

# A Novel Approach to Probabilistic Biomarker-Based Classification Using Functional Near-Infrared Spectroscopy

Tim Hahn,<sup>1,2\*</sup> Andre F. Marquand,<sup>3</sup> Michael M. Plichta,<sup>4</sup>  
Ann-Christine Ehlis,<sup>5</sup> Martin W. Schecklmann,<sup>6</sup> Thomas Dresler,<sup>2</sup>  
Tomasz A. Jarczok,<sup>2</sup> Elisa Eirich,<sup>2</sup> Christine Leonhard,<sup>2</sup> Andreas Reif,<sup>2</sup>  
Klaus-Peter Lesch,<sup>2</sup> Michael J. Brammer,<sup>3</sup> Janaina Mourao-Miranda,<sup>3,7</sup>  
and Andreas J. Fallgatter<sup>5</sup>

<sup>1</sup>Department of Cognitive Psychology II, Johann Wolfgang Goethe University Frankfurt/Main, Frankfurt am Main, Germany

<sup>2</sup>Department of Psychiatry, Psychosomatics and Psychotherapy, University of Wuerzburg, Fuechsleinstr. 15, Wuerzburg, Germany

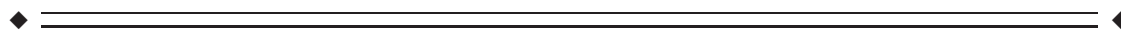
<sup>3</sup>Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, KCL, London, United Kingdom

<sup>4</sup>Department of Psychiatry, Division for Imaging in Psychiatry, Central Institute of Mental Health, Mannheim, Germany

<sup>5</sup>Department of Psychiatry and Psychotherapy, University of Tuebingen, Calwerstr. 14, Tuebingen, Germany

<sup>6</sup>Department of Psychiatry, Psychosomatics and Psychotherapy, University of Regensburg, UniversitaetsstraÙe 84, Regensburg, Germany

<sup>7</sup>Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College, London, United Kingdom



**Abstract:** Pattern recognition approaches to the analysis of neuroimaging data have brought new applications such as the classification of patients and healthy controls within reach. In our view, the reliance on expensive neuroimaging techniques which are not well tolerated by many patient groups and the inability of most current biomarker algorithms to accommodate information about prior class frequencies (such as a disorder's prevalence in the general population) are key factors limiting practical application. To overcome both limitations, we propose a probabilistic pattern recognition approach based on cheap and easy-to-use multi-channel near-infrared spectroscopy (fNIRS) measurements. We show the validity of our method by applying it to data from healthy controls ( $n = 14$ ) enabling differentiation between the conditions of a visual checkerboard task. Second, we show that high-accuracy single subject classification of patients with schizophrenia ( $n = 40$ ) and healthy controls ( $n = 40$ ) is

Additional Supporting Information may be found in the online version of this article.

Tim Hahn and Andre F. Marquand contributed equally to this work.

Contract grant sponsor: Wellcome Trust Career Development Fellowship (JMM); Contract grant number: WT086565/Z/08/Z; Contract grant sponsor: Wellcome Trust and EPSRC (AM); Contract grant number: WT088641/Z/09/Z.

\*Correspondence to: Tim Hahn, Department of Cognitive Psychology II, Johann Wolfgang Goethe University Frankfurt/Main MertonstraÙe 17, 60325 Frankfurt am Main, Germany. E-mail: TimHahn@gmx.net

Received for publication 6 June 2011; Revised 8 September 2011; Accepted 23 September 2011

DOI: 10.1002/hbm.21497

Published online 16 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

possible based on temporal patterns of fNIRS data measured during a working memory task. For classification, we integrate spatial and temporal information at each channel to estimate overall classification accuracy. This yields an overall accuracy of 76% which is comparable to the highest ever achieved in biomarker-based classification of patients with schizophrenia. In summary, the proposed algorithm in combination with fNIRS measurements enables the analysis of sub-second, multivariate temporal patterns of BOLD responses and high-accuracy predictions based on low-cost, easy-to-use fNIRS patterns. In addition, our approach can easily compensate for variable class priors, which is highly advantageous in making predictions in a wide range of clinical neuroimaging applications. *Hum Brain Mapp* 34:1102–1114, 2013. © 2012 Wiley Periodicals, Inc.

**Key words:** disease prevalence; single subject classification; schizophrenia; temporal classification

## INTRODUCTION

Recently, the interest in pattern recognition approaches to the analysis of clinical neuroimaging data has increased substantially. A crucial advantage of multivariate pattern recognition algorithms is that they provide predictions on the level of individual subjects while making use of information encoded by correlations between voxels (Marquand et al., 2010; Mourao-Miranda et al., 2005). It is this multivariate nature of pattern recognition algorithms that results in increased sensitivity over univariate methods (Ecker et al., 2009; Marquand et al., 2010) and has led to numerous applications in clinical research. Specifically, pattern recognition algorithms have shown their potential for high-accuracy classification at the level of individual subjects involving a number of disorders such as Alzheimer's disease (Tripoliti et al., 2009), autism (Ecker et al., 2009), attention deficit hyperactivity disorder (Zhu et al., 2008), schizophrenia (Kawasaki et al., 2007; for a review see Demirci et al., 2008) and major depressive disorder (Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008).

From this, the question arises why none of these promising developments have found their way into practical application. In our view, two of the most important factors limiting the widespread application of machine learning techniques in clinical practice are first, that they commonly rely on expensive neuroimaging techniques (mainly MRI) not well tolerated by many patients and – though increasingly available in many hospitals – complexity of biomarker research and application currently still requires cooperation with academic research institutions. Second, most existing applications do not provide mechanisms to accommodate variability in prior class frequencies. In other words, they are not well suited to making predictions in contexts where the frequency of each class differs from that observed in the training dataset. One reason this is important for clinical studies is to accommodate disease prevalence: for most psychiatric disorders, training a classifier based on a training set with class frequencies that accurately reflect disease prevalence is suboptimal, as it can be difficult for the classifier to adequately characterize the patient group. Instead, many neuroimaging studies

train classifiers based on artificially balanced datasets containing approximately equal numbers of patients and controls. This approach is suitable for making predictions in new datasets with similar class frequencies but it becomes problematic if the objective is to make predictions in contexts where the class frequencies are different (Bishop, 2007). For example, if a classifier is trained to diagnose a relatively rare disease using an artificially balanced dataset, the classifier will usually result in an unacceptably high type I error rate if applied to subjects drawn randomly from the general population. Overcoming both these obstacles are crucial requirements necessary for machine learning techniques to become widely used in clinical practice.

To address these issues, we propose a probabilistic spatiotemporal pattern recognition approach based on high-temporal-resolution functional near-infrared spectroscopy (fNIRS) data. Our approach enables predictions based simultaneously on the spatial pattern and temporal dynamics of the BOLD response and has highly desirable characteristics for clinical research studies: fNIRS contains information useful for basic and clinical research (Suto et al., 2004), it is substantially less expensive than alternative imaging modalities (e.g. fMRI), it is well tolerated and easily transportable (e.g. to the patient's bedside). Our approach is also fully probabilistic, which means predictions can be easily corrected to compensate for variable class priors (e.g. disease prevalence). Ultimately, this ensures that inference remains coherent even if the frequency of each class in the test dataset is substantially different from the class frequencies observed in the training dataset (Bishop, 2007).

We are not aware of any studies that apply pattern recognition methods to fNIRS data, nor of any studies that employ the temporal and spatiotemporal classification approaches proposed here. Therefore, we first provide a proof of concept application to validate our approach, where we apply the proposed temporal and spatiotemporal classifiers to fNIRS data obtained during a visual checkerboard task and then assess whether the resulting accuracies are higher over the primary visual cortex than over adjacent areas. We then apply the algorithm to the

diagnostic classification of schizophrenia using spatiotemporal patterns of brain activity during an n-back working memory task. Finally, we demonstrate how our approach can be used to compensate for variable class priors. We hypothesize that combining spatial and temporal information will allow for high-accuracy single-subject classification of schizophrenic patients and controls based on fNIRS data and that compensating for class priors will improve the predictive value of the classifier in unbalanced datasets.

resonance imaging (fMRI). Compared to fMRI, fNIRS is less restrictive, less expensive and less sensitive to motion artefacts and highly accepted by participants as well as easily transportable. Therefore it is optimally suited for claustrophobic or elderly participants (Herrmann et al., 2008; Zeller et al., 2010), psychiatric patients (Fallgatter et al., 2004; Schecklmann et al., 2008b) and children (Dresler et al., 2009). The method has proven to be highly reliable (Plichta et al., 2006a, 2007b; Schecklmann et al., 2008a); however, it comes at the cost of only being able to measure cortical structures (for an in-depth description of fNIRS, see Hoshi, 2003; Obrig and Villringer, 2003) (Fig. 1).

## MATERIALS AND METHODS

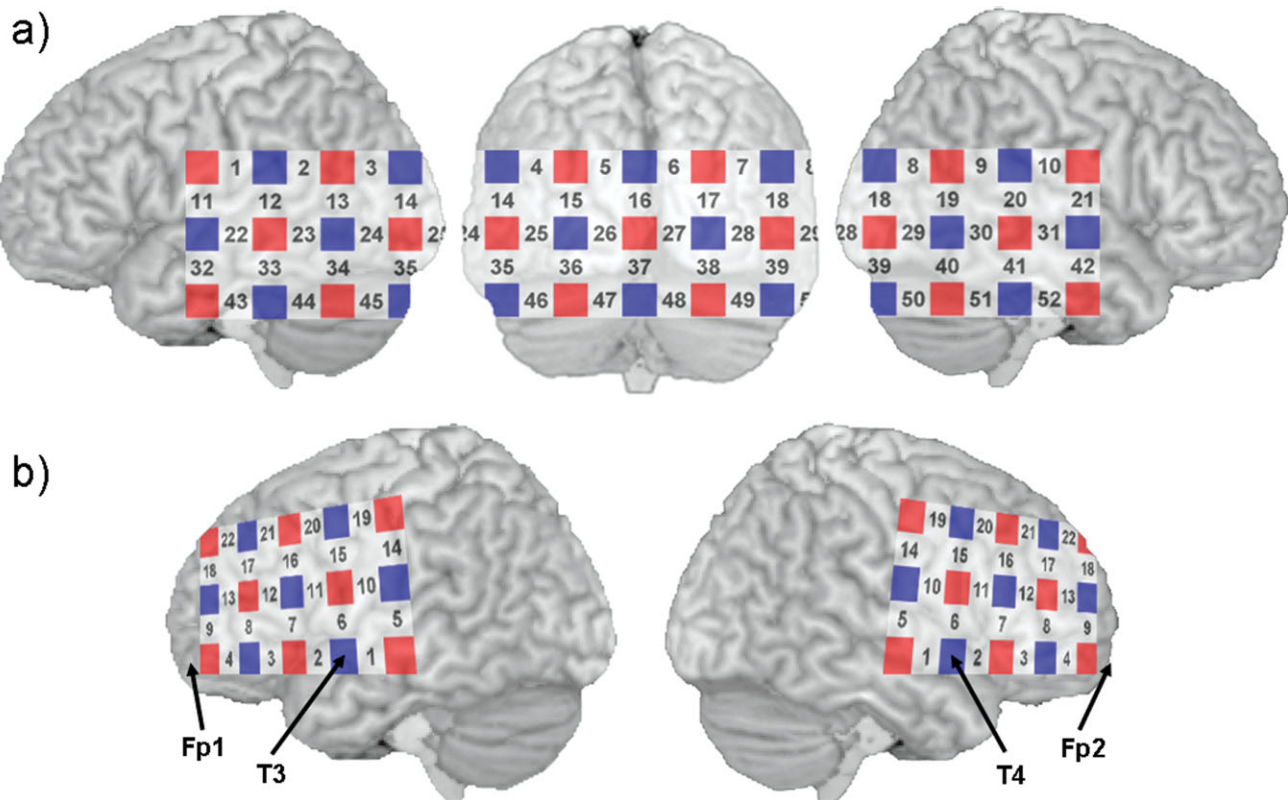
### Functional Near-Infrared Spectroscopy

Near-infrared spectroscopy (NIRS), an optical imaging technique (see Obrig and Villringer, 2003), uses near-infrared light (i.e. light with 650–850 nm wavelengths) that readily penetrates biological tissue (e.g. skin, skull, brain tissue). Basically, fNIRS measures changes similar to the blood oxygen level-dependent (BOLD) effect in functional magnetic

### Visual Checkerboard Paradigm

#### Participants

Fifteen subjects were examined (mean age =  $25.3 \pm 2.6$  years; 8 females). All subjects had normal or corrected to normal vision and no history of neurological or psychiatric disorders. One subject was excluded from analysis because of problems with localizing the visual cortex during the



**Figure 1.**

Schematic arrangement of the fNIRS probe set (red squares: emitters; blue squares: detectors; numbers: measurement channels). For the visual checkerboard paradigm, a functional localizer was used to position the probeset directly over the primary visual cortex. (a) Figure illustrates the approximate position

over the occipital cortex. (b) For the n-back task, the inferior row of the left probe set was oriented towards T3 and Fp1 (T4 and Fp2 for the right side) according to the international 10–20 system (Jasper, 1958).

functional localizer task. All subjects gave written informed consent. The investigation was in accordance with the latest version of the Declaration of Helsinki and was approved by the Ethics Committee of the University of Würzburg. Data from this sample have been published previously (Plichta et al., 2007a).

### **Task description and procedures**

A simple checkerboard paradigm was chosen as it has previously shown strong and robust activation effects in the primary visual cortex accessible with fNIRS (Hahn et al., 2010; Plichta et al., 2007a, 2006b). A parametric design with four levels of luminance intensities (Michelsen contrasts: 0, 8, 40 and 97%) was conducted, based on findings from fMRI studies (Avidan et al., 2002; Boynton et al., 1999; Heeger and Ress, 2002) which have demonstrated that an increase of stimulus contrast levels results in a monotonic increase of the primary visual cortex (V1) activity. In an event-related paradigm, a series of checkerboards was presented, each for 2 s, reversing in contrast at 8 Hz (according to Ozus et al., 2001) followed by an inter-stimulus interval (ISI) of a uniform grey colour screen of variable duration (ISI = 4–9 s, average: 6.5 s). Further details can be found in Plichta et al. (2007a).

### **fNIRS data acquisition and extraction**

Measurements were conducted with the ETG-4000 Optical Topography System (Hitachi Medical Co., Japan) using a 52-channel array of optodes, covering the occipital cortex (interoptode distance 30 mm, sampling rate 10 Hz). Initially, the probe set was placed with its lowest row centre optode at the subjects'inion.

Pre-processing of the data went as follows: First, the raw time-series at each channel was low-pass filtered with a cut-off frequency of 0.7 Hz. Then, the time-segment starting at the beginning of each trial and ending after 18 s was extracted. Following this, the average time-course over all trials of the same condition was calculated, yielding one time-segment for each channel and condition. Finally, the baseline calculated from the first 3 s preceding the onset of the trial was subtracted from each time point (these steps are in accordance with Plichta et al., 2007a).

## **N-back Task**

### **Participants**

Forty unrelated chronic schizophrenic patients according to ICD-10 criteria (23 male; mean age =  $42 \pm 12$  years; mean duration of illness = 16 years  $\pm 11$  years) participated. All patients were recruited at the Department of Psychiatry, Psychosomatics and Psychotherapy, University of Würzburg, and treated with typical ( $n = 11$ ) and/or atypical antipsychotics ( $n = 31$ ) as well as antidepressants ( $n = 20$ ) and/or benzodiazepines ( $n = 11$ ). None of the

subjects showed significant neurological co-morbidity, mental retardation or other somatic disorders. Diagnoses were made by an extensive, semi-structured interview analogous to the AMDP interview (AMDP, 2000) conducted by an experienced psychiatrist. Furthermore, 40 controls (19 male; mean age =  $40 \pm 16$  years) without history for any axis-I diagnosis or use of psychotropic medication participated.

Both groups were comparable for age ( $T = 0.712$ ;  $df = 78$ ;  $p = 0.479$ ) and gender ( $\chi^2 = 0.802$ ;  $df = 1$ ;  $p = 0.370$ ). Subjects gave written consent after detailed explanation of the procedures. The study was approved by the Ethics Committee of the University of Würzburg, and was in accordance with the Declaration of Helsinki. Data from this sample have been published previously (Reif et al., 2011).

### **Task description and procedures**

All subjects performed a letter n-back task. For the 1-back condition [low working memory (WM) load], they were instructed to press a response button, whenever a letter presented on a screen was identical to the preceding letter. For the 2-back condition (high WM load), they had to respond whenever a letter was identical to the one two trials before. The tasks were performed alternately in a block-wise fashion and separated by 30-s resting segments during which participants were instructed to sit still. Letters were presented in pseudorandomized order with a presentation time of 300 ms and an inter-stimulus interval of 1700 ms. Both task conditions were conducted three times each, resulting in three 30-s task segments for both paradigms. A 10-s baseline period preceded the first task. For both tasks, a total of 12 target trials appeared (for further details, see Reif et al., 2011).

### **fNIRS data acquisition and extraction**

fNIRS measurements were conducted with the same machine (see section fNIRS data acquisition and extraction), now using two 22-channel arrays of optodes, covering frontal areas on the left and right side of the head ( $12 \times 6$  cm each). The arrays were localized by placing the second to last optode of the bottom row onto T3/T4 on the left and right side of the head, respectively, and orienting the array towards Fp1/Fp2 (according to the International 10/20 system for electrode placement). The fNIRS probe set therefore covered lateral parts of the fronto-temporal cortex. Both arrays consisted of eight light emitters and seven photodetectors. Slow changes in the fNIRS signal unrelated to functional stimulation were removed by a linear fitting procedure, using a pre-stimulus baseline (the mean across 10 s before the active task period) and a post-stimulus baseline (the mean across the last 10 s of the 20-s rest segment) as recommended for removing long lasting physiological effects unrelated to brain activity due to functional stimulation (Ehrlis et al., 2005). Then, the time-segment starting at the beginning of each block and ending after 30 s was extracted.

Data were averaged according to the specific task condition (2-back, 1-back) channel-wise for each participant, yielding one time segment for each channel and condition (these pre-processing steps are in accordance with the procedure outlined by Reif et al. (2011)).

### Temporal Pattern Recognition of fNIRS Time-Series Data

#### Data pre-processing and Gaussian process classification

After the paradigm-specific pre-processing described above, the resulting time-course for each condition of each paradigm was mean-corrected (i.e. the mean of the time-series was subtracted from each data point). From this pre-processed time-course, the feature vector (with dimensionality equal to the number of time-points in the respective segment) was constructed.

We employed Gaussian process classifiers (GPCs) as the primary classification approach in this study because they are fully probabilistic prediction devices and therefore provide predictions that can be easily adjusted to compensate for variable class priors (see section Accommodating Variable Class Priors). In contrast, Support Vector Machines (SVMs) do not provide probabilistic predictions although ad hoc methods have been proposed to convert SVM predictions to probability scores (e.g. Platt, 1999). However, such methods are not ideal as they are not guaranteed to accurately approximate the true predictive probability density (Tipping, 2001). A second advantage of GPCs is that they provide an elegant mechanism for automatic parameter optimization. We do not review the details here, but in practice this is achieved by finding the model parameters that are most likely to have generated the data by maximizing the Bayesian model evidence (see Rasmussen and Williams, 2006). The main advantage of this approach is that multiple model parameters can be learned efficiently from the data without needing to evaluate model performance across a range of parameter settings. In contrast SVMs do not provide any methods for automatic parameter optimization, so all free parameters must be either set heuristically or optimized by cumbersome and computationally demanding methods, such as nested cross-validation, ultimately allowing for only a very limited number of parameters to be learned from the data. This feature of GPC models underpins the spatiotemporal classification approach presented in this paper, where we use GPCs to automatically learn a linear combination of individual optodes that best predicts the class labels (see Supporting Information for details).

An in-depth treatment of GP inference has been presented elsewhere and we do not repeat the details here (see Marquand et al., 2010; Rasmussen and Williams, 2006). In brief, GPC can be considered as a Bayesian extension of logistic or probit regression, where a sigmoidal likelihood function models the probability of having

observed each class label (conventionally denoted by  $p(y|\mathbf{x})$  where  $y \in \{-1, 1\}$  is the class label and  $\mathbf{x}$  is a data vector). A latent regression function models relationships between the data points and inference proceeds by (i) placing a Gaussian prior over the latent function, (ii) finding the optimal covariance parameters for the data, (iii) computing the posterior function distribution and integrating out the latent function to produce probabilistic predictions. Class predictions must sum to one, thus for binary classification it is sufficient to learn the latent function for a single class (i.e.  $p(y = 1|\mathbf{x})$ ). The predictions for the second class can then be derived by  $p(y = -1|\mathbf{x}) = 1 - p(y = 1|\mathbf{x})$ .

We used a customized version of the Gaussian processes for machine learning (GPML) toolbox for Matlab (The Mathworks, Natick, MA) (<http://www.gaussianprocess.org/gpml>) for all GPC inference. We employed a similar approach to earlier work (Marquand et al., 2010) in that we used a probit likelihood function to model the class labels, and used the Expectation propagation algorithm (Minka, 2001; Rasmussen and Williams, 2006) to compute the posterior distribution for the latent function. For classification of schizophrenic patients and controls (Experiment 2), we applied ANOVA-based feature selection (Guyon and Elisseeff, 2003) to select the time points of maximum group differences between patients and controls in the training set using a nested (3-way) cross-validation procedure as implemented in the PROBID toolbox ([www.brain-map.co.uk/probid](http://www.brain-map.co.uk/probid)), which ensures that the features were selected using the training set only. First the  $t$ -value and degrees of freedom were calculated for each time-point in the training set. Then the  $t$ -map was converted into a  $p$ -map, and time-points with a  $p$ -value higher than the threshold (uncorrected  $p = 0.05$ ) were discarded.

As noted, an important property of GP models is that they permit the flexible specification of prior covariance functions (i.e. kernels), which we utilized to construct the temporal and spatiotemporal classification models (see Supporting Information) Briefly, for the spatial classifier, we used a simple linear covariance function where data dimensions correspond to time points and the covariance function measures correlations between subjects across the whole time course. For the spatiotemporal classifier, we used the GPC to construct an optimal linear combination of covariance functions, each constructed from an individual optode. To achieve this, we first estimated an independent linear covariance function for each optode derived from the whole time course (which models the temporal pattern) then learned a weighted combination of these to train the spatiotemporal classifier (which models the spatial pattern). Note that all model parameters were learned using the training set only.

#### Leave-one-out cross-validation

The leave-one-out cross-validation (LOO-CV) done in order to assess generalizability and to predict a sample's

probability of being in class 1 (e.g. to be a patient; which we denote by  $p_{GP} = p(y = 1|\mathbf{x})$ ) was conducted as follows: In each leave-one-out run, we used data from all but one sample per class ( $S-1$  of the  $S$  samples per class) to train the classifier. Subsequently, the  $p_{GP}$  of the remaining pair of samples (e.g. one time-segment from a patient and one from a control in one condition or the time-segment from the 97% contrast condition and one from the 8% condition from the same subject), which was so far unseen by the algorithm, was calculated. This procedure was repeated  $S$  times, each time leaving out a different pair of samples, yielding each sample's  $p_{GP}$  for each contrast. Thus, for each contrast, we obtained the probability that each sample was labelled as a member of class 1 (e.g. a patient or a member of the 97% contrast checkerboard condition). This procedure was repeated for the data obtained at each one of the channels. Then, each sample's  $p_{GP}$  was thresholded at 0.50 to obtain binary class predictions. Accuracy was calculated as the ratio of correct predictions over number of cases.

To establish whether the observed GP classification accuracies are statistically significant, we ran each GP classifier 1000 times with randomly permuted labels and counted the number of permutations which achieved equal or greater accuracy than the one observed with the true labels. The p-value was then calculated by dividing this number by 1000.

Finally, spatial accuracy maps were obtained by mapping the accuracy calculated for each channel independently and applying false discovery rate (FDR) correction for multiple comparisons. Accuracies which did not survive FDR correction ( $P < 0.05$  corrected) were set to 0 in the resulting maps to show only channels with significant accuracy.

## Accommodating Variable Class Priors

### Background

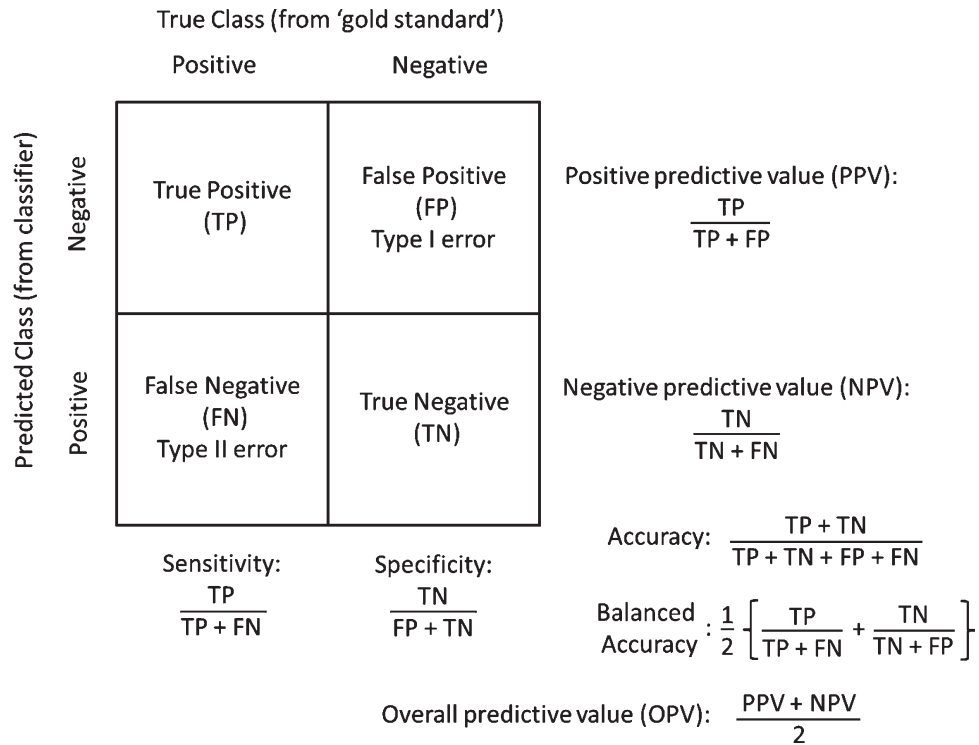
In this section, we illustrate why it is essential to accommodate variable class priors if a classifier is to be useful for predicting disease state or disease outcome in clinical practice. We adopt a didactic perspective and provide a simple synthetic example to illustrate, because accuracy measurements from diagnostic tests are surprisingly commonly misinterpreted even amongst medical professionals (see Gigerenzer and Edwards, 2003).

In practice, it can be difficult to train a classifier using a dataset with class frequencies that accurately reflect the prevalence of the disorder, particularly in neuroimaging, where data are expensive to acquire and therefore the total number of subjects available is usually small. For example, for a disease with prevalence of 5% a representative training set will contain one patient for every 20 controls. Thus it is likely that a classifier trained on such a dataset will obtain a strong bias toward detecting the larger class while not adequately characterizing the

smaller class (since it has only seen a few examples). Instead, most neuroimaging studies employ artificially balanced training sets that contain approximately equal numbers of each group with the view to applying this trained classifier to a different population. Further, neuroimaging studies most commonly report sensitivity, specificity and accuracy as the primary evaluation metrics. However, it should be emphasized that these measures alone do not completely describe the performance of the classifier. Instead, a more complete representation of classifier performance can be obtained by considering the entire confusion matrix that summarizes the errors made by the classifier on the test set (Fig. 2).

As defined in Figure 2, the sensitivity of the classifier can be stated as the probability of obtaining a positive test result (i.e. classifier prediction) given that the disease is present. Similarly, the specificity is the probability of obtaining a negative test result given that the disease is absent. These are useful to determine if the classifier is performing above chance level, but for diagnosis it is the positive predictive value (PPV) and negative predictive value (NPV) that are more important. Put simply, we wish to know what the probability of having the disease is given we have a positive test result (i.e. the PPV) or the probability of not having the disease given we have a negative result (NPV). Importantly, the PPV and NPV are both sensitive to the prior class frequencies. To see this, consider again training a classifier to diagnose a disease with prevalence of 5%. Assume that this classifier is trained on a perfectly balanced dataset as described above where it obtains a sensitivity and specificity of 80%, yielding a confusion matrix similar to that shown in Table I (where both PPV and NPV can also be seen to be 80%). If this classifier is then applied to diagnose 100 cases from the general population, approximately 95 of these can be expected to be controls and approximately five patients. From the sensitivity and specificity, we see that this classifier can be expected to diagnose  $0.8 \times 5 = 4$  of the five patients correctly and  $0.8 \times 95 = 76$  of the 95 controls correctly. So far this seems reasonable, but given 80% specificity and the class priors noted above, the classifier can be expected to mistakenly predict that 20% of healthy control subjects have the disease, which amounts to  $0.2 \times 95 = 19$  subjects, as shown in the confusion matrix in Table II. In this case, the PPV of the test is much poorer and given a positive test result, the probability of having the disorder is only  $4/(4 + 19) \approx 0.21$ .

To summarize this section, it is important to consider the PPV and NPV in addition to the sensitivity and specificity to describe the performance of a classifier. Further, the PPV and NPV from a classifier trained on an approximately balanced dataset (e.g. derived from LOO-CV) do not reflect those of a population where the classes are unbalanced. In such a situation, the PPV on the unbalanced population will be overestimated by the balanced classifier if the positive class is smaller (i.e. the type I error rate will be underestimated) and the NPV will be



**Figure 2.**

Confusion matrix summarizing the errors made by the classifier on the test set.

overestimated if the negative class is smaller (i.e. type II error rate will be overestimated).

**Adjusting probabilistic predictions to accommodate prior frequencies**

As noted, one of the main motivations for employing the GP classifiers proposed in this study is that they furnish probabilistic predictions, which unlike categorical predictions (for example, those provided by SVM) can be easily adjusted to compensate for different or variable class priors. In practice this can be achieved by a simple three-step procedure (Bishop, 2007): (i) dividing the predictive probabilities for each class by their frequency in the training set, (ii) multiplying them by their frequencies in the test set then and (iii) normalizing the class pre-

dictions so that they sum to one. Intuitively this operation has the effect of making the classifier more likely to predict the larger class by a factor determined by the relative class frequencies. The converse is also true in that the classifier is only permitted to predict the smaller class when predictive confidence is high.

**Proof of concept demonstration**

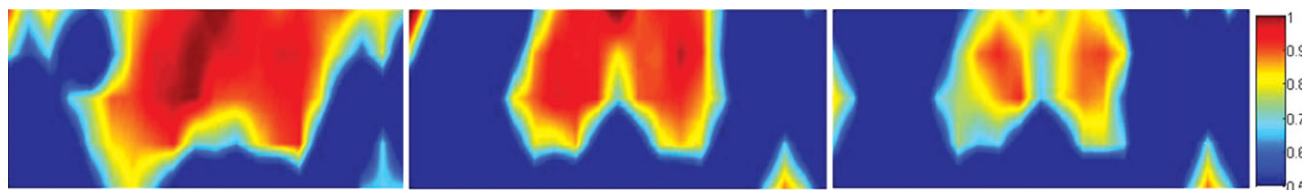
Ideally, the approach outlined above should be evaluated using new data acquired directly from the target population, which unfortunately was not available for the present sample. Instead, we demonstrate the approach by simulating an unbalanced test set by repeatedly sub-sampling the dataset described in section N-back Task (2-back condition only). This involved repeatedly performed the following steps:

**TABLE I. Performance of a hypothetical classifier (balanced dataset)**

		True class	
		Patient	Control
Predicted class	Patient	80	20
	Control	20	80

**TABLE II. Performance of a hypothetical classifier (unbalanced dataset)**

		True class	
		Patient	Control
Predicted class	Patient	4	19
	Control	1	76



**Figure 3.**

Spatial accuracy maps for classification of the 97%-contrast (left panel), the 40%-contrast (middle panel), and the 8%-contrast condition (right panel) versus the no-contrast condition ( $P < 0.05$ , FDR-corrected). Highest accuracies are consistently found over the primary visual cortex.

1. Randomly split the data in half, each half containing an equal number of patients and controls (and therefore the same total number of subjects).
2. Train a balanced classifier to discriminate patients from controls based on one half of the data (training set)
3. Perform an inner CV loop on the remaining data (validation set):
  1. Draw a test set from the validation set which contains more patients than controls by factor  $F$
  2. Apply classifier trained in step 2 to the unbalanced test set
  3. Compute predictions and performance metrics
  4. Adjust predictions for class frequencies and recompute performance metrics
4. Repeat steps 3a –d until each patient has been sampled once

We employed the above procedure, so that the test dataset was unbalanced by a factor of  $F = 5, 10$  or  $20$  (i.e. having 5, 10 or 20 controls for every patient which corresponds to disease prevalence of 20, 10 and 5%, respectively) and report the average performance over 10 random splits of the data for each level of this factor. For comparison, we also computed the predictive performance on the entire validation set at once which shows the accuracy that would have been obtained if the same subjects were tested using a balanced design (i.e. split-half CV). For each classification approach, we report the sensitivity, specificity, PPV, NPV, accuracy, balanced accuracy (Brodersen et al., 2010) and the overall predictive value (OPV), which we define as the mean of PPV and NPV (Fig. 2). The accuracy describes the proportion of correct predictions overall and balanced accuracy describes the mean proportion of correct predictions for each class (thereby accommodating the different class frequencies). If the class frequencies in the test set are equal (as for the LOO-CV approach described in section Leave-one-out cross-validation), the accuracy and balanced accuracy are equivalent.

It is unreasonable to assume that the chance levels for a classifier applied to an unbalanced test dataset are equal to 50%, so we employed permutation testing to (i) derive empirical chance level for each of the classifier performance measures noted above and (ii) to assess whether the classifiers exceeded chance for the balanced accuracy and

OPV (the two most appropriate summary measures for unbalanced datasets). We achieved this by retraining each classifier 1000 times with training labels randomly assigned to either of the two groups. This provided a distribution of accuracies under the null hypothesis that the classifier did not exceed chance. We report the mean of this distribution as the empirical chance level for each performance measure and we assessed significance by (i) counting the number of times the permuted balanced accuracy or OPV was greater than or equal to their true values and (ii) dividing each of these numbers by 1000.

## RESULTS

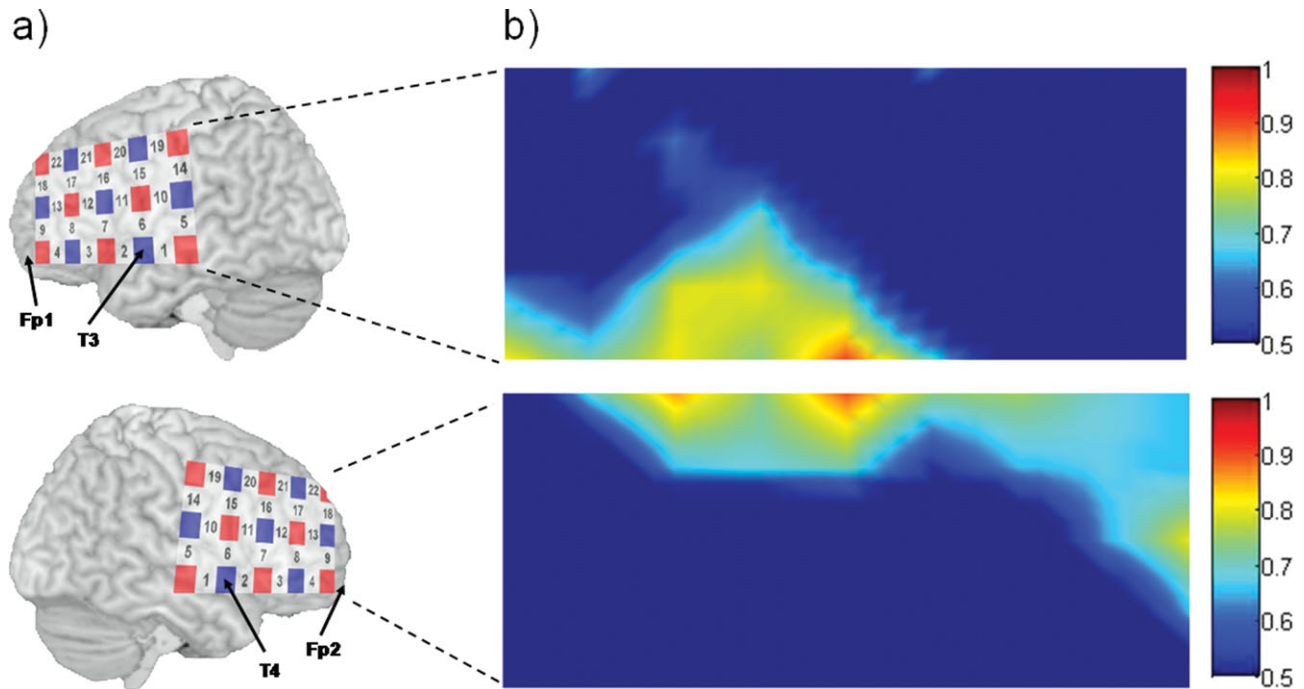
### Visual Checkerboard Paradigm

Independent GP classification based on the temporal information from each of the 52 fNIRS channels revealed significant accuracies for the classification of all three contrast conditions versus the no-contrast condition (Fig. 3). As hypothesized, the maximum classification accuracy was found at channels located over the primary visual cortex (channels 26, 25 and 15 for high-, medium- and low-contrast intensity, respectively) as previously identified using the functional localizer (see section Task description and procedures). Likewise, a Wilcoxon signed rank test revealed that the mean accuracy of all channels over the visual cortex region of interest (ROI) was higher than the mean accuracy of the channels outside this ROI. This effect was present in all three contrast conditions (97% vs. 0% contrast:  $Z = 2.54$ ;  $P < 0.011$ ; 40% vs. 0% contrast:  $Z = 2.52$ ;  $P < 0.012$ ; 8% vs. 0% contrast:  $Z = 2.38$ ;  $P < 0.018$ ). The median accuracy of all GP classifiers based on channels over the visual cortex was at 93%, 91 and 84% for high-, medium- and low-contrast intensity, respectively.

### Classification in the Context of Schizophrenia

Using the temporal information from the two probe sets each comprising the 22 fNIRS channels for both conditions, revealed significant accuracies for the classification of schizophrenic patients versus healthy controls. Significant classification accuracy was found for the 2-back





**Figure 4.**

(a) Schematic arrangement of the fNIRS probe set (red squares: emitter; blue squares: detectors; numbers: measurement channels). (b) Spatial accuracy maps for the classification of schizophrenic patients and healthy controls (2-back condition;  $P < 0.05$ , FDR-corrected).

condition at channels located over inferior fronto-temporal areas, extending into the dorsolateral prefrontal cortex showing accuracies as high as 78% (channels 2, 3, 4, 7, 8, 9, 12, 17, 20, 22 in the left probe set and channels 1, 2, 3, 4, 6, 7, 8, 9, 13, 18 in the right probe set; Fig. 4).

In the 1-back condition, a number of channels also showed significant accuracy (channels 14, 19, 22 in the left probe set and channels 1, 2, 17, 18, 21 in the right probe set).

Integrating spatial and temporal information using the spatiotemporal classification approach leads to an accuracy of 76% ( $P < 0.001$ ; sensitivity = 80%; specificity = 72.5%, PPV = 73.8%, NPV = 76.3%) for the 2-back condition and to an accuracy of 67.5% ( $P < 0.001$ ; sensitivity = 70%;

specificity = 65%, PPV = 66.7%, NPV = 68.4%) for the 1-back condition.

### Accommodating Variable Class Priors

Classifier performance measures derived from the artificially unbalanced datasets are summarized in Tables III–V. The split-half accuracy was overall slightly lower than the LOO-CV accuracy, probably because of the smaller training set split half CV entails. As expected, NPV increases monotonically and PPV decreases as the test set becomes more unbalanced with PPV becoming very poor for the

**TABLE III. Classifier performance for a balanced training set and unbalanced test set (five patients for every control)**

	Sensitivity	Specificity	PPV	NPV	Accuracy	Bal. Acc.	OPV
Split half	73.0 (1.0)	69.0 (0.7)	70.4 (0.4)	72.4 (0.8)	71.0 (0.4)	71.0 (0.4)	71.4 (0.4)
Not adjusted	71.5 (1.1)	73.1 (0.9)	38.1 (1.1)	87.6 (2.0)	73.2 (0.7)	72.3 (0.5)	62.8 (1.3)
Adjusted	58.5 (0.6)	87.1 (0.7)	55.3 (1.6)	91.3 (0.1)	82.3 (0.5)	72.8 (0.3)	73.3 (0.8)
Chance: n. adj.	49.9	50.0	16.6	83.3	50.0	49.9	58.3
Chance: adj.	24.4	75.4	16.6	83.3	66.9	49.9	58.3

All statistics are reported as percentages and numbers in brackets indicate standard errors across 10 random splits of the data. Chance levels on the unbalanced test set for the adjusted and non-adjusted classifiers are also reported. PPV = positive predictive value, NPV = negative predictive value, OPV = overall predictive value.

**TABLE IV. Classifier performance for a balanced training set and unbalanced test set (ten patients for every control)**

	Sensitivity	Specificity	PPV	NPV	Accuracy	Bal. Acc.	OPV
Split half	74.0 (1.0)	67.5 (0.9)	69.7 (0.4)	72.9 (0.9)	70.8 (0.4)	70.8 (0.4)	71.0 (0.5)
Not adjusted	74.0 (1.0)	68.3 (0.9)	20.1 (0.6)	96.2 (0.2)	68.8 (0.8)	71.1 (0.5)	58.1 (0.3)
Adjusted	51.5 (0.7)	88.5 (0.7)	40.7 (2.5)	89.4 (1.4)	85.4 (0.6)	70.0 (0.5)	65.0 (1.6)
Chance: n. adj.	50.1	50.0	9.1	90.1	50.0	50.0	54.6
Chance: adj.	16.2	83.5	9.0	90.1	77.4	49.9	54.4

All statistics are reported as percentages and numbers in brackets indicate standard errors across ten random splits of the data. Chance levels on the unbalanced test set for the adjusted and non-adjusted classifiers are also reported. PPV = positive predictive value, NPV = negative predictive value, OPV = overall predictive value.

most unbalanced dataset. Further, the PPV derived from the (balanced) split half CV approach becomes increasingly unrepresentative of the performance of the classifier for the unbalanced data. Crucially, when the predictions are adjusted to compensate for class priors in the test set, PPV improves in every case and the NPV of the adjusted classifier remains approximately equivalent to the non-adjusted classifier. The accuracy from the classifier also increases substantially after the predictive probabilities have been adjusted, indicating an increase in the total number of correct predictions. However, this is not of interest from a diagnostic perspective since it mainly reflects the imbalance in the test set given that the adjusted classifier is more likely to correctly predict membership of the larger class. In contrast, the balanced accuracy remains virtually identical for all classifiers, indicating that the average proportion of correct predictions for each class remained constant before and after the adjustment of the predictive probabilities. At first glance, the sensitivity of the adjusted classifier appears relatively poor but, it should be emphasized that the sensitivity reflects the proportion of correct predictions the classifier obtained against the weight of evidence that most cases are expected to be derived from the larger class. To emphasize this point, comparison with the empirical chance levels (Table III) shows that the sensitivity and specificity of the adjusted classifier exceeded chance in every case.

The permutation test for the balanced accuracy on the unbalanced test datasets showed that both the adjusted and the non-adjusted classifiers exceeded chance level for

$F = 5, 10$  and  $20$ . Further, the permuted balanced accuracy never exceeded its true value, corresponding to a significance of  $P = 0.001$  for each classifier. The permutation test for the OPV showed that in all cases the adjusted classifier produced greater OPV than would be expected by chance ( $P = 0.002, P = 0.006$  and  $P = 0.001$  respectively for  $F = 5, 10$  and  $20$ ). In contrast, the OPV for the non-adjusted classifier did not exceed chance for any level of  $F$  ( $P = 0.08, P = 0.07$  and  $P = 0.36$ , respectively), which as expected reflects the poor performance of the non-adjusted classifier in an unbalanced test setting.

## DISCUSSION

Using the newly developed temporal pattern classifier for fNIRS, we showed that temporal information is sufficient to differentiate between conditions in a visual checkerboard task. Further validating the approach, high accuracies were specifically found at channels over the primary visual cortex while classification accuracy was lower in other regions. Most importantly, GP classification based solely on the temporal dynamics during an n-back task allowed for significant single-subject classification of schizophrenic patients and healthy controls. Furthermore, the maximum accuracy of 78% and the combined accuracy of 76% obtained from spatial integration are in the range of the highest ever achieved in biomarker-based classification of schizophrenic patients. In addition, our accuracy estimation is based on one of the largest schizophrenic samples

**TABLE V. Classifier performance for a balanced training set and unbalanced test set (20 patients for every control)**

	Sensitivity	Specificity	PPV	NPV	Accuracy	Bal. Acc.	OPV
Split half	71.0 (0.7)	68.5 (1.0)	70.9 (0.5)	70.4 (0.5)	69.8 (0.5)	69.8 (0.5)	70.7 (0.4)
Not adjusted	71.0 (0.7)	69.5 (0.9)	11.0 (0.2)	94.9 (1.1)	69.1 (0.8)	70.3 (0.4)	53.0 (0.6)
Adjusted	48.5 (1.1)	90.0 (0.7)	29.3 (2.8)	97.2 (0.0)	89.8 (0.4)	69.3 (0.3)	63.3 (1.4)
Chance: n. adj.	50.4	50.0	4.8	95.3	50.0	50.1	52.4
Chance: adj.	14.8	85.2	4.8	95.2	81.9	50.0	52.4

All statistics are reported as percentages and numbers in brackets indicate standard errors across 10 random splits of the data. Chance levels on the unbalanced test set for the adjusted and non-adjusted classifiers are also reported. PPV = positive predictive value, NPV = negative predictive value, OPV = overall predictive value.

ever considered in functional biomarker research of this disorder (for a review and methodological considerations regarding previous studies, see Demirci et al., 2008). Finally, we show that the predictions derived from our probabilistic classification approach can be easily adjusted to compensate for variable class priors, which is highly advantageous if it is necessary to make predictions in different populations from those with which the classifier was trained.

Specifically considering the results of the visual checkerboard experiment, the algorithm appears to be sensitive to task-related differences in the temporal dynamics of the BOLD response. As expected, accuracies were higher over the visual cortex than over other regions. While these results demonstrate sufficient sensitivity and specificity of the algorithm, a number of questions arise: Generally, the effects observed in the visual checkerboard experiment are relatively large and regularly observable in the time-series of single trials with the naked eye. It remains to be investigated how reliability impacts classification. This is particularly important as pattern recognition requires a more or less reliable pattern in single samples (i.e. for every subject and condition). Generally, most classifiers including the one used in our approach only assume that the training samples are independent and identically distributed.

Additionally, it might be of interest which properties of the time-series drive classification. In analogy to the spatial weight-mapping method described by Marquand et al. (2010), temporal weight-mapping could quantify the contribution of each time-point to the formation of the decision boundary. When the distribution of the samples with respect to the decision boundary is of interest, temporal g-maps in analogy to the spatial g-maps described in Marquand et al. (2010) could also be of interest (an example of each temporal map can be found in Supporting Information, Fig. 1). One should, however, be aware that the maps represent a multivariate temporal pattern, and one ought to be cautious about interpreting time-points or time-periods in isolation. Coefficient scores (weights) for each time-point should thus be interpreted in the context of the entire multivariate pattern. Nonetheless, if hypotheses about the temporal pattern exist, discrimination mapping might prove helpful in visualizing what the classifier has learned.

As we perform GP classification at each fNIRS channel independently, it also needs to be noted that the same accuracies might result from different temporal patterns. In standard GLM group analyses, a similar phenomenon is evident, for instance, when one or more derivatives are included in the model. In such cases, the same beta-value might be estimated even though the latency or the dispersion of the subjects' hemodynamic response can differ.

Differentiating between schizophrenic patients and healthy controls, the temporal pattern recognition algorithm demonstrates its suitability for predicting group membership for individual subjects. Further validating the approach, it yields above chance accuracies in those regions which have previously been associated with the

n-back task in group analyses comparing schizophrenic patients and controls (e.g. Reif et al., 2011). In particular, temporal information in the inferior fronto-temporal areas is relevant for classification. As other approaches using for instance large anatomical (Davatzikos et al., 2005; Job et al., 2006; Kawasaki et al., 2007; Nakamura et al., 2004; Takayanagi et al., 2010) or functional MRI data sets (Calhoun et al., 2008; Pokrajac et al., 2005; Shen et al., 2010) do not substantially outperform our approach which is based on a comparatively small amount of data per subject, using temporal information in biomarker-based classification appears highly promising.

Notwithstanding the need for further investigation of the properties of our approach using different experimental designs (i.e. segment lengths, particularly in rapid event-related designs), the advantages of temporal pattern classification appear substantial: From a methodological point of view, the approach can be considered model-free in the sense that no prior assumptions about parameters of the hemodynamic response (such as shape or latency) are required. The GP classifier merely learns to differentiate temporal patterns which can then be applied to predict class memberships based on new data. In contrast to ordinary regression analysis, temporal (auto-) correlations within the data are therefore not only unproblematic in GP classification, but temporal structure (i.e. regular patterns or oscillations) can be learned and utilized to increase classification accuracy. In fact, performing GP classification on the Fourier transformed (frequency) data is equivalent to conducting the analysis on the original data. In this context, future studies could furthermore use specific priors to test hypotheses concerning the discriminative power of specific parts of the frequency spectrum.

A key feature of the probabilistic approach to classification is that it aims to be coherent about managing uncertainty throughout the analysis, which ultimately provides predictions that accurately reflect the confidence of the classifier decision. We demonstrated in this work that a simple numerical adjustment of these probabilistic predictions enables the classifier to easily compensate for variable class priors. For the proof-of-concept application we presented here, the adjusted classifier showed significant predictive value even in the most unbalanced test set. Further, without this adjustment the classifier was unable to produce above chance predictive value in any of the unbalanced test settings, despite achieving impressive sensitivity and specificity and a balanced accuracy significantly above chance. This approach is potentially beneficial for improving the predictive value (i.e. limiting the error rate) of the classifier in a wide range of clinical neuroimaging contexts where the class priors may be variable (e.g. predicting relapse, remission or treatment response in addition to diagnosis). Other advantages of this approach are that it does not require the class frequencies in the test set to be specified during training, and there is no need to retrain the classifier if the class frequencies in the test set change. Note that this adjustment

cannot be performed if the classifier does not provide probabilistic predictions even if the classifier provides a numerical decision function that serves as a surrogate for classifier confidence (since the decision function may be scaled arbitrarily). There is also no reason why this approach is limited to GPC and it is suitable for any classifier that provides probabilistic predictions. However we favour GPC in practice because in previous work, we found that it produced more accurate predictions for fMRI data than alternative probabilistic classifiers such as the relevance vector machine (Marquand et al., 2010). In addition, the capability of GPC to automatically estimate parameters from the data is very useful in a wide range of neuroimaging contexts and underpins the spatiotemporal classification approach proposed in this paper.

In addition, the temporal classifier can be used to obtain spatial maps showing where in space temporal dynamics are most informative. As the temporal classifiers are run independently at each spatial location of interest, the resulting spatial maps (i.e. the accuracy maps) are not multivariate in nature and can thus be interpreted more easily than those commonly obtained using spatial or spatial-temporal pattern recognition approaches (Mourao-Miranda et al., 2007). While more computationally demanding, our approach thereby addresses a fundamental issue regarding the interpretation of spatial maps obtained from other biomarker-based machine learning approaches to neuroimaging.

From the point of view of practical application, the algorithm in combination with fNIRS measurements enables the analysis of sub-second, multivariate temporal patterns within the BOLD response and thus allows for an investigation of potentially clinically relevant temporal dynamics differentiating patients and healthy individuals or task related mental states within patient populations. As our approach provides the classification probability for each participant, knowledge about prevalence of a disorder can easily be incorporated. Considering the ecological validity of fNIRS (absence of (1) horizontal position, (2) head-fixation, (3) anxiety inducing environment in a narrow scanner, (4) loud measurement sequences) combined with the low sensitivity for movement artefacts and the ease and low cost with which fNIRS can be used with virtually any patient group “at bedside”, biomarker-based classification as described here might be of particular importance for practical applications in the future. Most importantly, these practical applications in the field of psychiatry can not only consist in the differentiation between patients with a certain disorder from healthy controls, but in the prediction of the success of a specific therapy in an individual patient.

## ACKNOWLEDGMENTS

AM gratefully acknowledges support from the King’s College Annual Fund and the King’s College London Centre of Excellence in Medical Engineering. The project was partially

supported by the German Research Foundation (SFB TRR 58, subproject C4) and the IZKF of the University of Wuerzburg. The authors declare no conflicts of interest.

## REFERENCES

- Arbeitsgemeinschaft fMuDidP. 2000. Das AMDP-System. Manual zur Dokumentation psychiatrischer Befunde. Göttingen, Germany: Hogrefe.
- Avidan G, Harel M, Hendler T, Ben-Bashat D, Zohary E, Malach R (2002): Contrast sensitivity in human visual areas and its relationship to object recognition. *J Neurophysiol* 87(6):3102–3116.
- Bishop CM (2007): *Pattern Recognition and Machine Learning*. Berlin: Springer. p.738.
- Boynton GM, Demb JB, Glover GH, Heeger DJ (1999): Neuronal basis of contrast discrimination. *Vision Res* 39(2):257–269.
- Brodersen KH, Cheng Soon O, Stephan KE, Buhmann JM (2010): The balanced accuracy and its posterior distribution. *Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR 2010)*, August 23–26, Istanbul, Turkey.
- Calhoun VD, Maciejewski PK, Pearlson GD, Kiehl KA (2008): Temporal lobe and “default” hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. *Hum Brain Mapp* 29(11):1265–1275.
- Davatzikos C, Shen D, Gur RC, Wu X, Liu D, Fan Y, Hughett P, Turetsky BI, Gur RE (2005): Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch Gen Psychiatry* 62(11):1218–1227.
- Demirci O, Clark VP, Magnotta VA, Andreasen NC, Lauriello J, Kiehl KA, Pearlson GD, Calhoun VD (2008): A review of challenges in the use of fMRI for disease classification/characterization and a projection pursuit application from multi-site fMRI schizophrenia study. *Brain Imag Behav* 2(3):147–226.
- Dresler T, Obersteiner A, Schecklmann M, Vogel AC, Ehlis AC, Richter MM, Plichta MM, Reiss K, Pekrun R, Fallgatter AJ (2009): Arithmetic tasks in different formats and their influence on behavior and brain oxygenation as assessed with near-infrared spectroscopy (NIRS): a study involving primary and secondary school children. *J Neural Transm* 116(12):1689–1700.
- Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, Brammer MJ, Murphy C, Murphy DG (2009): Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *NeuroImage* 49(1):44–56.
- Ehlis AC, Herrmann MJ, Wagener A, Fallgatter AJ (2005): Multi-channel near-infrared spectroscopy detects specific inferior-frontal activation during incongruent Stroop trials. *Biol Psychol* 69(3):315–331.
- Fallgatter AJ, Ehlis A, Wagener A, Michel T, Herrmann MJ (2004): Near-infrared spectroscopy in psychiatry. *Nervenarzt* 75(9):911–916.
- Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, Brammer MJ (2008): Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry* 63(7):656–662.
- Gigerenzer G, Edwards A (2003): Simple tools for understanding risks: From innumeracy to insight. *Br Med J* 327(7417):741–744.
- Guyon I, Elisseeff A (2003): An introduction to variable and feature selection. *J Machine Learn Res* 3:1157–1182.

- Hahn T, Heinzel S, Plichta MM, Reif A, Lesch K-P, Fallgatter AJ (2010): Neurovascular coupling in the human visual cortex is modulated by cyclooxygenase-1 (COX-1) gene variant. *Cerebral Cortex* 68:459–464.
- Hahn T, Marquand AF, Ehliis AC, Dresler T, Kittel-Schneider S, Jarczok TA, Lesch KP, Jakob PM, Mourao-Miranda J, Brammer MJ, Fallgatter AJ (2011): Integrating neurobiological markers of depression. *Arch Gen Psychiatry* 68:361–368.
- Heeger DJ, Ress D (2002): What does fMRI tell us about neuronal activity? *Nat Rev Neurosci* 3(2):142–151.
- Herrmann MJ, Langer JB, Jacob C, Ehliis AC, Fallgatter AJ (2008): Reduced prefrontal oxygenation in Alzheimer disease during verbal fluency tasks. *Am J Geriatr Psychiatry* 16(2):125–135.
- Hoshi Y (2003): Functional near-infrared optical imaging: utility and limitations in human brain mapping. *Psychophysiology* 40(4):511–520.
- Jasper HA (1958): The ten-twenty system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* 10:371–375.
- Job DE, Whalley HC, McIntosh AM, Owens DG, Johnstone EC, Lawrie SM (2006): Grey matter changes can improve the prediction of schizophrenia in subjects at high risk. *BMC Med* 4:29.
- Kawasaki Y, Suzuki M, Kherif F, Takahashi T, Zhou SY, Nakamura K, Matsui M, Sumiyoshi T, Seto H, Kurachi M (2007): Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *NeuroImage* 34(1):235–242.
- Marquand AF, Mourao-Miranda J, Brammer MJ, Cleare AJ, Fu CH (2008): Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. *Neuroreport* 19(15):1507–1511.
- Marquand A, Howard M, Brammer M, Chu C, Coen S, Mourao-Miranda J (2010): Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage* 49(3):2178–2189.
- Minka TP (2001): A Family of Algorithms for Approximate Bayesian Inference. Massachusetts Institute of Technology.
- Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M (2005): Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage* 28(4):980–995.
- Mourao-Miranda J, Friston KJ, Brammer M (2007): Dynamic discrimination analysis: a spatial-temporal SVM. *NeuroImage* 36(1):88–99.
- Nakamura K, Kawasaki Y, Suzuki M, Hagino H, Kurokawa K, Takahashi T, Niu L, Matsui M, Seto H, Kurachi M (2004): Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. *Schizophr Bull* 30(2):393–404.
- Obrig H, Villringer A (2003): Beyond the visible—imaging the human brain with light. *J Cereb Blood Flow Metab* 23(1):1–18.
- Ozuz B, Liu HL, Chen L, Iyer MB, Fox PT, Gao JH (2001): Rate dependence of human visual cortical response due to brief stimulation: An event-related fMRI study. *Magn Reson Imag* 19(1):21–25.
- Platt J. 1999. Probabilities for SV machines. In: Sholkopf B, Burges C, Smola A, editors. *Advances in Large Margin Classifiers*. Cambridge: The MIT press. p 61–74.
- Plichta MM, Herrmann MJ, Baehne CG, Ehliis AC, Richter MM, Pauli P, Fallgatter AJ (2006a): Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable? *Neuroimage* 31(1):116–124.
- Plichta MM, Herrmann MJ, Ehliis AC, Baehne CG, Richter MM, Fallgatter AJ (2006b): Event-related visual versus blocked motor task: Detection of specific cortical activation patterns with functional near-infrared spectroscopy. *Neuropsychobiology* 53(2):77–82.
- Plichta MM, Heinzel S, Ehliis AC, Pauli P, Fallgatter AJ (2007a): Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data: A parametric validation study. *Neuroimage* 35(2):625–634.
- Plichta MM, Herrmann MJ, Baehne CG, Ehliis AC, Richter MM, Pauli P, Fallgatter AJ (2007b): Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: reproducibility of activation? *Hum Brain Mapp* 28(8):733–741.
- Pokrajac D, Megalooikonomou V, Lazarevic A, Kontos D, Obradovic Z (2005): Applying spatial distribution analysis techniques to classification of 3D medical images. *Artif Intell Med* 33(3):261–280.
- Rasmussen CE, Williams CKI. 2006. *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT press.
- Reif A, Schecklmann M, Eirich E, Jacob CP, Jarczok TA, Kittel-Schneider S, Lesch KP, Fallgatter AJ, Ehliis AC (2011): A functional promoter polymorphism of neuronal nitric oxide synthase moderates prefrontal functioning in schizophrenia. *Int J Neuropsychopharmacol* 14:1–11.
- Schecklmann M, Ehliis AC, Plichta MM, Fallgatter AJ (2008a): Functional near-infrared spectroscopy: a long-term reliable tool for measuring brain activity during verbal fluency. *Neuroimage* 43(1):147–155.
- Schecklmann M, Ehliis AC, Plichta MM, Romanos J, Heine M, Boreatti-Hummer A, Jacob C, Fallgatter AJ (2008b): Diminished prefrontal oxygenation with normal and above-average verbal fluency performance in adult ADHD. *J Psychiatr Res* 43(2):98–106.
- Shen H, Wang L, Liu Y, Hu D (2010): Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *NeuroImage* 49(4):3110–3121.
- Suto T, Fukuda M, Ito M, Uehara T, Mikuni M (2004): Multichannel near-infrared spectroscopy in depression and schizophrenia: Cognitive brain activation study. *Biol Psychiatry* 55(5):501–511.
- Takayanagi Y, Kawasaki Y, Nakamura K, Takahashi T, Orikabe L, Toyoda E, Mozue Y, Sato Y, Itokawa M, Yamasue H, Kasai K, Kurachi M, Okazaki Y, Matsushita M, Suzuki M (2010): Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. *Prog Neuropsychopharmacol Biol Psychiatry* 34(1):10–17.
- Tipping M (2001): Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* 1:211–244.
- Tripoliti EE, Fotiadis DI, Argyropoulou M, Manis G (2009): A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data. *J Biomed Inform* 43:307–320.
- Zeller JB, Herrmann MJ, Ehliis AC, Polak T, Fallgatter AJ (2010): Altered parietal brain oxygenation in Alzheimer's disease as assessed with near-infrared spectroscopy. *Am J Geriatr Psychiatry* 18(5):433–441.
- Zhu CZ, Zang YF, Cao QJ, Yan CG, He Y, Jiang TZ, Sui MQ, Wang YF (2008): Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *NeuroImage* 40(1):110–120.