Identifying psychosis spectrum disorder from experience sampling data using machine learning approaches

Daniel Stamate[1,7*], Andrea Katrinecz[1*], Daniel Stahl[2], Simone J.W. Verhagen[3], Philippe A.E.G. Delespaul[3], Jim van Os[3,4,5], Sinan Guloksuz[3,6]

[1]Data Science & Soft Computing Lab, and Department of Computing, Goldsmiths, University of London, London, UK

[2]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

[3]Department of Psychiatry and Neuropsychology, Maastricht University Medical Centre, The School for Mental Health and Neuroscience, Maastricht, the Netherlands

[4]Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Centre Utrecht, Utrecht, the Netherlands

[5]King's College London, King's Health Partners, Department of Psychosis Studies, Institute of Psychiatry, London, UK

[6]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

[7]Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, University of Manchester, Manchester, UK.

[*] Daniel Stamate and Andrea Katrinecz are joint first authors.

Corresponding author: Sinan Guloksuz
Department of Psychiatry and Psychology, Maastricht University Medical Centre, PO BOX 616, 6200 MD, Maastricht, the Netherlands
Tel: +31-43-388-4071
Fax: +31-43-388-4122
Email: sinan.guloksuz@maastrichtuniversity.nl

Word Count
Abstract: 249
Main text only:
References:
Tables: 2
Figures: 3

Abstract

The ubiquity of smartphones have opened up the possibility of widespread use of the Experience Sampling Method (ESM). 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 indicator for later mental ill-health. In this study, ESM data of patients with psychosis spectrum disorder and controls were used to examine daily life emotions and higher order patterns thereof. We attempted to determine whether aggregated ESM data, in which statistical measures represent the distribution and dynamics of the original data, were able to distinguish patients from controls in a predictive modelling framework. Variable importance, recursive feature elimination, and ReliefF methods were used for feature selection. Model training, tuning, and testing were performed in nested cross-validation, 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 performance was studied using Monte Carlo simulations. The results provide evidence that patterns in emotion changes can be captured by applying a combination of these techniques. Acceleration in the variables anxious and insecure was particularly successful in adding further predictive power to the models. The best results were achieved by Support Vector Machines with radial kernel (accuracy=82% and sensitivity=82%). This proof-of-concept work demonstrates that synergistic machine learning and statistical modeling may be used to harness the power of ESM data in the future.

1. Introduction

The evolution of mobile technology opens up new avenues for psychosis research. The ubiquitous smartphones and wearables provide researchers with convenient, low-cost tools for collecting rich longitudinal data of various measures: momentary mental states through the Experience Sampling Method (ESM), cognition, physiological parameters (e.g., sleep, heart-rate), mobility traces via global positioning system (GPS), and smart-phone user data (e.g., patterns of social media use and texting) (Torous and Keshavan, 2018).

Prior to advances in mobile technology, ESM (also called Ecological Momentary Assessment [EMA]) has been extensively used in psychosis research to unravel contextualized dynamical changes. These studies have shown that the method is feasible and valid to assess psychotic experiences in daily life, offering the unique opportunity to link symptom variability to environmental experiences (Oorschot et al., 2009). ESM research helped unravel differences between stress-sensitivity in psychotic patients, ultra-high risk patients and healthy controls, particularly showing a stronger influence of momentary stress on affective and psychotic symptoms in the early phases of psychosis (van der Steen et al., 2017). With technological advances, clinical implementations of ESM are possible (Hartmann et al., 2015) and personalized networks can be created revealing in-debt insight in how momentary experiences and symptoms are related, informing treatment strategies for psychosis on an n=1 level (Bak et al., 2016). ESM, a structured diary approach, offers several advantages over traditional self-report questionnaires used in mental healthcare and research, including reduced recall bias and assessment error due to the frequent recording of momentary mental states at multiple time points, which increase the validity and reliability of the method (Verhagen et al., 2016). Recently, technology has expanded to smartphone use (Pot-Kolder et al., 2018) enabling to obtain larger samples with increased response rates, high accuracy, and less bias. We recently discussed how ESM could be utilized to address several clinical challenges across diagnosis, and proposed this method as a regular, low-cost and high-impact mHealth tool for clinical practice, involving patients in the process of diagnosis and treatment (van Os et al., 2017b). In this conceptualization paper**,** an example was shown how change in momentary mental states and reactivity to contextualized factors (current location) could be used to differentiate depressed patients from controls (van Os et al., 2017b).

However, the intense ESM assessment produces a massive amount of information at an individual level. The complexity of data poses a challenge for traditional statistical approaches to provide optimal solutions for prediction and classification purposes using ESM data. In this regard, statistical machine learning (ML) may offer improved solutions for harnessing ESM data to gain insight into psychosis at a person-level. To the best of our knowledge, apart the study that we propose, there exists no research that applied ML to ESM data.

In fact, aside from the neuroimaging field (Winterburn et al., 2017), the exploration of ML in the field of clinical psychosis research is still at its early stage (Iniesta et al., 2016; Schnack, 2017), but the field is rapidly adopting these contemporary methods to handle challenges of psychosis research (Tandon and Tandon, 2018). With a pragmatic approach, the first ML applications aim to identify predictors of treatment, clinical outcomes, and diagnostic classifications (Alghamdi et al., 2016; Stamate et al., 2017; Cannon et al., 2016; Fusar-Poli et al., 2017; Jauhar et al., 2018; Koutsouleris et al., 2016; Mothi et al., 2018; Stamate et al., 2018b). These early findings demonstrate the potential of ML in handling complex data in clinical psychosis research.

Given that, in this proof-of-concept study and in our preliminary work that this study extends (Stamate et al., 2017), we leveraged a large ESM dataset —for the first time— to build a predictive modelling approach to differentiating patients with psychosis spectrum disorder (PSD) from controls and aimed to explore whether it is possible to capture the patterns and the dynamics in changes in intensity rating of emotions (velocity, acceleration) by applying ML algorithms to ESM data.

2. Methods

2.1. Study sample

Data were derived from the pooled ESM-MERGE dataset, which consisted of 510 variables and 98,480 observations. The original data came from 11 studies, however, as 4 of them did not provide the emotion variables we were interested in, we finally chose to include 7 studies in our dataset. Detailed information on characteristics of individual studies included in the pooled ESM-MERGE dataset were provided elsewhere (Wigman et al., 2015). The 7 studies used in our work are the following: Aripiprazole study (Lataster et al., 2011), Maastricht Coping Study (Bak et al., 2009), Maastricht Psychosis Study (Myin-Germeys et al., 2001), ZAPP study (Thewissen et al., 2008), Genetic Risk and Outcome of Psychosis study (Collip et al., 2011), Stress-reactivity in Psychosis study (Lataster et al., 2013) and PREVENT study (van der Steen et al., 2017). Although the data was collected through different studies, the conditions were very similar, which made it possible for us to use them as one large merged dataset. ESM is characterised by uncontrolled settings, and there is no site involved. Moreover, a major requirement of ESM data is that all interventions are minimised, and data collection happens while participants live their normal everyday life. To rate the intensity of emotion, the same Likert scale with 7 levels was provided in all studies, in which the minimum, maximum and middle levels were also expressed verbally to help a uniform understanding. We kept 2 categories from the outcome variable *status*: *patients with psychotic disorders* and *controls*. This reduction retained 472 individuals including 260 patients diagnosed with psychotic disorders and 212 controls.

2.2. Experience Sampling Method

The Experience Sampling method (ESM), is a validated, structured diary approach to capture momentary mental states (emotions) in the context of daily life, using repeated assessments and alerting participants by means of prompts (Delespaul, 1995; Oorschot et al., 2009; van Os et al., 2017a).

The participants answered a set of questions (emotion, current context, company, and appraisal of the current situation) at quasi-random moments signaled by the beep of their smart phone, 10 times a day for a period of six consecutive days, which resulted in a maximum of 60 observations from each individual. Our main interest was to examine ten of the emotion (momentary mental state) variables: three positive (c*heerful, relaxed* and *satisfied),* six negative (*anxious, down, guilty, insecure, irritated, lonely)* and one psychosis specific item (*suspicious)*. The wording of the questions was as follows: “At this moment I feel (e.g. insecure)”. All these variables were measured on a Likert scale with an uneven scale of 1-7 representing the intensity of the feeling (1 = *not at all* to 7 = *very*).

2.3. Variables

The 10 emotion variables, demographic variables *age* and *sex*, and outcome variable *status* were included in the predictive model training.

2.4. Data pre-processing

Only data from the first six ESM days were used to ensure that solely initial patient records were considered (not monitoring treatment outcome).

As a result of the beep schedule (semi-random beeps were programmed between 7.30 AM and 10.30PM) and participants not always being able to respond to beeps due to their varying day rhythms, there were natural missings in the data set (Delespaul, 1995). 75% of the observations were fully complete, i.e. included data for all emotion variables, which is an expected completion rate for this type of intensive ESM data collection (Delespaul, 1995). Looking at all emotion variables, 27.3 % of patient data and 20.5% of control data were missing, which provided us with sufficient data for both groups, even though the missing data related to patients were slightly higher. The amount of missing data was nearly the same in all emotion variables, as typically data was missing for a whole row due to not responding to a beep, which means that the effect of missing data on the various variables was similar, with no risk of causing a bias. We were only interested in obtaining sufficient consecutive observations per participant that could be used for aggregation by examining short term emotion changes. The missing data did not affect the analysis, as the ML algorithms were not directly applied to the raw dataset, but to an aggregated version of it.

2.5. Person-level data aggregation

As highlighted by previous work, fluctuation in the intensity rating of emotion is able to characterize patients versus controls (van Os et al., 2017b). Our aim was to capture these characteristics and use them for classification. As nested data can not be directly used as input for the ML modeling, for each individual we used aggregation on the emotion variables for the values to be represented by statistics.

To prepare our dataset for aggregation, we first grouped the data by individuals, and sorted the observations to keep the longitudinal order within each person. These 10 columns will be referred as the “base” data. Additionally, we introduced “velocity” (changes in the intensity rating of emotion between successive beeps) and “acceleration” (the change rate of the velocity) to represent the dynamics in the data (Supplementary Table S1). Velocity was defined as the difference between the value of the respective beep and the previous beep in the base data. To consider only short-time changes in the intensity rating of emotion, only differences for consecutive beeps within a day were calculated, otherwise N/A was recorded. Acceleration was defined as the difference between the velocity of the respective beep and the previous beep, while the absolute value of acceleration (irrespective of the direction of change) was defined to focus on the size of the change. We note that velocity only considers emotion levels at two consecutive beeps, and acceleration considers three. If an emotion was changing in the same rate in the same direction (e.g., both positive increases on the Likert scale) within three consecutive beeps, the acceleration score was small, but if an emotion changed direction (e.g. a decrease instead of an increase on the following moment), the acceleration score 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’).

Velocity, acceleration and absolute value of acceleration was calculated for each base emotion variables, resulting in 40 columns nested within each individual. Four different datasets were created including subsets of the above 40 columns, which we will refer to as below:

Base=Base data (10 columns)

Velo=Base data+velocity (20 columns)

Acc=Base data+velocity+acceleration (30 columns)

Acc_abs=Base data+velocity+ absolute value of acceleration. (30 columns)

Following this, data aggregation to the person-level was carried out on all four versions of the emotion variables (i.e., base, velocity, acceleration, and absolute value of acceleration). Each person was therefore represented by one row of descriptive statistics reflecting the distribution of the data within that person's observations. Two different rules,version 1 (V1) and version 2 (V2) as defined below, were applied to the above four datasets, thereby creating eight different aggregated datasets:

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.

Sex and age as predictors and status as outcome variable were also added to each dataset. These eight aggregated datasets were used for further processing(Fig.1), and they will be referred to as "aggregated datasets".

2.6. ‘No variance’ and high correlation removal

Some 0.5 quantile (median) velocity and acceleration variables showed almost no variation with most observations being zero, they were considered non-informative and thus removed to eliminate noise.

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

2.7. Feature selection

Some models such as Logistic Regression, Neural Networks and Support Vector Machines are negatively affected by a large number of variables (Kuhn and Johnson, 2013), therefore we performed feature selection on datasets prior to developing the predictive models. Note that we did not create new versions of the eight datasets at this stage, as the feature selection was part of the modeling script and was applied to the training set (Fig.1).

Three feature selection methods were considered for dimensionality reduction: (i) Feature ranking by importance using Learning Vector Quantization (LVQ) with repeated sampling, (ii) Recursive Feature Elimination (RFE) built on the Random Forest algorithm, (iii) ReliefF (Kononenko, 1994) feature selection with permutation test (Rudolph, 1995) based on 2000 random permutations.

In a separate process from modeling, we also utilised the feature selection methods and applied them to the best performing dataset to gain a better understanding of which predictors have strong associations with the outcome variable status.

2.8. Principal component analysis

As an alternative dimensionality reduction method, principal component analysis (PCA) was performed on the eight aggregated datasets. As many variables had a skewed distribution, Box-Cox transformation (Kuhn and Johnson, 2013) was first applied. The most important 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 new datasets were created, which we will refer to as “PCA datasets”. The number of principal components used in the new datasets varied between 8 and 15.

Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gaussian Process (GP), and Neural Network (NN) algorithms were applied to the new datasets to determine whether there was a significant association between the outcome variable status and the principal components.

We further extended the data analyses process by creating eight additional datasets, where PCA datasets were combined with the original data sets; we will refer to these sets as “mixed PCA” datasets. They were only used for model building with the RF algorithm. Adding PCA variables to a decision tree-based model such as random forest allows using linear combinations of variables (given in this case by the principal components) in the decision trees’ test nodes. In this case the decision borders do not have to be parallel with the variable axes, allowing more flexibility in predicting the outcome variable with the RF model.

2.9. Machine Learning based predictive modeling process

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 RF, SVM with linear, polynomial and radial kernels, GP with linear, polynomial and radial kernels, LR with and without stepwise model selection by Akaike Information Criterion, and NN with one hidden layer. Models were tuned in the inner cross-validation based on maximizing the area under the ROC curve (AUC). ROC analysis was then applied 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. Figure 1 provides an illustration of the ML-based prediction modeling process.

2.10. Monte Carlo simulation

The stability of the models was tested using Monte Carlo simulation. The method involved a number of repetitions of the nested cross-validation (100 times). Performance metrics of accuracy, AUC, Cohen's kappa statistic, sensitivity, and specificity were evaluated and recorded in each experiment, and those models were chosen that consistently provided the best results.

2.11. Hardware and software

Monte Carlo simulation involving model tuning as part of the nested cross-validation is a 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. 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, and kernlab.

3. Results

3.1. Predictive modelling

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

It's also noticeable that datasets including acceleration information were more likely to produce a higher performance predictive model than sets with only base and velocity data.

The top 20 models were based on algorithms such as RF, GP, and SVM with radial and polynomial kernels. Many models built with principal components also achieved good results. This confirms that patterns with predictive value exist in the data, as several different techniques were able to capture them. The best performing feature selection technique was the ReliefF method (Kononenko, 1994), 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, and with the V2 aggregation rule applied (Figure 2). The best result was produced by a SVM model with radial kernel (SVM Radial) 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 SVM model with polynomial kernel (SVM Poly) on the dataset after the same correlation removal and feature selection process. The third best performer was a GP model with radial kernel (GP Radial), performed on the principal components of the dataset. The performances of the best three models are displayed in Table 1, Figure 2 and 3. Predictive models obtained with the other algorithms, i.e. RF, NN and LR, were behind in performance. Generally, the RF algorithm worked best on the datasets comprising the principal components or a mixture of principal components and base emotion variables, and achieved accuracy results around 73%. NN and LR algorithms performed with around 70% accuracy.

3.2. Feature analysis

Feature analysis was carried out on the acceleration dataset with the V2 aggregation rule applied, as this dataset proved to be the most successful in predicting the outcome variable status. Table 2 shows the top results of feature selection with variable importance of the LVQ model, RFE, 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 highlighted importance of any of the velocity measures. The ReliefF method showed only velocity-related variables amongst the top ten, and no acceleration measures.

According to our results, accelaration often better distinguishes patients with a psychotic disorder from controls than velocity, which means that quick jumps in emotion (i.e., a large value in velocity) is more likely to normally occur in both statuses, while only patients with a psychotic disorder are likely to have these dramatic changes also in the opposite direction within the short period of three beeps time (Stamate et al., 2017).

As suggested by Table 2, the most informative statistical measure was the 0.9 quantile, and for acceleration variables also the interquartile range. The most frequently 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 some predictive information in their level form. The least power was shown by irritated, guilty and sex (Table 2).

4. Discussion

To the best of our knowledge, this is the first study that applied machine learning to ESM data to build predictive modeling for psychosis spectrum disorder (we reported preliminary results of this study in the extended abstract Stamate et al., 2017). With the aim of establishing feasibility and preliminary performance of predictive modeling for efficient ESM use in clinical trials, this proof-of-concept study leverages a subset of existing ESM data from patients with PSD and healthy controls to generate an actionable output.

4.1. Interpretation of findings

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

By evaluating the discriminative power of variables across models, we revealed that the level of emotions shows good predictive power for several emotion items: anxious, insecure, suspicious, down, lonely, and cheerful. Rather than relying only on the intensity rating of emotion, in this study we also attempted to inspect the effect of mood changes onto our predictive 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 research, which showed that the fluctuation in the intensity rating of emotion could distinguish patients suffering from mental illness from controls (van Os et al., 2017b). Feature selection methods, variable importance, and the most successful models highlighted that acceleration often better represents the dynamics of mood changes than velocity in predictive models. 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 was especially successful in adding further predictive power to the models.

4.2. Limitations

Although this proof-of-concept study used a large ESM dataset, it has several limitations that need to be addressed in forthcoming phases of the work.

First, the current algorithms should be applied to completely independent datasets for external validation. We should however note here that the methodology that we employed based on nested cross-validation which is repeated 100 times in a Monte Carlo simulation as illustrated in Figure 1, leads to an extended internal validation. The latter could provide good estimates of the performance of the predictions that would be obtained on an external dataset if it were available, if we reasonably assume that the external dataset's underlying variable distributions would not change with respect to the used dataset. This is one of the main advantages of the use of the nested cross-validation, although it is more computationally intensive especially if repeated in a Monte Carlo simulation as in our approach.

Second, the current predictive modeling approaches could be tested against multi-level models (repeated observations at each beep nested within days that were further nested within individuals). In the future, we aim to leverage readily accessible data from a large general population cohort consisting of six days of ESM data and a wide-range of clinical, behavioral, genetic, environmental variables collected from over 800 participants (Pries et al., 2017).

4.3. Implications for clinical and research practice

There is a growing interest in mobile health and smartphone applications for mental disorders (Torous et al., 2017). Owing to increasing access to affordable mobile devices and development of computational data processing for uncovering patterns at an individual level, ESM may emerge as an extremely cost-effective and easy to implement mHealth tool at large.

Several feasibility studies showed that mobile device delivered ESM assessments are well accepted even among patients with serious mental illness, with similar completion rates to those in other diseases(Edwards et al., 2016; Kimhy et al., 2006; Moitra et al., 2017).

ESM data are able to capture the ability to bounce back from mental disturbance (e.g., suddenly feeling scared as an effect of hearing voices). For instance, healthy people normally show resilience and stability, returning to their baseline state rapidly. On the contrary, consistent with our current findings, patients and population at-risk spend a longer time in the altered state, not only with lower mood levels than controls but also with higher reactivity to the daily-life stressors (van de Leemput et al., 2014; van Os et al., 2017b). Similarly, a greater autocorrelation (i.e., the similarity of states as a function of a time-lag between them) in positive emotions is associated with a better recovery rate and prevention of depressive episodes (Hohn et al., 2013). Therefore, ESM data can be very useful in the assessment of subtle and transitory manifestations of behavioral and emotional changes that may be early warning signs of psychopathology, with the ability to capture contextualized momentary variation of mental states (Nelson et al., 2017).

Given the sufficient performance of generic ESM items (excluding the psychopathology specific mental state item: “suspicious”) in current models, the future project will seek to extend the current analyses attempting to build a detection system for early recognition of psychosis using mobile ESM technology. The future project will also attempt to tackle an important challenge that poses a threat to sustained engagement to ESM application: burdensome data collection procedure. Although fine-grained assessment of various context-dependent emotions and behaviors enhances delineating the current mental state, a fraction of participants often stop using mobile applications after several weeks of use. To improve application usability, we propose three solutions to be tested in future implementation trials: (I) using an abbreviated version with a particular focus on fluctuating emotional items rather than an extensive list of emotional items; (II) reducing the number of beeps per day; (III) leveraging sensor and user data from mobile devices to construct a two-level detection system.

Modern mobile platforms (e.g., smartphones and wearables) produce a massive amount of personal data from sleep and activity levels to social media use and basic phone usage (text and voice call); however, this raw material alone is of minimal value for predicting mental health problems due to low specificity and high between-person variance (Mohr et al., 2017). In this regard, a two-level detection system combining an abbreviated ESM technique and unobtrusive collection of smartphone user data may offer a feasible yet accurate active learning algorithm for predicting clinical vulnerability and mental health problems. In this system, once these two layers of data are successfully paired after a personal optimization period aimed to maximize within-individual consistency, active learning algorithms processing passive user data in the background can be trained to activate an ESM protocol when deviations from personal routine were observed. A recent study analyzing sensor data collected by a smartphone up to 8.5 months in a sample of 21 patients with schizophrenia showed that passive sensing data (e.g., levels of physical activity, changes in sleep rhythm, voice-call activity) were associated with responses to ESM items, and therefore might be of use in predicting mental health outcomes (Wang et al., 2016). By applying ESM approach along with wearable technology for measuring cardiac autonomic regulation over 36 hours, researchers demonstrated that momentary increases in autonomic arousal forecasted increases in ESM-measured auditory hallucinations severity (Kimhy et al., 2017).

4.4. Conclusion

In this proof-of-concept work, we demonstrated that machine learning 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. We are still in the very early stages of mobile-health implementation in psychiatry; however, if successfully developed, a mobile platform for early detection has the potential to help achieve major translational goals, such as early recognition of mental problems, accelerating access to care, and personalized monitoring of relapse. Given the high rates of mobile platform use in adolescents and young adults, the degree of the impact would be higher in the target population for early detection.

References

Alghamdi, W., Stamate, D., Vang, K., Stahl, D., Colizzi, M., Tripoli, G., Quattrone, D., Ajnakina, O., Murray, R.M., Di Forti, M., 2016. A Prediction Modelling and Pattern Detection Approach for the First-Episode Psychosis Associated to Cannabis Use, Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 825-830.

Bak, M., Delespaul, P., Krabbendam, L., Huistra, K., Walraven, W., Van Os, J., 2009. Capturing coping with symptoms in people with a diagnosis of schizophrenia: introducing the MACS-24. Int. J. Methods Psychiatr. Res. 18(1), 4-12.

Bak, M., Drukker, M., Hasmi, L., van Os, J., 2016. An n= 1 clinical network analysis of symptoms and treatment in psychosis. PLoS One 11(9), e0162811.

Cannon, T.D., Yu, C., Addington, J., Bearden, C.E., Cadenhead, K.S., Cornblatt, B.A., Heinssen, R., Jeffries, C.D., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Seidman, L.J., Tsuang, M.T., Walker, E.F., Woods, S.W., Kattan, M.W., 2016. An Individualized Risk Calculator for Research in Prodromal Psychosis. Am. J. Psychiatry 173(10), 980-988.

Collip, D., Nicolson, N., Lardinois, M., Lataster, T., Van Os, J., Myin-Germeys, I., 2011. Daily cortisol, stress reactivity and psychotic experiences in individuals at above average genetic risk for psychosis. Psychol. Med. 41(11), 2305-2315.

Delespaul, P., 1995. Assessing Schizophrenia in Daily Life: The Experience Sampling Method, Department of Psychiatry and Psychology. Maastricht University Medical Centre, Maastricht.

Edwards, C.J., Cella, M., Tarrier, N., Wykes, T., 2016. The optimisation of experience sampling protocols in people with schizophrenia. Psychiatry Res. 244, 289-293.

Fusar-Poli, P., Rutigliano, G., Stahl, D., Davies, C., Bonoldi, I., Reilly, T., McGuire, P., 2017. Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. JAMA Psychiatry 74(5), 493-500.

Hartmann, J.A., Wichers, M., Menne-Lothmann, C., Kramer, I., Viechtbauer, W., Peeters, F., Schruers, K.R., van Bemmel, A.L., Myin-Germeys, I., Delespaul, P., 2015. Experience sampling-based personalized feedback and positive affect: a randomized controlled trial in depressed patients. PLoS One 10(6), e0128095.

Hohn, P., Menne-Lothmann, C., Peeters, F., Nicolson, N.A., Jacobs, N., Derom, C., Thiery, E., van Os, J., Wichers, M., 2013. Moment-to-moment transfer of positive emotions in daily life predicts future course of depression in both general population and patient samples. PLoS One 8(9), e75655.

Iniesta, R., Stahl, D., McGuffin, P., 2016. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol. Med. 46(12), 2455-2465.

Jauhar, S., Krishnadas, R., Nour, M., Cunningham-Owens, D., Johnstone, E., Lawrie, S., 2018. Is there a symptomatic distinction between the affective psychoses and schizophrenia? A machine learning approach. Schizophr. Res.

Kimhy, D., Delespaul, P., Corcoran, C., Ahn, H., Yale, S., Malaspina, D., 2006. Computerized experience sampling method (ESMc): assessing feasibility and validity among individuals with schizophrenia. J. Psychiatr. Res. 40(3), 221-230.

Kimhy, D., Wall, M.M., Hansen, M.C., Vakhrusheva, J., Choi, C.J., Delespaul, P., Tarrier, N., Sloan, R.P., Malaspina, D., 2017. Autonomic Regulation and Auditory Hallucinations in Individuals With Schizophrenia: An Experience Sampling Study. Schizophr. Bull.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of RELIEF, European conference on machine learning. Springer, 171-182.

Koutsouleris, N., Kahn, R.S., Chekroud, A.M., Leucht, S., Falkai, P., Wobrock, T., Derks, E.M., Fleischhacker, W.W., Hasan, A., 2016. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. The Lancet Psychiatry 3(10), 935-946.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer.

Lataster, J., Collip, D., Ceccarini, J., Hernaus, D., Haas, D., Booij, L., van Os, J., Pruessner, J., Van Laere, K., Myin-Germeys, I., 2013. Familial liability to psychosis is associated with attenuated dopamine stress signaling in ventromedial prefrontal cortex. Schizophr. Bull. 40(1), 66-77.

Lataster, J., Myin-Germeys, I., Wichers, M., Delespaul, P.A., van Os, J., Bak, M., 2011. Psychotic exacerbation and emotional dampening in the daily life of patients with schizophrenia switched to aripiprazole therapy: a collection of standardized case reports. Therapeutic advances in psychopharmacology 1(5), 145-151.

Mohr, D.C., Zhang, M., Schueller, S.M., 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annu. Rev. Clin. Psychol. 13, 23-47.

Moitra, E., Gaudiano, B.A., Davis, C.H., Ben-Zeev, D., 2017. Feasibility and acceptability of post-hospitalization ecological momentary assessment in patients with psychotic-spectrum disorders. Compr. Psychiatry 74, 204-213.

Mothi, S.S., Sudarshan, M., Tandon, N., Tamminga, C., Pearlson, G., Sweeney, J., Clementz, B., Keshavan, M.S., 2018. Machine learning improved classification of psychoses using clinical and biological stratification: Update from the bipolar-schizophrenia network for intermediate phenotypes (B-SNIP). Schizophr. Res.

Myin-Germeys, I., van Os, J., Schwartz, J.E., Stone, A.A., Delespaul, P.A., 2001. Emotional reactivity to daily life stress in psychosis. Arch. Gen. Psychiatry 58(12), 1137-1144.

Nelson, B., McGorry, P.D., Wichers, M., Wigman, J.T.W., Hartmann, J.A., 2017. Moving From Static to Dynamic Models of the Onset of Mental Disorder: A Review. JAMA Psychiatry 74(5), 528-534.

Oorschot, M., Kwapil, T., Delespaul, P., Myin-Germeys, I., 2009. Momentary assessment research in psychosis. Psychol. Assess. 21(4), 498.

Pot-Kolder, R.M., Geraets, C.N., Veling, W., van Beilen, M., Staring, A.B., Gijsman, H.J., Delespaul, P.A., van der Gaag, M., 2018. Virtual-reality-based cognitive behavioural therapy versus waiting list control for paranoid ideation and social avoidance in patients with psychotic disorders: a single-blind randomised controlled trial. The Lancet Psychiatry 5(3), 217-226.

Pries, L.K., Guloksuz, S., Menne-Lothmann, C., Decoster, J., van Winkel, R., Collip, D., Delespaul, P., De Hert, M., Derom, C., Thiery, E., Jacobs, N., Wichers, M., Simons, C.J.P., Rutten, B.P.F., van Os, J., 2017. White noise speech illusion and psychosis expression: An experimental investigation of psychosis liability. PLoS One 12(8), e0183695.

Rudolph, P., 1995. Good, Ph.: Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Series in Statistics, Springer-Verlag, Berlin—Heidelberg—New York: 1994, x, 228 pp., DM 74, 00; ōS 577.20; sFr 74.–. ISBN 3-540-94097-9. Biometrical Journal 37(2), 150-150.

Schnack, H.G., 2017. Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). Schizophr. Res.

Stamate, D., Katrinecz A., Alghamdi W., Stahl, D., Delespaul, P., van Os, J., Guloksuz, S., 2017.

Predicting Psychosis Using the Experience Sampling Method with Mobile Apps, Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on. IEEE, 667-673.

Stamate, D., Alghamdi, W., Ogg, J., Hoile, R., Murtagh, F., 2018a. A Machine Learning Framework for Predicting Dementia and Mild Cognitive Impairment. Machine Learning and Applications (ICMLA), 2018 17th IEEE International Conference on. IEEE, 671- 678.

Stamate, D., Alghamdi, W., Stahl, D., Zamyatin, A., Murray, R. M., Di Forti, M., 2018b. Can Artificial Neural Networks Predict Psychiatric Conditions Associated with Cannabis Use? Artificial Intelligence Applications and Innovations (AIAI), 2018 14th IFIP International Conference on, Springer, 311-322.

Tandon, N., Tandon, R., 2018. Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia. Oxford University Press US.

Thewissen, V., Bentall, R.P., Lecomte, T., van Os, J., Myin-Germeys, I., 2008. Fluctuations in self-esteem and paranoia in the context of daily life. J. Abnorm. Psychol. 117(1), 143.

Torous, J., Keshavan, M., 2018. A new window into psychosis: The rise digital phenotyping, smartphone assessment, and mobile monitoring. Schizophr. Res.

Torous, J., Onnela, J.P., Keshavan, M., 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Transl Psychiatry 7(3), e1053.

van de Leemput, I.A., Wichers, M., Cramer, A.O.J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E.H., Viechtbauer, W., Giltay, E.J., Aggen, S.H., Derom, C., Jacobs, N., Kendler, K.S., van der Maas, H.L.J., Neale, M.C., Peeters, F., Thiery, E., Zachar, P., Scheffer, M., 2014. Critical slowing down as early warning for the onset and termination of depression, Proc. Natl. Acad. Sci. U. S. A., 87-92.

van der Steen, Y., Gimpel-Drees, J., Lataster, T., Viechtbauer, W., Simons, C., Lardinois, M., Michel, T., Janssen, B., Bechdolf, A., Wagner, M., 2017. Clinical high risk for psychosis: the association between momentary stress, affective and psychotic symptoms. Acta Psychiatr. Scand. 136(1), 63-73.

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., Drukker, M., Reininghaus, U., Jacobs, N., Lataster, T., 2017a. The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. Depress. Anxiety 34(6), 481-493.

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., Drukker, M., Reininghaus, U., Jacobs, N., Lataster, T., Simons, C., PhD, E.-M.I., Lousberg, R., Guloksuz, S., Leue, C., Groot, P.C., Viechtbauer, W., Delespaul, P., 2017b. The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. Depress Anxiety 34(6), 481-493.

Verhagen, S.J.W., Hasmi, L., Drukker, M., van Os, J., Delespaul, P.A.E.G., 2016. Use of the experience sampling method in the context of clinical trials., Evid Based Mental Health. BMJ Publishing Group Ltd, Royal College of Psychiatrists and British Psychological Society, pp. 86-89.

Wang, R., Aung, M.S., Abdullah, S., Brian, R., Campbell, A.T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E.A., 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 886-897.

Wigman, J.T., van Os, J., Borsboom, D., Wardenaar, K.J., Epskamp, S., Klippel, A., Viechtbauer, W., Myin-Germeys, I., Wichers, M., 2015. 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., 2015/03/26 ed, 1-13.

Winterburn, J.L., Voineskos, A.N., Devenyi, G.A., Plitman, E., de la Fuente-Sandoval, C., Bhagwat, N., Graff-Guerrero, A., Knight, J., Chakravarty, M.M., 2017. Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. Schizophr. Res.

Acknowledgements

We would like to thank all researchers that substantially contributed to the conception and design of the original experience sampling studies and/or contributed to preparing the data from these multiple experience sampling studies for merging them into a combined ESM data-set. In alphabetical order: M. Bak, D. Collip, N. Nicolson, I. Myin-Germeys, N. Geschwind, C. Henquet, N. Jacobs, M. Janssens, M. Lardinois, J. Lataster, T. Lataster, C. Menne-Lothmann, M. Van Nierop, M. Oorschot, C.J.P. Simons, V. Thewissen, and M. Wichers. We also thank all research personnel that provided general administrative support in preparing the ESM data-set, in particular T. Driesen.

[FOR ALL FIGURES: PLEASE USE SEPARATELY SUBMITTED HIGH RESOLUTION FIGURES FOR PUBLICATION]

Figure 1. Illustration of the machine learning process applied to one of the aggregated, PCA or mixed PCA datasets. Input Dataset represents one of the eight aggregated datasets or one of the eight datasets that were created from the aggregated datasets using Principal Component Analysis (PCA datastes) or one of the eight datasets that include variables from both the aggregated datasets and PCA datasets (mixed PCA datasets). Support Vector Machines (SVM), Gaussian Processes (GP), Neural Networks (NN), Logistic Regression (LR) and Random Forests (RF) predictive models are trained and tuned / optimized in an inner 10-fold cross-validation. Prior to model training, the important predictors are selected with feature selection techniques such as ReliefF or Recursive Feature Elimination (RFE) on the training folds. The outer 5-fold cross-validation embeds the model post-processing with ROC (Receiver Operating Characteristic) curves and evaluates model performances. The double cross-validation called nested cross validation, repeated 100 times in a Monte Carlo simulation, is used also to reliably validate the predictive models.

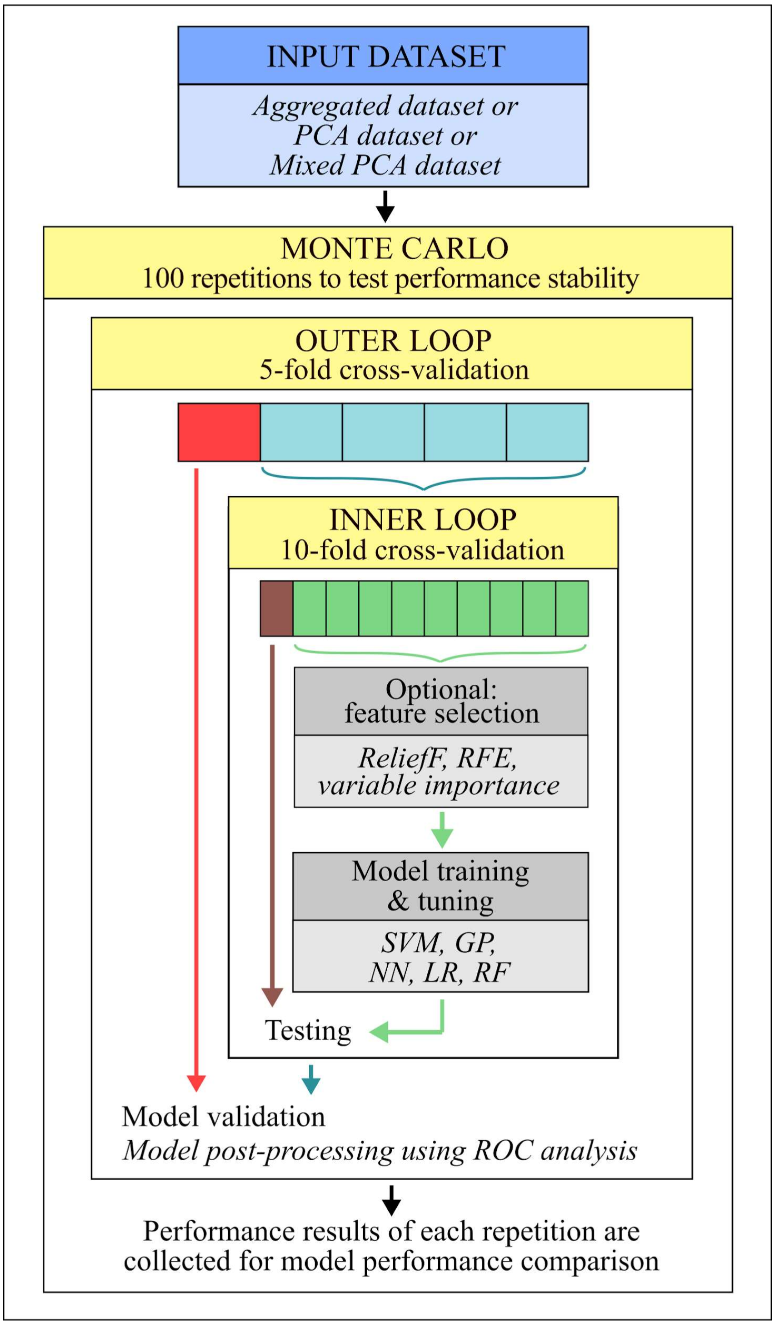

INPUT DATASET
Aggregated dataset or
PCA dataset or
Mixed PCA dataset
MONTE CARLO
100 repetitions to test performance stability
OUTER LOOP
5-fold cross-validation
INNER LOOP
10-fold cross-validation
Optional:
feature selection
ReliefF, RFE,
variable importance
Model training
& tuning
SVM, GP,
NN, LR, RF
Testing
Model validation
Model post-processing using ROC analysis
Performance results of each repetition are
collected for model performance comparison

Figure 2. Boxplots showing performance result of the top three models in Monte Carlo 100 experiments: SVM Radial (Support Vector Machines with Radial kernel), SVM Poly (Support Vector Machines with Polynomial kernel) and GP Radial (Gaussian Process with Radial kernel).

Figure 2a. Accuracy

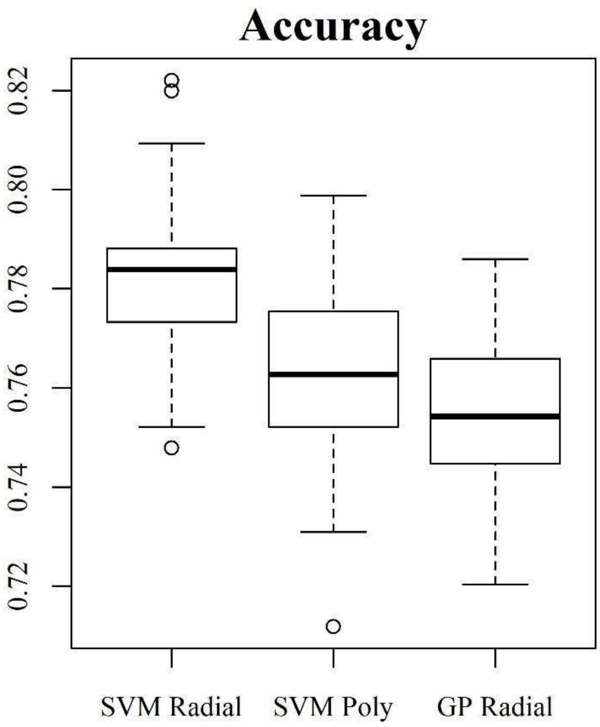


Figure 2b. Sensitivity

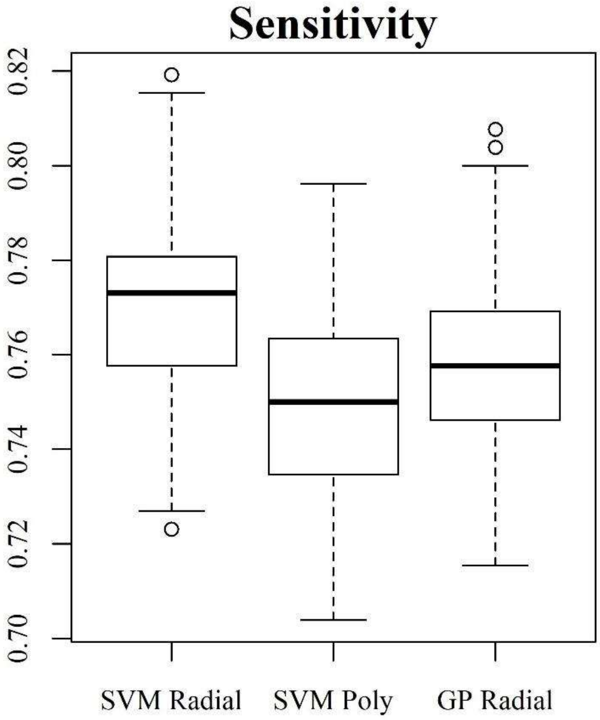


Figure 2c. Specificity

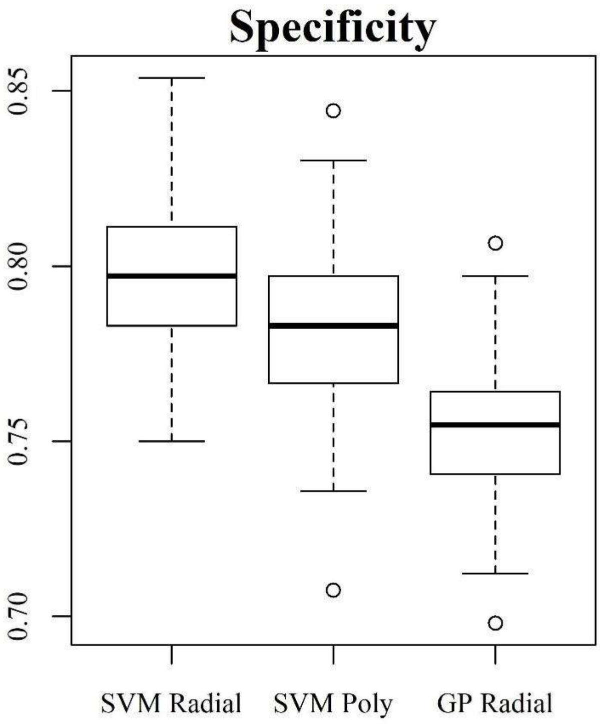


Figure 2d. Kappa

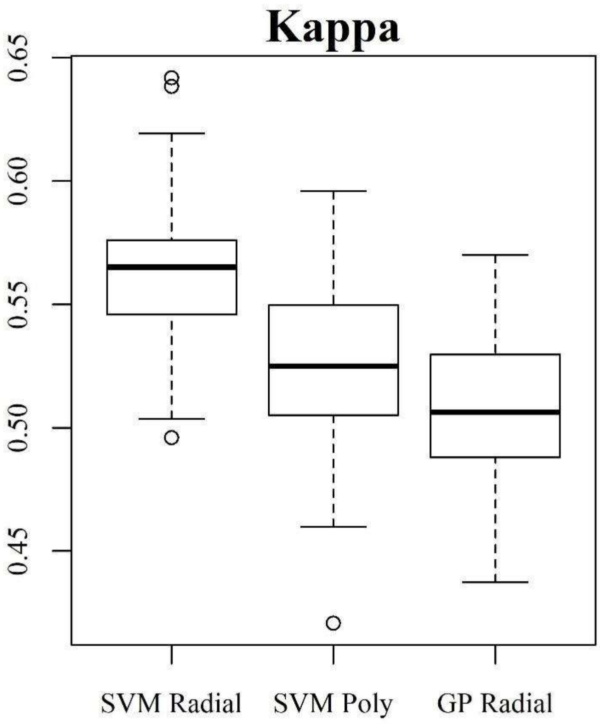
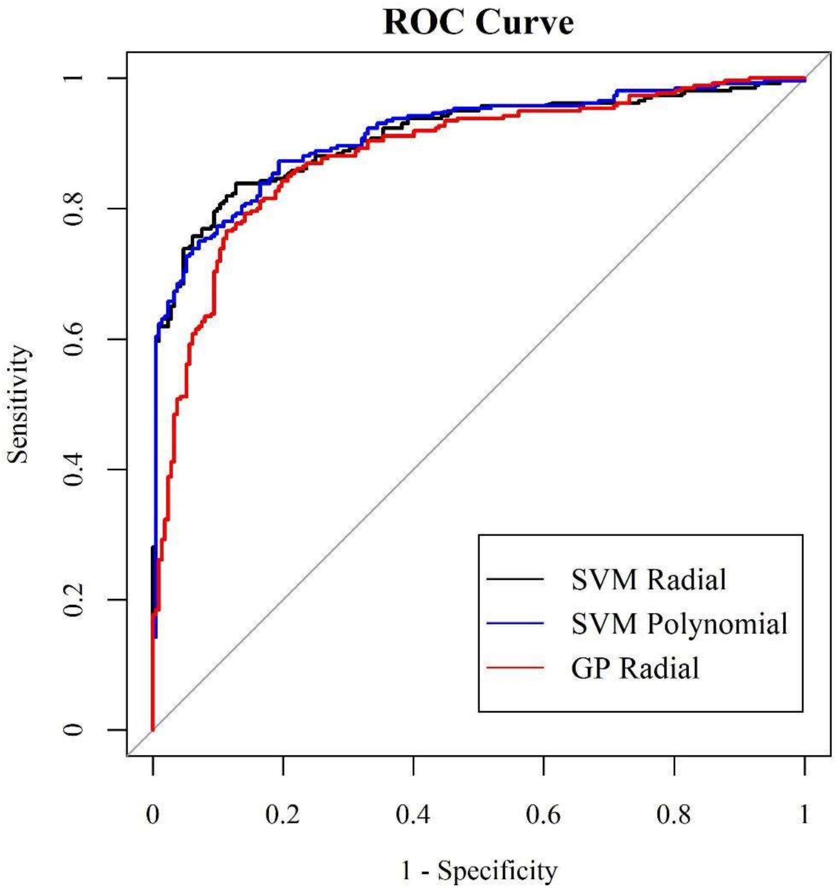


Figure 3. ROC curves of the three top models



Model	AUC	Sens	Spec	Acc	Kappa	AUC*	Sens*	Spec*	Acc*
SVM Radial	0.8639	0.8192	0.8255	0.8220	0.6419	0.8459	0.7706	0.7957	0.7819
SVM Poly	0.8435	0.7885	0.8113	0.7987	0.5959	0.8300	0.7481	0.7828	0.7637
GP Radial	0.8216	0.7808	0.7925	0.7860	0.5700	0.8157	0.7582	0.7535	0.7561

Table 1. Performances for the three top models

All the best three predictive models' results were achieved by the datasets including base, velocity and acceleration data, and with the V2 aggregation rule applied.

*Monte Carlo (n=100) experiments average results. AUC, Sens, Spec, Acc stand for Area Under Curve, Sensitivity, Specificity and Accuracy, respectively.

Table 2. Variable rank with Learning Vector Quantization (LVQ), Recursive Feature Elimination (RFE) and ReliefF feature selection methods applied on the dataset including base, velocity and acceleration data in normal values, with V2 aggregation applied

Rank	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			