Journal of the Neurological Sciences 347 (2014) 262–267

Contents lists available at ScienceDirect

Journal of the Neurological Sciences

journal homepage: www.elsevier.com/locate/jns

Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms

Jorne Laton [a,*], Jeroen Van Schependom [a,b], Jeroen Gielen [a], Jeroen Decoster [c], Tim Moons [c], Jacques De Keyser [a], Marc De Hert [c], Guy Nagels [a,b,c,d]

[a] Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Laarbeeklaan 101, 1090 Brussel, Belgium
[b] Faculté de Psychologie et des Sciences de l'Education, Université de Mons, Place du Parc 20, 7000 Mons, Belgium
[c] UPC KU Leuven - Campus Kortenberg, Department of Neurosciences, KU Leuven, Leuvensesteenweg 517, 3070 Kortenberg, Belgium
[d] National MS Center Melsbroek, Vanheylenstraat 16, 1820 Melsbroek, Belgium

ARTICLE INFO

Article history:
Received 27 June 2014
Received in revised form 3 October 2014
Accepted 8 October 2014
Available online 16 October 2014

Keywords:
Schizophrenia
EEG
P300
MMN
Pattern classification

ABSTRACT

Objective: The diagnostic process for schizophrenia is mainly clinical and has to be performed by an experienced psychiatrist, relying primarily on clinical signs and symptoms. Current neurophysiological measurements can distinguish groups of healthy controls and groups of schizophrenia patients. Individual classification based on neurophysiological measurements mostly shows moderate accuracy.
We wanted to examine whether it is possible to distinguish controls and patients individually with a good accuracy. To this end we used a combination of features extracted from the auditory and visual P300 paradigms and the mismatch negativity paradigm.
Methods: We selected 54 patients and 54 controls, matched for age and gender, from the data available at the UPC Kortenberg. The EEG-data were high- and low-pass filtered, epoched and averaged. Features (latencies and amplitudes of component peaks) were extracted from the averaged signals. The resulting dataset was used to train and test classification algorithms. First on separate paradigms and then on all combinations, we applied Naïve Bayes, Support Vector Machine and Decision Tree, with two of its improvements: Adaboost and Random Forest.
Results: For at least two classifiers the performance increased significantly by combining paradigms compared to single paradigms. The classification accuracy increased from at best 79.8% when trained on features from single paradigms, to 84.7% when trained on features from all three paradigms.
Conclusion: A combination of features originating from three evoked potential paradigms allowed us to accurately classify individual subjects as either control or patient. Classification accuracy was mostly above 80% for the machine learners evaluated in this study and close to 85% at best.

1. Introduction

Schizophrenia is a complex psychiatric disorder. The diagnostic process is mainly clinical and has to be performed by an experienced psychiatrist, who relies to a large extent on clinical signs and symptoms. Clinicians and patients would benefit from biomarkers that can help distinguish schizophrenia from normal controls.

Current neurophysiological measurements can distinguish between groups of healthy controls and groups of schizophrenia patients. Mismatch Negativity (MMN) amplitude is reduced in schizophrenia patients compared to healthy controls [1]. A reduction in MMN amplitude was also shown in people experiencing very early stages of a psychotic illness and those in an at-risk mental state [2]. Overall, reduction in MMN amplitude is one of the most consistent findings in schizophrenia [2].

In the evaluation of potential biomarkers, individual classification value is more important than group level effects. Individual classification based on neurophysiological measurements only shows moderate accuracy. Greenstein et al. applied the Random Forest machine learner on data from 74 anatomic brain MRI subregions obtained from 98 childhood onset schizophrenia patients and 99 age, sex and ethnicity-matched healthy controls [3]. Patients and controls were classified on a combination of brain regions with an accuracy of 73.7%. Johannesen et al. investigated classification of healthy controls and schizophrenia- and bipolar disorder patients based on different sets of features: (1) P50 suppression, P300 latency and P300 amplitude; (2) N100 amplitude; (3) evoked spectral power and (4) P50 and P300 hemisphere asymmetry [4]. They achieved 71% accuracy when classifying schizophrenia patients and healthy controls with the P50 and P300 endophenotypes.

* Corresponding author at: Center for Neurosciences (C4N), UZ Brussel, Vrije Universiteit Brussel, Laarbeeklaan 101, 1090 Brussel, Belgium. Tel.: +32 2 477 64 10.
E-mail addresses: Jorne.Laton@vub.ac.be (J. Laton), Jeroen.Van.Schependom@vub.ac.be (J. Van Schependom), Jeroen.Gielen@vub.ac.be (J. Gielen), Jeroen.Decoster@uzleuven.be (J. Decoster), Tim.Moons@opzgeel.be (T. Moons), Jacques.DeKeyser@uzbrussel.be (J. De Keyser), Marc.De.Hert@uc-kortenberg.be (M. De Hert), Guy.Nagels@vub.ac.be (G. Nagels).

http://dx.doi.org/10.1016/j.jns.2014.10.015
0022-510X/

Using the N100 and spectral power improved the accuracy to 79% for classification of subjects as either patient or control.

In this study, we wanted to examine whether it is possible to distinguish between schizophrenia patients and normal controls at the individual level with a good accuracy, using machine classifiers. To this end we used a combination of features from the mismatch negativity paradigm, the auditory- and the visual P300 paradigm.

2. Methods

2.1. Participants

Fifty-four patients with schizophrenia or schizoaffective disorder ('Schizo'; 36 male; age: 40.5 ± 10.1) and 54 healthy non-medicated control participants ('Norm'; 36 male; age: 37.6 ± 14.1) were recruited, matched for age and gender. Patients were recruited in the UPC (University Psychiatric Centre KULeuven, campus Kortenberg), where they were diagnosed by a semi-structured interview (OPCRIT v4.0). All participants have given written informed consent. Detailed demographic data can be found in Table 1.

2.2. Recording conditions

EEGs were recorded using a 64-channel ANT digital EEG measure station (ANT, The Netherlands). Ag-AgCl electrodes were arranged in an electrode cap using the international 10/10 system. Signals were digitised at a sampling frequency of 256 Hz and stored for offline analysis.

2.3. Paradigms and procedures

The P300 paradigms are attention related, requiring a response to their "target" stimulus. The MMN is similar but requires no directed attention since it is performed passively. The grand averages for both groups and for the different stimuli are shown in Fig. 1.

2.3.1. Auditory P300 (P300a)

The paradigm consists of 't' (target), 'd' (distractor) and 's' (standard) tones. The target and the distractor differ from the standard tone (1000 Hz) respectively by a higher (1500 Hz) and a lower frequency (500 Hz) and all have a duration of 100 ms and a loudness of 70 dB. These stimuli are presented pseudo-randomly, with a distribution of 80% standard, 10% target and 10% distractor and an inter-stimulus interval randomised between 1 and 1.5 seconds. In total 400 tones are provided per test, with a total test time of 540 seconds.

2.3.2. Visual P300 (P300v)

For the visual P300 a target (Square, side 106 pixels), distractor (Circle, diameter 176 pixels) and standard figure (Square, side 158 pixels) are presented in full blue (RGB = 0, 0, 255) in the centre of a black (RGB = 0, 0, 0) background with a resolution of 1024 by 768 pixels. The distribution of the stimuli and the total test time are the same as for the auditory P300.

Table 1
Demographic data.

	Patients	Controls	P
Amount of participants	54	54	
Male	36	36	
Age (years): mean ± std	40.5 ± 10.1	37.6 ± 14.1	0.22
Age (years): range	[22.4, 60.5]	[15.1, 64.4]	
Education (years): mean ± std	12.6 ± 1.80	14.8 ± 2.11	4.84×10^{-5}
Disease duration (years): mean ± std	14.8 ± 9.04	–	
Disease duration (years): range	[1, 40]	–	

2.3.3. MMN

The paradigm consists of 'd' (duration deviant), 'f' (frequency deviant) and 's' (standard) tones. The duration and frequency deviant tones differ from the standard tone (100 ms, 1000 Hz) respectively by a longer duration (250 ms) and a higher frequency (1500 Hz) and all have a loudness of 70 dB. These stimuli are presented in pseudo-random order, with a distribution of 90% standard, 5% duration deviant and 5% frequency deviant and an inter-stimulus interval of 300 ms. In total 1800 tones are provided per test, with a total test time of 733 seconds.

2.4. EEG signal pre-processing

Offline signal processing was done in Matlab [5] using SPM8 [6]. First, three Butterworth filters were applied: a high-pass filter with a cut-off frequency at 0.1 Hz, removing DC, a low-pass filter with a cut-off frequency at 30 Hz followed by a band-stop filter with range of 48 Hz to 52 Hz, removing 50 Hz mains hum.

Then, signals were epoched with a [-200, 800] ms peristimulus interval for the P300 paradigms, and a [-100, 500] ms peristimulus interval for the MMN, after which baseline correction and rereferencing to linked ears were performed.

This was followed by artefact rejection, using three criteria: absolute maximum (>80 μV), peak-to-peak maximum (>120 μV) and flat segment rejection. Then, the epochs were averaged into stimulus specific responses for each individual patient. Subsequently, to remove high frequencies that might have arisen by this averaging, the abovementioned low-pass filter was reapplied and baseline correction was performed.

2.5. Feature detection

We designed an algorithm in Matlab for automatic detection of the latencies and amplitudes of a set of peaks (see Table 2) in channels Fz, Cz and Pz. The parameters that are entered into the algorithm, are the averaged evoked potentials, a time interval in which the peak is to be detected, and the direction of the peak (positive or negative). A peak is defined as a time point that has a higher amplitude than its neighbouring time points, so multiple peaks can be detected in an interval. The latency and amplitude of the peak with the largest absolute value is returned. The possibility exists that no peak is detected at all, e.g. when the interval only contains a rising edge. To handle missing values that might occur when no local maxima are detected, we added a 'force' option to the algorithm. In theory, the derivative of the signal is zero when the signal reaches a local maximum. This option 'force' therefore returns the time point with the smallest derivative in the specified interval.

The choice of the minimum and maximum value for each intervals in which the algorithm should search for the peak, was data-driven. We always started with a [-50, +50] milliseconds interval around the average latency of the peak, measured on the grand averages of the dataset (see Fig. 1). When the deviation of the latency of a peak was larger than 50 milliseconds for at least one subject, which was the case for the P300 peaks in both the auditory and the visual P300 paradigm, this interval was extended as much as necessary to contain correct data from all subjects. The limit to extending the detection interval of a peak was the presence of a larger peak nearby, e.g. N100 and N200, or P200 and P300. Optimisation consisted of extending intervals where necessary whilst ensuring as little overlap as possible between intervals of peaks that point in the same direction. The peaks and their detection intervals are shown in Table 2.

The complete feature set consisted of amplitude and latency of each of these peaks in channels Fz, Cz and Pz, which yielded six features per peak. With four peaks in the target- and in the distract event of the auditory and visual P300 paradigm and two peaks in the frequency- and in the duration deviant event of the mismatch negativity paradigm, 120 features were measured in total for each subject. For both P300

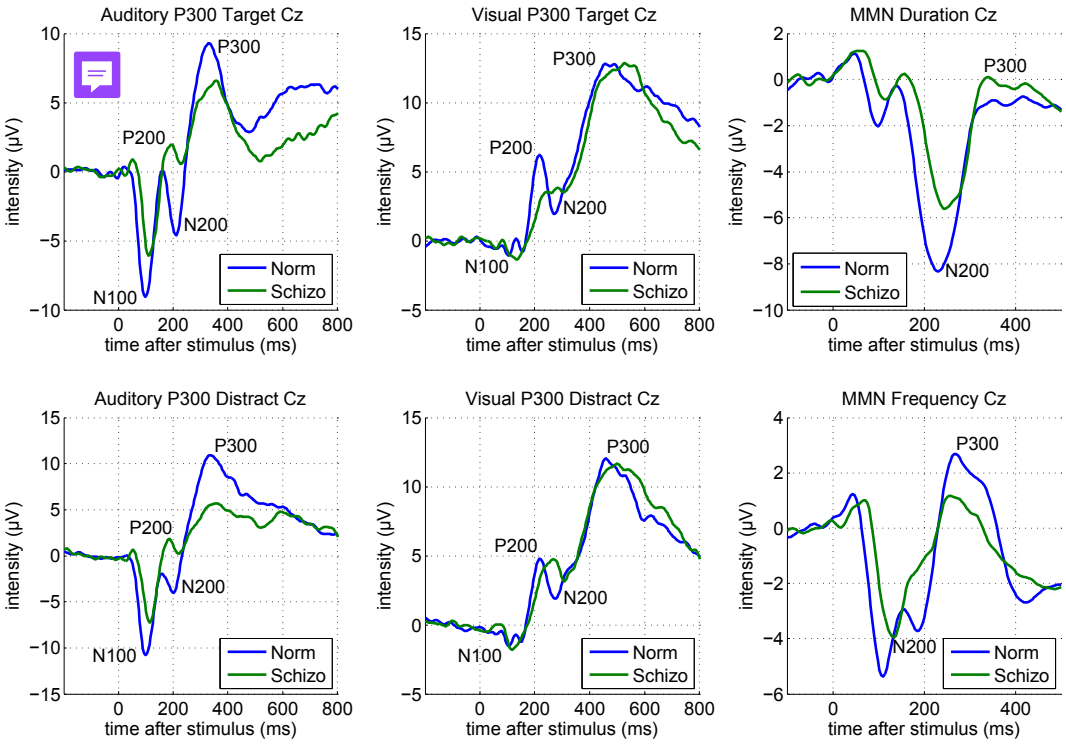

Fig. 1. Left column: grand averages of patient group (green) and control group (blue) of target and distractor event of the auditory P300 paradigm. Middle column: grand averages of target and distractor event of the visual P300 paradigm. Right column: duration and frequency deviant events of the mismatch negativity paradigm.

paradigms, detection of features was performed directly in the averaged target and distractor responses, while for the MMN the frequent (standard) event average was subtracted from the deviant event average before detection.

2.6. Feature pre-processing

This was the last analysis phase performed in Matlab. The feature values underwent two further steps before classification started: standardisation and normalisation. In the standardisation step, for each instance the feature mean was subtracted from the individual subject's feature value and this result was divided by the standard deviation of that feature. This is the z-score [7]. Transforming features into their z-scores ensures that all features have equal weight in training. If features do not have equal weights, classifiers can get biased [7]. In the normalisation step, the sigmoid function was applied to these z-scores, rescaling them all to values between 0 and 1. The sigmoid function is approximately linear between -2 and 2, saturates to 0 on the negative side and to 1 on the positive side. This property reduces the influence of z-scores larger than 2 in absolute value, which lie outside the 95% confidence interval. We did not delete any features, nor any subjects in this feature processing phase, features were only 'flattened' by the sigmoid function when they were likely to represent outliers.

2.7. Classification

Using the R development environment [8], we then applied three well-known classifiers: Naïve Bayes, Support Vector Machine and Decision Tree. We also applied two improvements on Decision Tree: Adaptively Boosted Decision Tree and Random Forest.

Three parameters were calculated to evaluate classifier performance: sensitivity, specificity and percentage correctly classified (PCC).

To estimate the sensitivity, specificity and accuracy of a classifier, the average and standard deviation of 100 randomly generated tenfold cross-validation estimates were calculated. Patient and control data were collected into one dataset, which was then randomly divided into ten subsets or folds with the same proportion of patients and controls as in the complete dataset. In each iteration of a tenfold cross-validation, nine folds were used for training the classifier, while the remaining fold was used to test the trained classifier. This step was repeated until each fold had been used once for testing. The test results for each fold were then combined into a total estimate of the classifier's performance. The repeated tenfold cross-validation also allowed for the calculation of the standard deviation of the classifier's performance indices.

Table 2
Detection intervals in channels Fz, Cz and Pz for every peak, intervals were initialised at [-50, 50] ms around its average latency. These intervals were extended where necessary and possible, especially for the P300 peaks, as their latencies showed larger variance than the others.

Paradigm	Event	Peak	Detection Interval (ms after stimulus)
Auditory P300	target	N100	[50, 160]
		P200	[100, 250]
		N200	[150, 310]
		P300	[260, 450]
	distractor	N100	[50, 150]
		P200	[100, 260]
		N200	[140, 300]
		P300	[260, 450]
Visual P300	target	N100	[50, 200]
		P200	[150, 300]
		N200	[200, 400]
		P300	[350, 600]
	distractor	N100	[50, 200]
		P200	[150, 300]
		N200	[200, 400]
		P300	[400, 600]
Mismatch Negativity	duration deviant, relatively to standard	N200	[150, 300]
		P300	[250, 450]
	frequency deviant, relatively to standard	N200	[100, 250]
		P300	[200, 400]

2.7.1. Naïve Bayes

Derived from Bayes' theorem, this is one of the best known and at the same time one of the simplest classifiers [7,9]. Despite its simplicity and naïvety it has proven to perform surprisingly well on arbitrarily large feature sets. The simplicity follows from the naïvety, which is introduced by assuming the random variables (features) are independent from each other. This is of course seldom the case in real-life datasets, but even then it has shown to be able to achieve high classification accuracy. We implemented this classifier using the R package "e1071" [10].

2.7.2. Decision tree

Decision tree classifiers are well-known machine learners that generate a tree structure in which each node represents a split of the dataset based on a feature's value [7,9]. Using all feature values of an instance, this tree is traversed and the class at the end of the path is the final classification of that instance. The generation of such trees is taken care of by an algorithm based on information gain calculation for individual features. When the tree is fully built, it needs pruning, which is basically cutting branches in the tree that add too little extra accuracy. Cutting these branches reduces the risk of overfitting and use of random features.

The decision tree is a very easy to understand and an intuitive model, which makes it highly interpretable and easy to implement. A large drawback, however, is the fact that the algorithm generating it may have trouble handling a too large number of features, especially if many of these do not contain information relevant to the classification problem, because in these conditions the algorithm tends to overfit. Inevitably, useless features might accidentally seem to predict a certain subset and can therefore wrongly be incorporated in the model. Pruning reduces this problem, but needs well-considered tuning of its parameters. It is also impossible to manually select the most useful features, as that should be the aim of the classifier itself. We implemented this classifier using the R package "rpart" [11,12] with default settings.

2.7.3. Adaptive boosting

AdaBoost is a meta-classifier designed for boosting the performance of existing (weak) classifiers, for which an error rate of slightly less than 0.5 (in the binary case) is already enough to apply iterative boosting [7,13,14]. Per iteration the weight of incorrectly classified instances is increased relatively to the correctly classified instances. Then the classifier is retrained with these new instance weights. This forces the classifier to focus more on the misclassified subjects. The only parameter to be tuned is the number of iterations, for which higher settings give better performance up to a certain point, at the cost of longer computation time. We used the R package "ada" [13] to apply this classifier and set the number of iterations to 100.

2.7.4. Random forest

A Random Forest is a collection of randomly generated decision trees. For each of these individual trees, a subset of features is selected at random from the complete feature set and then the tree is built. This resolves a major problem of decision trees: training a decision tree is difficult as there is a large number of features. After a fixed number of trees has been built, the prediction is based on majority voting of all trees. Most of these trees will be useless and giving random responses, but on average these will cancel each other out, which results in only the relevant trees adding to the prediction. We used the R package "randomForest" [15,16] with default parameters for this classifier. We also used the R package "FSelector" [17] for the built-in Random Forest Importance function, to generate a feature importance ranking, based on weights.

2.7.5. Support vector machine

The Support Vector Machine is a machine learning algorithm, used to analyse data and recognise patterns. Given a training set, a model is built that assigns the instances to one of the two categories. This can be represented as points in space, the two categories are hereby divided by a gap that is as wide as possible. New instances are mapped into that space, the side of the gap on which they fall determines to which category they belong. We used the R package "e1071" [10] and tuned its parameters with the built-in tune function.

3. Results

3.1. Feature ranking

To get a global idea of the relevance of our features, we calculated two measures: the significance and the importance; the latter is used for selection in Random Forest. These rankings are just informative, we did not use these to perform a separate step of feature selection in classification, because seemingly irrelevant features may very well become useful when they are combined.

Names for the features in the tables are constructed by concatenating the following parts:

- paradigm: auditory P300 = "P300a", visual P300 = "P300v", mismatch negativity = "MMN"
- stimulus type:

  P300 paradigms: target = "t", distract = "d"

  MMN paradigms: frequency deviant = "f", duration deviant = "d"
  - peak: the "00" are omitted from the peak labels, so we get "N1" for the N100 peak and so on
  - channel: "Fz", "Cz", "Pz"
  - latency = "L", amplitude = "A"

3.2. Classification accuracy

We measured the performance of our classifiers using features originating from only one paradigm in Table 5; from all possible combinations of two paradigms in Table 6 and from the combination of all three paradigms in Table 7.

4. Discussion

The feature rankings (see Tables 3 and 4) show that the features extracted from the auditory P300 paradigm are the most important ones. If all paradigms were equally valuable, we could expect the same distribution in a top 20 ranking as in the whole feature set, being 40% features from the auditory P300 paradigm, 40% from the visual P300 paradigm and 20% from the MMN paradigm. In both rankings shown in Tables 3

Table 3
Ranking of 20 most significantly different features.

Rank	Feature	Significance
1	P300at N1CzL	2,41E-07
2	P300vt P2CzL	4,30E-06
3	P300vd P2CzL	5,47E-06
4	P300ad N1PzA	6,51E-06
5	P300vt P2PzL	7,36E-06
6	P300ad N1PzL	9,36E-06
7	P300ad N1CzA	3,59E-05
8	P300at N1PzL	4,42E-05
9	P300at N2CzL	5,63E-05
10	P300ad N1CzL	6,96E-05
11	P300ad P3PzA	8,38E-05
12	P300ad N1FzL	1,13E-04
13	P300at P2CzL	1,56E-04
14	P300at N1CzA	1,79E-04
15	MMNf P3CzA	1,95E-04
16	P300at P2FzL	2,17E-04
17	MMNd N2CzL	2,97E-04
18	P300ad P2CzA	3,68E-04
19	P300at N2CzA	3,79E-04
20	MMNd N2FzL	4,82E-04

Table 4
Ranking of 20 most important features according to Random Forest Importance.

Rank	Feature	Importance
1	P300ad N1PzA	14,46
2	P300vd P2CzL	13,06
3	P300ad N1CzA	11,93
4	P300ad N1FzL	10,55
5	P300at N1PzL	8,06
6	P300ad N1PzL	7,75
7	P300vt P2PzL	7,29
8	P300ad N1CzL	7,08
9	P300at N1CzL	6,92
10	P300ad P2CzL	6,84
11	P300vt P2CzL	6,28
12	P300ad P2FzL	5,39
13	P300at N1CzA	5,37
14	P300at P2CzL	4,96
15	P300at N2CzL	4,84
16	P300ad P2PzL	3,82
17	MMNf N2FzA	3,73
18	MMNf P3CzA	3,65
19	MMNd N2FzL	3,62
20	P300at P3PzA	3,56

and 4, we see that 14 features originate from the auditory P300 paradigm, while we would expect 8. This strongly suggests that the auditory P300 contains the most valuable information. Furthermore, the visual P300 paradigm has only three features in this ranking, so we can state that the auditory P300 paradigm is the more valuable of the two. The mismatch negativity also appears three times, which is almost as much as expected. From these rankings we can conclude that the auditory P300 paradigm has the highest value, the MMN paradigm has medium value and the visual P300 has relatively low value. However, we did not exclude features, assign weights or perform any kind of feature selection for classification, since we assumed that a combination of features from different paradigms would be more valuable.

Classification was performed on features from one, two and all three paradigms. When features from only a single paradigm were used, the best classification accuracies were found for the auditory P300 paradigm, followed by the visual P300 paradigm. Mismatch negativity seemed to perform the worst, which may be due to the lower number of features compared to the P300 paradigms. Adding features from a second paradigm showed an improvement for the combinations "auditory P300 & visual P300" and "visual P300 & MMN", but not for "MMN & auditory P300". MMN and auditory P300 are both auditory. Perhaps the combination of information from both stimulus modalities drives the improvement in the two other combination sets. Combining the features of all three paradigms further improved the performance for Adaboost and particularly for Random Forest, which outperformed all other classifiers. Interesting to note is that the training time for Random Forest was ten times lower than for Adaboost: less than 4 seconds for one cross-validation run compared to 39 seconds. Naïve Bayes and Support Vector Machine stabilised at the best result found for the combinations of two paradigms, without further improvement when we looked at the combination of features from all three paradigms. The

Table 6
Averages and standard deviations of accuracies estimated by 100 random cross-validation runs of each classifier using features extracted from two paradigms.

Paradigms:	Auditory P300 & Visual P300		Visual P300 & Mismatch Negativity		Mismatch Negativity & Auditory P300	
Classifier	Average	SD	Average	SD	Average	SD
Naïve Bayes	0.804	0.013	0.795	0.015	0.784	0.009
Decision Tree	0.782	0.033	0.654	0.036	0.735	0.034
Adaboost	0.816	0.020	0.787	0.023	0.798	0.021
Random Forest	0.826	0.022	0.799	0.019	0.806	0.021
Support Vector Machine	0.800	0.015	0.813	0.017	0.794	0.014

Decision Tree performance dropped slightly when we moved from combinations of two paradigms to a combination of all three paradigms. This might be due to this classifier's problems in dealing with large sets of features.

In summary, we show that a combination of features from oddball and mismatch evoked potential paradigms can be used to distinguish between patients with schizophrenia and healthy controls with an accuracy close to 85%.

This is an improvement on previous results from similar studies on schizophrenia. Neuhaus et al. also worked on attention-related evoked potentials and tested nine classifiers with cross-validation on data obtained from a group of 40 schizophrenia patients and 40 healthy controls [18]. Their feature set consisted of amplitudes and latencies of the N1 peak in channels P3, Pz, P4, O1 and O2 and of the P3 peak in channels Fz, Cz, P3, Pz and P4, from which we can conclude they had a total of 20 features. They also used Naïve Bayes with equal variance, yielding 77.2% total accuracy, and Naïve Bayes with unequal variance, yielding 78.3%, but their best result of 78.9% was achieved by the Mahalanobis classifier, which is based on the Mahalanobis distance definition.

In contrast to this study, we also used a correlate of spontaneous attention, the MMN, which might explain our higher classification accuracy. In general, it is a good idea not to include too many features, but the set of 120 features we used to train the classifiers proved not to be too high for adequate classification training.

More recently, Neuhaus et al. performed a very similar study on a group of 144 schizophrenia patients and 144 matched controls using a combination of the click-conditioning and the oddball paradigm [19]. Features were also latencies and amplitudes, but extracted from all 32 channels. This is one of the main differences with respect to our study: they had 64 features per peak, where we only had six, as we only worked with three channels. To deal with their large amount of features, Neuhaus et al. performed feature selection in which features were ranked according to their strength and the most promising were selected to build a model. Their highest classification accuracy was 77.7%, obtained with Naïve Bayes.

Feature selection was not performed in our study, since we had a relatively small amount of features. We also wanted to avoid feature selection based on rankings such as those shown in Tables 3 and 4, because these rankings can lead to wrong conclusions. Features that, when

Table 5
Averages and standard deviations of accuracies estimated by 100 random cross-validation runs of each classifier using features extracted from one paradigm.

Paradigm:	Auditory P300		Visual P300		Mismatch Negativity	
Classifier	Average	SD	Average	SD	Average	SD
Naïve Bayes	0.784	0.011	0.784	0.017	0.714	0.017
Decision Tree	0.743	0.033	0.606	0.039	0.635	0.034
Adaboost	0.795	0.018	0.764	0.025	0.711	0.026
Random Forest	0.798	0.020	0.765	0.020	0.750	0.023
Support Vector Machine	0.795	0.016	0.774	0.021	0.729	0.019

Table 7
Averages and standard deviations of accuracies estimated by 100 random cross-validation runs of each classifier using features extracted from all three paradigms.

	Auditory P300 & Visual P300 & Mismatch Negativity					
Classifier	PCC AV	PCC SD	Sens AV	Sens SD	Spec AV	Spec SD
Naïve Bayes	0.803	0.011	0.817	0.012	0.789	0.020
Decision Tree	0.774	0.033	0.778	0.050	0.769	0.045
Adaboost	0.835	0.020	0.847	0.030	0.823	0.028
Random Forest	0.847	0.021	0.844	0.024	0.850	0.031
Support Vector Machine	0.799	0.015	0.808	0.021	0.791	0.023

used separately, are not able to separate subjects into two classes, can in some cases succeed in this separation when they are combined, provided that the features are linearly independent. This is an argument against feature selection in a separate step preceding classifier training, as was done by Neuhaus et al., and this is a potential explanation for the marked difference in classification accuracy between Neuhaus et al.'s study and ours.

Using P50 suppression and P300 amplitude and latency, Johannesen et al. obtained a classification accuracy of 71% in separating schizophrenia patients from normal controls, which was further improved using the N100 amplitude and the evoked spectral power, to 79% correctly classified instances [4]. In a single-trial analysis of auditory evoked responses, in particular the P50 Paired Clicks paradigm, Iyer et al. claim to obtain 100% accuracy [20]. This result was obtained on a relatively small group of test subjects (13 schizophrenia patients, 14 + 6 controls). Smaller subject groups might increase the chance of extreme classification results. However, the methods proposed in this study remain interesting, especially the majority voting and the use of single trials instead of averages, as this deals with problems such as latency variance caused peak flattening. Also, the variance of the latency over all epochs could very well be a feature on itself.

In the functional magnetic resonance imaging field, using the Auditory Oddball task (AOD) and rest data, Du et al. were able to obtain 98% classification accuracy on a group of 28 schizophrenia patients and 28 healthy controls [21]. Using leave-one-out cross-validation, they applied the Euclidian distance classifier algorithm on individual features to obtain a maximum accuracy of 93% on single components of AOD data and 91% on rest data. Feature combination in a majority voting framework further improved the results to 98% for AOD data and 93% on rest data.

Our classification results based on EEG data are not as good as Du et al.'s results, but we feel our approach has several advantages in its different stages. The features we use, for instance P300 latency or N100 amplitude, are more easily understood by a clinician than extracted single components of more complex data. The tree classifier and random forest classifier are also more easily understood by clinicians using our findings, compared to a Euclidean distance classifier. We feel that clinicians will be more inclined to accept the use of more transparent classification algorithms over the use of more complex classifiers, which they will regard as a black box.

In the field of anatomical brain measures, Ota et al. investigated grey matter and cerebrospinal fluid volume in a group of female schizophrenia patients and healthy controls and achieved an accuracy of 72% [22], while Greenstein et al. worked with regional cortical thickness, bilateral hippocampus, thalamus and lateral ventricle volumes and the Random Forest classifier to achieve 73.7% [3]. Mandl et al. combined schizophrenia patients and healthy controls with their healthy siblings which succeeded in increasing prediction rate for increased risk to 72%, using volumetric brain and white matter fibre bundle measures and IQ as classification features [23]. With an accuracy of at least 10% less, we can state that anatomical measures are not as accurate in diagnosing or predicting schizophrenia as functional measures.

Our features were extracted from evoked potential recordings. Electroencephalography has a number of considerable advantages over fMRI, time resolution being of course the most prominent. But also its relatively high portability, flexibility in possible task designs and the cost of an EEG-measurement station compared to an MRI scanner are valuable advantages.

A further step is to use source localisation [18] as an extra step before feature extraction. In this way, the location of peak activities could become a third measure next to latency and amplitude and could then also be used for training.

5. Conclusion

A combination of features originating from three evoked potential paradigms (auditory P300, visual P300 and MMN) allowed us to accurately classify individual subjects as either "normal control" or "patient with schizophrenia". Classification accuracy was above 80% for most machine learners evaluated in this study and close to 85% for a Random Forest classifier trained with features from all three evoked potential paradigms.

Conflict of interest

There is no conflict of interest.

Acknowledgements

We gratefully acknowledge the technical and nursing assistance of Mr Ignace Dewulf in obtaining the recordings for this study. GN holds the "Biogen Idec — National MS Center Melsbroek chair for neurophysiological research in multiple sclerosis" and the "Merck-Novartis joint research chair for pattern recognition in multiple sclerosis" at the Vrije Universiteit Brussel.

References

[1] Lin Y-T, Liu C-M, Chiu M-J, Liu C-C, Chien Y-L, Hwang T-J, et al. Differentiation of schizophrenia patients from healthy subjects by mismatch negativity and neuropsychological tests. PLoS One 2012;7:e34454.
[2] Atkinson RJ, Michie PT, Schall U. Duration mismatch negativity and P3a in first-episode psychosis and individuals at ultra-high risk of psychosis. Biol Psychiatry 2012;71:98–104.
[3] Greenstein D, Malley JD, Weisinger B, Clasen L, Gogtay N. Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. Front Psychiatry 2012;3:53.
[4] Johannesen JK, O'Donnell BF, Shekhar A, McGrew JH, Hetrick WP. Diagnostic specificity of neurophysiological endophenotypes in schizophrenia and bipolar disorder. Schizophr Bull 2013;39:1219–29.
[5] MATLAB. version 8 (R2012b) 2012.
[6] Ashburner J, Barnes G, Chen C. SPM8 Manual. Funct Imaging Lab Inst Neurol 2012.
[7] Witten I, Frank E. Data mining: practical machine learning tools and techniques; 2005.
[8] R Core Team. R: a language and environment for statistical computing; 2013.
[9] Mitchell TM. Machine learning; 1997.
[10] Dimitriadou E. Package "e1071." … Packag … 2009.
[11] Therneau T, Atkinson B, Ripley B, Ripley M. Package "rpart"; 2014.
[12] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC Press; 1984.
[13] Culp M, Johnson K, Michailides G. ada: an r package for stochastic boosting. J Stat Softw 2006;17:1–27.
[14] Freund Y, Schapire R, Abe N. A short introduction to boosting. J Jpn Soc Artif Intell 1999;14:771–80.
[15] Breiman L, Cutler A. Package " randomForest "; 2011.
[16] Breiman L. Random Forests; 2001 5–32.
[17] Romanski P, Kotthoff M. Package "FSelector"; 2013.
[18] Neuhaus AH, Popescu FC, Grozea C, Hahn E, Hahn C, Opgen-Rhein C, et al. Single-subject classification of schizophrenia by event-related potentials during selective attention. Neuroimage 2011;55:514–21.
[19] Neuhaus AH, Popescu FC, Rentzsch J, Gallinat J. Critical evaluation of auditory event-related potential deficits in schizophrenia: evidence from large-scale single-subject pattern classification. Schizophr Bull 2013;40:1062–71.
[20] Iyer D, Boutros NN, Zouridakis G. Single-trial analysis of auditory evoked potentials improves separation of normal and schizophrenia subjects. Clin Neurophysiol 2012;123:1810–20.
[21] Du W, Calhoun VD, Li H, Ma S, Eichele T, Kiehl KA, et al. High classification accuracy for schizophrenia with rest and task FMRI data. Front Hum Neurosci 2012;6:145.
[22] Ota M, Sato N, Ishikawa M, Hori H, Sasayama D, Hattori K, et al. Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry. Psychiatry Clin Neurosci 2012;66:611–7.
[23] Mandl RCW, Brouwer RM, Cahn W, Kahn RS, Hulshoff Pol HE. Family-wise automatic classification in schizophrenia. Schizophr Res 2013;149:108–11.