11 independent studies using the *PsyMate* mobile application [6]. The outcome variable *status* had 10 categories, and only two of them were kept for this analysis: *psychotic patients* and *controls*. This reduction retained 472 individuals including 260 patients with psychosis and 212 controls. The participants answered a set of questions 10 times a day for a period of six days, which resulted in 60 observations from each individual.

B. Variables

The variables extracted from the original dataset were the following:

- ***Subjective predictor variables***: our main interest was to examine ten of the emotion (momentary mental state) variables; seven negative and three positive feelings were extracted from the original dataset: *anxious, down, guilty, insecure, irritated, lonely, suspicious,* and *cheerful, relaxed* and *satisfied,* respectively. All these variables were measured on a Likert scale with an uneven scale of 1-7 representing the intensity of the feeling, which was treated as an interval-level scale in the analysis. An uneven scale is a symmetric scale with a neutral option represented by the middle value, in our case by level 4, which is also labelled verbally. This helps the interpretation of the scale levels to be unambiguous, bringing more reliability and validity into the study.
- ***Demographic predictor variables***:, *age* and *sex* were also included as predictors. *Age* was expressed in a continuous numeric format. *Sex* included levels *female* and *male*.
- ***Outcome variable***: *status* variable with levels *psychotic patient* and *control.*
- Variables *subject number*, *day number* and *beep number* were only used as a help during the aggregation process.

C. Pre-processing

Invalid or missing *age* and *sex* values could be corrected based on other data related to the patients. *Age* was rounded down to the closest integer. Emotion related variables, originally expressed in strings, were converted to numeric data type. All data related to longer than six days were removed, to ensure that only initial patient records were considered (not monitoring treatment outcome).

As a result of patients not responding to beeps, there was a significant amount of missing data in the set. Only 75% of the data were complete cases, the rest were mostly beeps with missing data in all emotion variables. This did not affect our analysis, as the machine learning algorithms were not directly applied to this set, but to an aggregated version of it.

D. *Data aggregation*

As previous researches have already highlighted, the variance in emotion changes is able to characterise patients vs controls [2]. Our aim was to capture these characteristics, and use them for classification.

Velocity (the emotion changes between the successive beeps) and *acceleration* (the change rate of the emotion changes, i.e. the change in velocity) were introduced to represent the dynamics in the data.

- *Calculating velocity:* a new column was added for each emotion variable, where the difference between the value of the respective beep and the previous beep was recorded. Only differences for consecutive beeps within a day were calculated, to consider only short time emotion changes, in other cases NA was recorded.
- *Calculating acceleration:* a new column was added for each emotion variable, where the difference between the velocity of the respective beep and the previous beep was recorded.
- *Calculating the absolute value of acceleration:* a new column was added for each emotion variable, where the absolute value of the acceleration was recorded. This measure did not distinguish cases with opposite direction of speeding, but rather focused on the size of the change.

Following this, data aggregation was carried out on all four versions of the emotion variables (base, velocity, acceleration and absolute value of acceleration), replacing the 60 beeps of each individual with distribution representative statistics of those 60 observations. This way each person was represented by one row of descriptive statistics reflecting the distribution of the data within that person's observations.

Based on which aggregated variables were included in the datasets, four different types of sets were created:

- **Base** = Base data
- **Velo** = Base data + velocity
- **Acc** = Base data + velocity + acceleration
- **Acc_abs** = Base data + velocity + acceleration in absolute value

Two different rules were used to aggregate the beep collected values:

- *V1:* Six new measures were introduced to represent each variable: the minimum and maximum value of all observations, the 0.25, 0.5 and 0.75 quantiles, and the interquartile range within each person.
- *V2:* Four new measures were used to represent each variable: the 0.1, 0.5 and 0.9 quantiles, and the interquartile range.

The above two aggregation rules applied to the four types of datasets created eight different aggregated datasets. Apart from the aggregated data, the demographic details of *gender* and *age* as predictors, and variable *status* as outcome variable also was added to each of the eight datasets. These eight datasets were used for further exploration in this analysis.

TABLE I. CALCULATING VELOCITY, ACCELERATION AND ABSOLUTE VALUE OF ACCELERATION DEMONSTRATED ON ANXIOUS EMOTION

Subj.	Day	Beep	Status	Anx.	Anx. velo	Anx. acc	Anx. abs(acc)
244	2	1	patient	2	NA	NA	NA
244	2	2	patient	4	2	NA	NA
244	2	3	patient	2	-2	-4	4
244	2	4	patient	1	-1	1	1
244	2	5	patient	1	0	1	1
244	2	6	patient	4	3	3	3
244	2	7	patient	4	0	-3	3
244	2	8	patient	3	-1	-1	1
244	2	9	patient	4	1	2	2
244	2	10	patient	3	-1	-2	2

'No variance' and high correlation removal

Some median velocity and acceleration variables were found with almost no variation in the data, most observations being zero. These were non-informative and were removed to eliminate noise.

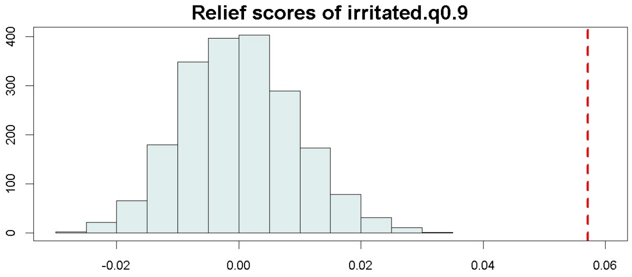
As positive and negative feelings usually come together, correlation was generally strong within the variables. Spearman rank correlation was computed on the variables in order to remove very high correlation as an option for pre-processing. Different cut-offs for correlation were tried such as 0.9, 0.85 and 0.8, to see which worked better for the predictive modelling performance.

E. Feature selection

Some models such as Logistic Regression, Neural Networks and Support Vector Machines are negatively affected by too large number of variables [7]. Three feature selection methods were tried for dimensionality reduction:

- feature ranking by importance using Learning Vector Quantization with repeated sampling
- a backwards feature selection method, recursive feature elimination built on the Random Forest algorithm
- ReliefF [8] feature selection with permutation test [9] based on 2000 random permutations. For instance, features with an observed Relief score corresponding to a distance of at least 1.96 standard deviations from the centre of the normal distribution built with the Relief scores repeatedly calculated 2000 times with permuted labels, were selected for further processing, based on the application of the permutation test with significance level alpha=0.05. The example of such a variable *irritated.q0.9* is indicated in Fig. 1.

FIGURE 1. THE DASHED LINE CORRESPONDING TO OBSERVED RELIEF SCORE=0.057 FOR IRRITATED.Q0.9 VARIABLE IS SUFFICIENTLY FAR AWAY FROM THE CENTRE OF NORMAL DISTRIBUTION OF THE VARIABLE'S RELIEF SCORES BASED ON 2000 PERMUTATIONS OF LABELS. THIS INDICATES PREDICTIVE POWER FOR IRRITATE.Q0.9.



As part of our objectives, we would like to gain more understanding of what variables have strong associations to the classes, to be used as practical clinical information. To achieve this, feature selection was performed on the best performing datasets, and the results were compared.

F. Principal component analysis

As an alternative dimensionality reduction method, PCA was performed on the eight datasets. As a large number of variables had a skewed distribution, Box-Cox transformation [7] was applied in this process to correct for skewness. A number of principal components were selected to cover over 80% of the variance in the data. The coordinates for the new dimensions were calculated for each row, and with the outcome variable *status* added, eight PCA datasets were created. The number of principal components used in the new datasets were between 8 and 15.

Logistic Regression, Support Vector Machine, Random Forest, Gaussian Process and Neural Network algorithms were applied to the new datasets to determine whether there was a significant association between the classes and the principal components.

We enlarged the data analyses process by creating 8 additional datasets, where PCA datasets were combined with the original sets, which were then used for model building with the Random Forest algorithm. Adding PCA variables to a decision tree based model allows using linear combinations of variables (given in this case by the principal components) in the test nodes. In this case the decision borders do not have to be parallel with the variable axes, allowing more flexibility in computing a class.

G. Machine learning techniques

Model training and tuning, and testing were performed in a nested cross-validation, comprising a 5-fold outer cross validation, and a 10-fold inner cross validation. Models were based on algorithms such as Random Forests, Support Vector Machines (linear, polynomial and radial kernel), Gaussian Processes (linear, polynomial and radial kernel), Logistic Regression (with and without stepwise model selection by *Akaike Information Criterion (AIC)*), and Neural Networks (with one hidden layer). Models were tuned in the inner cross validation based on the AUC criterion. ROC analysis was used to post-process these models by further splitting the hold-out folds from the outer cross validation, and using parts of these hold-out data for finding the best probability cut-offs for balancing sensitivity and specificity, and the other parts for testing the models.

H. *Monte Carlo simulation*

The stability of good models was tested using Monte Carlo simulation. The method involved a number of repetitions of the nested cross validation – in our case 100 times. Performance metrics of accuracy, area under the curve (AUC), Cohen's kappa statistic, sensitivity and specificity were evaluated and recorded in each experiment. The results were visualised using boxplots to capture the performance capability and stability of models. Finally, those models were chosen that consistently provided the best results.

I. *Hardware and software*

Monte Carlo simulation in our framework involving model tuning as part of the nested cross validation, is computationally expensive procedure, therefore a robust framework was required. Parallel processing was performed on a data analytics cluster of 11 servers with Xeon processors and 832GB fast RAM. The R software was used with a number of packages, including *caret*, *pROC*, *MASS*, *e1071*, *CORElearn*, *randomForest*, *ggplot2*, *data.table*, *mclust*, *stringi*, *spatstat*, *plyr*, *DMwR*, *arm*, *AppliedPredictiveModeling*, *doParallel*, *kernlab* and *H2O*.

III. RESULTS

A. *Predictive modelling*

The modelling results have brought some interesting and consistent findings.

It stands out, that nearly all of the 20 best performing models were based on datasets produced by *V2* aggregation. This indicates that minimum and maximum values of mood ratings are not the most important characteristics to distinguish patients and healthy people. Secondly, the 0.1 and 0.9 quantiles of the mood ratings are more informative to predict classes than the first and third quartiles.

Another remarkable aspect is that datasets also including acceleration information, especially in its normal form, were more likely to produce a successful model, than sets with only base data and velocity data.

The top 20 models were based on algorithms such as Random Forest, Gaussian Process, and Support Vector Machines with radial and polynomial kernels. Many models built with principal components also achieved good results. This confirms that there exists a pattern in the data, as several different techniques were able to capture it.

The best performing feature selection technique was the ReliefF method [8], therefore it was our major feature selection method in this study.

All the best three results were achieved by the datasets including base, velocity and acceleration data in normal values, and with *V2* aggregation applied. The very best result was produced by a Support Vector Machine with radial kernel on the dataset with Spearman correlation over 0.9 removed and feature selection performed by the ReliefF method, in which only variables with an observed Relief score corresponding to a p-value lower than 0.1 in the permutation test were retained. The second best result was achieved by a Support Vector Machine with polynomial kernel on the dataset after the same correlation removal and feature selection process. The third best performer was a Gaussian Process algorithm with radial kernel (GP Radial), performed on the principal components of the dataset. The performances of the best 3 models are displayed in *Tables II* and *III*.

TABLE II. BEST RESULTS OF THE 3 TOP PERFORMING MODELS

Method	**Variables**	**Aggregation**	**Dim Red**	**AUC**	**Sens**	**Spec**	**Accur**	**Kappa**
SVM Radial	*acc*	*V2*	*ReliefF+ Corr rem*	0.8639	0.8192	0.8255	0.8220	0.6419
SVM Poly	*acc*	*V2*	*ReliefF+ Corr rem*	0.8435	0.7885	0.8113	0.7987	0.5959
GP Radial	*acc*	*V2*	*PCA*	0.8216	0.7808	0.7925	0.7860	0.5700

TABLE III. MONTE CARLO 100 EXPERIMENTS
AVERAGE RESULTS OF THE 3 TOP PERFORMING MODELS

Method	**Variables**	**Aggregation**	**Dim Red**	**AUC**	**Sens**	**Spec**	**Accur**	**Kappa**
SVM Radial	*acc*	*V2*	ReliefF+ Corr rem	0.8459	0.7706	0.7957	0.7819	0.5623
SVM Poly	*acc*	*V2*	ReliefF+ Corr rem	0.8300	0.7481	0.7828	0.7637	0.5264
GP Radial	*acc*	*V2*	PCA	0.8157	0.7582	0.7535	0.7561	0.5093

FIGURE 2. BOXPLOTS SHOWING PERFORMANCE RESULTS OF THE TOP 3 MODELS IN MONTE CARLO 100 EXPERIMENTS

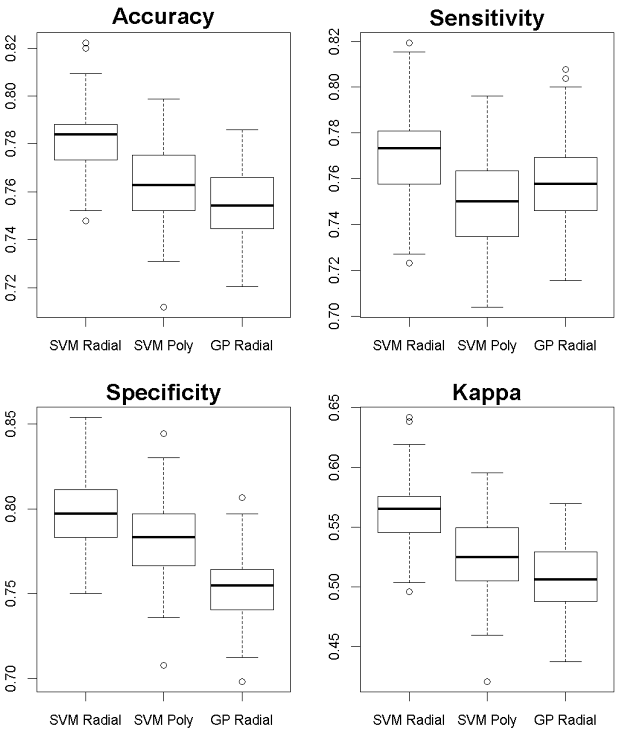
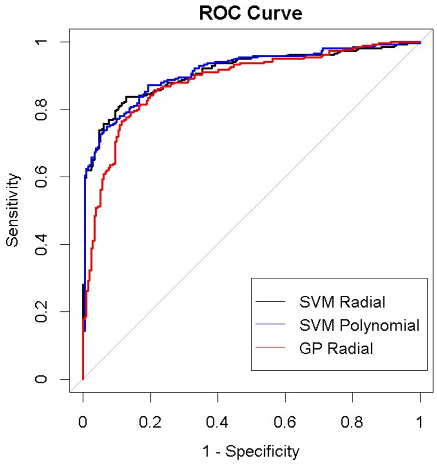


FIGURE 3. ROC CURVES OF THE 3 TOP MODELS



The performances of the top 3 models are reflected also in Figures 2 and 3.

Random Forest was slightly behind in performance, achieving accuracy results around 73%. Generally, the algorithm worked best on the datasets comprising also the principal components. Neural Networks and Logistic Regression algorithms performed with around 70% accuracy.

B. *Feature analysis*

Feature analysis was carried out on the *acc* dataset with *V2* aggregation applied, as this dataset was the most successful in predicting classes. *Table IV* shows the top results of feature selection with variable importance of learning vector quantisation model, recursive feature elimination, and ReliefF methods applied to this dataset. Both the varImp(LVQ) and RFE methods show several acceleration measures amongst the top features, while none of them have highlighted importance of any velocity measures. The ReliefF method showed only velocity related variables amongst the top ten, and no acceleration measures.

Acceleration is calculated as the difference between two consecutive velocity values (i.e. emotion changes), therefore velocity only considers emotion levels at two consecutive beeps, and acceleration considers three. If an emotion was changing slowly in the same direction within three consecutive beeps, acceleration was small, but if an emotion changed direction, acceleration was higher. This way acceleration is able to capture *'emotion spikes'* ('up-and-downs'), while velocity only captures one step of *emotion change* ('up' or 'down'). Acceleration is useful with variables if quick jump in emotion (large value in velocity) can normally occur in both classes, but only patients with psychosis are likely to have these dramatic changes also in the opposite direction within the short period of three beeps time.

The most informative statistical measure was the *0.9 quantile*, and for acceleration variables also the *interquartile range*.

The most occurring variables were *anxious* and *insecure,* both in their emotional level as well as in their acceleration forms. *Suspicious* occurred in its emotional level and its velocity forms. *Cheerful*, *feeling down* and *lonely* carried information in their emotional level form. *Satisfied* and *relaxed* variables also held little predictive information in their level form. The least power was shown by *irritated*, *guilty* and *gender*.

TABLE IV. LEARNING VECTOR QUANTISATION (LVQ), RECURSIVE FEATURE ELIMINATION (RFE) AND RELIEFF FEATURE SELECTION RESULTS USING *V2 AGGREGATION DATA*

Score	varImp(LVQ)	RFE	ReliefF
1	acc.anxious.interq	cheerful.q0.1	cheerful.q0.1
2	insecure.q0.9	Age	relaxed.med
3	acc.anxious.q0.9	acc.anxious.interq	velo.guilty.q0.1
4	down.q0.9	satisfied.q0.1	relaxed.q0.9
5	lonely.q0.9	lonely.q0.9	velo.irritated.q0.9
6	cheerful.q0.1	acc.satisfied.inter	down.q0.9
7	anxious.q0.9	suspicious.q0.9	insecure.q0.9
8	acc.insecure.interq	acc.anxious.q0.9	velo.suspicious.interq
9	insecure.interq	acc.insecure.interq	suspicious.q0.9
10	down.interq	lonely.interq	velo.suspicious.q0.1
Common in top 20: cheerful.q0.1, insecure.q0.9			

IV. DISCUSSION AND CONCLUSION

Several machine learning methods were explored, and all of them were able to recognize patterns differentiating the two classes to a certain level, which shows that this is a sound ground for further exploration. These models were further tested with Monte Carlo experiments, and they consistently yielded adequate predictive power and stability. The best performing models were Support Vector Machines with radial kernel, achieving as high as 82% accuracy in some cases, and an average performance of 78% accuracy in Monte Carlo simulations with 100 repetitions.

By evaluating the discriminative power of variables across models, we found that the level of emotions shows good predictive power for a few variables, such as *anxious*, *insecure, suspicious, feeling down*, *lonely*, and *cheerful*. Rather than relying only on the experienced level of emotion, in this study we also attempted to inspect the effect of mood changes onto our model performance, therefore we implemented the measure of velocity (i.e. change in mood) and acceleration (i.e. change in velocity), and both were successful in increasing the predictive power of the models. This is consistent with previous researches, which showed that the variance in emotion changes was beneficial in predicting patients with psychosis. Feature selection methods, variable importance and the most successful models highlighted that acceleration often better represents the dynamics of mood changes in predictive models, than velocity. This indicates that inspecting mood changes in three steps rather than two, being able to capture successive 'up-and-downs' rather than individual 'ups' or 'downs', helps to yield better predictions. The acceleration in variables *anxious* and *insecure* were especially successful in adding predictive power to the models.

The forthcoming phases of the work envisage: (i) external validation of the algorithm by applying it to an independent dataset and (ii) refining the predictive modelling approach to embrace the multi-level structure of data, whereby repeated observations at each beep nested within days that were further nested within individuals.

Given the sufficient performance of generic ESM items of the PsyMate application (excluding psychopathology specific mental state: "suspicious") in current models, the future work will extend the current project by building a detection system for mental illness in general. This work will leverage data from a large general population cohort consisting of 6 days of ESM data and a wide-range of clinical, behavioural, genetic, environmental variables collected from over 800 participants.

Overall, this proof of concept work demonstrated that symbiotic machine learning and statistical models could harness the power of ESM data in predicting mental illness as a low-cost high-impact self-monitoring tool with the ease and convenience of current mobile technology.

REFERENCES

[1] M. Csikszentmihalyi, R. Larson, and S. Prescott, 'The ecology of adolescent activity and experience', *J. Youth Adolesc.*, vol. 6, no. 3, pp. 281–294, Sep. 1977.

[2] J. van Os *et al.*, 'The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice', *Depress. Anxiety*, vol. 34, no. 6, pp. 481–493, 2017.

[3] O. Ajnakina *et al.*, 'Utilising symptom dimensions with diagnostic categories improves prediction of time to first remission in first-episode psychosis', *Schizophr. Res.*

[4] W. Alghamdi *et al.*, 'A Prediction Modelling and Pattern Detection Approach for the First-Episode Psychosis Associated to Cannabis Use', in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 825–830.

[5] R. Iniesta, D. Stahl, and P. McGuffin, 'Machine learning, statistical learning and the future of biological research in psychiatry', *Psychol. Med.*, vol. 46, no. 12, pp. 2455–2465, 2016.

[6] J. T. W. Wigman *et al.*, 'Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach', *Psychol. Med.*, vol. 45, no. 11, pp. 2375–2387, 2015.

[7] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer New York, 2013.

[8] I. Kononenko, 'Estimating Attributes: Analysis and Extensions of RELIEF', 1994, pp. 171–182.

[9] P. I. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2000.