

# Permutations of functional magnetic resonance imaging classification may not be normally distributed

[Show less](#)

Mohammed S Al-Rawi, Adelaide Freitas, João V Duarte, Joao P Cunha, Miguel Castelo-Branco

First Published December 18, 2017 Research Article

## Article Information

Volume: 26 issue: 6, page(s): 2567-2585

Article first published online: December 18, 2017; Issue published: December 1, 2017

<https://doi.org/10.1177/0962280215601707>

Mohammed S Al-Rawi<sup>1, 2</sup>, Adelaide Freitas<sup>3</sup>, João V Duarte<sup>2</sup>, Joao P Cunha<sup>4, 5</sup>, Miguel Castelo-Branco<sup>1, 2</sup>

<sup>1</sup>The Institute of Nuclear Sciences Applied to Health, University of Coimbra, Coimbra, Portugal

<sup>2</sup>Visual Neuroscience Laboratory, IBILI – Institute for Biomedical Imaging and Life Sciences, Faculty of Medicine, University of Coimbra, Coimbra, Portugal

<sup>3</sup>Department of Mathematics, Center for Research & Development in Mathematics and Applications, University of Aveiro, Aveiro, Portugal

<sup>4</sup>IEETA, University of Aveiro, Aveiro, Portugal

<sup>5</sup>Department of Electrical and Computer Engineering, University of Porto, Porto, Portugal

Corresponding Author: Mohammed S Al-Rawi, The Institute of Nuclear Sciences Applied to Health, University of Coimbra, Azinhaga de Santa Comba, Coimbra 3000-548, Portugal.

Email: [al-rawi@uc.pt](mailto:al-rawi@uc.pt)

## Abstract

A fundamental question that often occurs in statistical tests is the normality of distributions. Countless distributions exist in science and life, but one distribution that is obtained via permutations, usually referred to as permutation distribution, is interesting. Although a permutation distribution should behave in accord with the central limit theorem, if both the independence condition and the identical distribution condition are fulfilled, no studies have corroborated this concurrence in functional magnetic resonance imaging data. In this work, we used Anderson–Darling test to evaluate the accordance level of permutation distributions of classification accuracies to normality expected under central limit theorem. A simulation study has been carried out using functional magnetic resonance imaging data collected, while human subjects responded to visual stimulation paradigms. Two scrambling schemes are evaluated: the first based on

permuting both the training and the testing sets and the second on permuting only the testing set. The results showed that, while a normal distribution does not adequately fit to permutation distributions most of the times, it tends to be quite well acceptable when mean classification accuracies averaged over a set of different classifiers is considered. The results also showed that permutation distributions can be probabilistically affected by performing motion correction to functional magnetic resonance imaging data, and thus may weaken the approximation of permutation distributions to a normal law. Such findings, however, have no relation to univariate/univoxel analysis of functional magnetic resonance imaging data. Overall, the results revealed a strong dependence across the folds of cross-validation and across functional magnetic resonance imaging runs and that may hinder the reliability of using cross-validation. The obtained  $p$ -values and the drawn confidence level intervals exhibited beyond doubt that different permutation schemes may beget different permutation distributions as well as different levels of accord with central limit theorem. We also found that different permutation schemes can lead to different permutation distributions and that may lead to different assessment of the statistical significance of classification accuracy.

**Keywords** [Permutation testing](#), [normality](#), [classification analysis](#), [Anderson–Darling test](#), [central limit theorem](#)

## 1 Introduction

Decoding mental states from patterns of brain activity in humans relies mainly on using classification techniques on brain images acquired via functional magnetic resonance imaging (fMRI).<sup>1–3</sup> The decoding devolves on MultiVoxel pattern analysis (MVPA) of fMRI data by using classifiers. There has been a non-trivial interest in this method with aims to uncover the brain mysteries and understanding how it works. The statistical significance of the classification procedure has been raised as an important issue to confirm successful mental decoding. Consequently, in neuroimaging studies where researchers usually work on constrained datasets that may suffer from low sample size and high dimensionality, classification analysis are usually performed via cross-validation. One can obtain a  $p$ -value of the classification accuracy by using the binomial distribution (the multinomial distribution if more than two classes).<sup>4</sup> This approach, however, fails because of using

cross-validation and another method has to be used to find the  $p$ -values of the classification accuracy.<sup>5,6</sup> Since the class distributional properties are unknown and the distributional properties of the test statistic are complex, permutation tests have been suggested to estimate the classification significance.<sup>2</sup> These methods are based on the null hypothesis that classes have identical distributions and an attempt is made to reject the hypothesis and prove otherwise. The significance of the classification can be estimated by measuring how far does the accuracy deviate from theoretical chance-level. Permutation testing is inspired from Fisher's exact test that he proposed in the 1930s.<sup>7</sup> When applied in pattern classification, the method simply estimates how much the test error of the classifier deviates from the permuted test error values of the same classifier. As it is computationally impractical to cover all the possible permutations as in Fisher exact test, applying Monte Carlo method is eminent. It is worth mentioning that permutation testing applied to neuroimaging has also been used in univariate analysis of fMRI data to detect peak activations in response to stimuli.<sup>8</sup>

When running permutation testing procedures to estimate the classification significance, a classifier is used to classify the data of each permutation and an empirical permutation distribution (PD) from the permuted classification accuracies is built. Thus, a PD represents an empirical estimate of the cumulative distribution of the test error (or the accuracy) of the classifier under the null hypothesis of independence between the data and the labels. The test error or the accuracy can be estimated using cross-validation in each iteration of the permutation procedure. As mentioned in Stelzer et al.,<sup>6</sup> the theoretical null distribution of the accuracy (percentage of correctly estimated labels) in each cross-validation fold is proportional to a binomial distribution. Consequently, in accordance with the central limit theorem (CLT), a PD calculated in a cross-validation classification manner should approximate to normal law if both the independence condition and the identical distribution condition among the validation folds are fulfilled. Moreover, Golland et al.<sup>9</sup> proved that the empirical leave-one-out cross-validation error for finite VC classifiers will concentrate to the theoretical value expected under the null hypothesis. Based on different cross-validation schemes over artificially generated random data and real clinical data, however, Noirhomme et al.<sup>10</sup> found cases where PDs can be matched to the

binomial distribution and cases where PDs are differed significantly from the binomial distribution.

fMRI data are usually acquired in several independent runs and they may not have enough samples to enable running the classifier without partitioning the data. One of the most renowned partitioning schemes is called cross-validation where the data are partitioned into several folds such that one fold is reserved for testing the classifier and the remaining folds are used in training the classifier. It is quite useful, if possible, to consider each run as a fold that enables performing leave-one-run out cross-validation. After finding the classification accuracy for each fold, the overall accuracy (or the error) is measured by taking the mean of the accuracy. Although there has been criticisms of using cross-validation in fMRI classification studies,<sup>5</sup> other works highlighted the importance of using cross-validation and did several analyses to show how useful permutation tests are.<sup>11,12</sup> Furthermore, several practice situations have widely accepted that possible dependence features associated with non-overlapping folds or cross-validation could be negligible.<sup>3</sup> Despite the fact that it is commonly accepted that every PD is normally distributed and centered on the theoretical chance-level, there has not been any study to confirm this. Hence, it is still not known whether any PD in neuroimaging studies follows a normal distribution or not, and since there are numerous classifiers to consider, it is also not known yet whether PDs obtained from which of these classifiers will accord to CLT. To that end, this work could be considered as another attempt to investigate the pros and cons of cross-validation and whether they can adequately retort on prediction from fMRI data.

It is clear that investigating the accordance of PDs to CLT is a vital topic to fMRI classification and could facilitate obtaining enhanced mental states decoding techniques. The motives behind measuring the normality of PDs are: comparison different classification models, inferring the dependence across fMRI runs, finding enhanced data partitioning schemes, spotlighting the method of cross-validation, and whether it can be considered as inadequate partitioning procedure.<sup>5</sup> To investigate the shapes of PDs in this work, several PDs are built via classifying fMRI data taken when the subjects responded to visual stimulation paradigms. Anderson–Darling test (AD-test)

will be used to infer whether a PD is normal or not, i.e., to test the null hypothesis that the random sample of the test error values of a classifier, or a system of classifiers, is generated by a normal distribution.

## 2 Methods

### 2.1 Testing normality

AD-test<sup>13,14</sup> has been considered as one of the most powerful normality tests in the literature.<sup>15</sup> AD-test statistic is defined as follows

$$AD = n \int_{-\infty}^{+\infty} (F_n(x) - \Phi(x))^2 [\Phi(x)(1 - \Phi(x))]^{-1} d\Phi(x) \quad (1)$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution,  $F_n(\cdot)$  is the empirical distribution function, and  $n$  is the sample size. When the parameters of the normal distribution are estimated from the empirical distribution, a normalization factor has been proposed for improving the power of the original formulation<sup>16</sup>

$$AD^* = AD(1 + 0.75/n + 2.25/n^2) \quad (2)$$

### 2.2 Classifier's test error

Cross-validation procedures are commonly used for evaluating the performance of classification models by estimating the expected level-of-fit of the models. These procedures are of extreme importance when the data has fewer samples, as in the case of fMRI classification analysis. One round of cross-validation involves partitioning the sampled data into two complementary subsets. From one subset (called the training set), the parameters of the model are estimated and from the other subset (called the testing set), the estimated model is validated. The validation is based on a distance measure of the quality of fit of the estimated model. This distance is calculated by comparing the labels observed from the sample belonging to the testing set and the labels predicted by the estimated model. For nominal labels, as it occurs in brain decoding experiments, the distance measure commonly computed is the so-called test error. In cross-validation methods, the test error is defined by the proportion of data belonging to the testing set incorrectly classified by the estimated model and represents an estimate of

the misclassification error rate of the classification model in the prediction of future data. In order to reduce variability, multiple rounds of cross-validation are performed using different partitions such that the final test error, given by the average of the test errors over the rounds, is the expected level-of-fit of the classification model, and thus, can be used for assessing its performance.

There are few types of cross-validation techniques depending on how the partition of the sampled data is made and the mostly used is called  $J$ -fold cross-validation. In the  $J$ -fold cross-validation, the original sampled data are randomly partitioned into  $J$  equal size subsamples and the cross-validation process is repeated  $J$  times such that, in each iteration, a single subsample of the  $J$  subsamples is used as a testing set, and the remaining  $(J-1)$  subsamples are used as a training set. To describe the mathematical formulation of the  $J$ -folds cross-validation, consider a linear classification model given by

$$y = Wx + b \tag{3}$$

where  $\mathbf{W}$  and  $\mathbf{b}$  are the model parameters and  $\mathbf{y}$  is the classifier's output, which indicates the nominal label predicted by the model for the observation  $\mathbf{x}$ . Let  $N$  be the number of subjects involved in the analysis. For subject  $i$ ,  $i = 1, 2, \dots, N$ , let a sampled data and its corresponding label set be denoted by  $x^{(i)}$  and  $Y^{(i)}$ , respectively. To assess the fit of the model shown in [equation \(3\)](#), the test error is computed as a fit measure using  $J$ -fold cross-validation on the data set of the subject  $i$ .

Consider the subject  $i$  and a partition  $\Omega = \cup_{j=1}^J \Omega_j$  of its original dataset in  $J$  subsets,  $\Omega_1, \Omega_2, \dots, \Omega_J$ . For the  $j$ th iteration of cross-validation, the dataset of the samples  $x^{(i)}$  and their corresponding labels  $Y^{(i)}$  are split into two sets: a training set defined by  $S_{1j} = \cup_{k=1, k \neq j}^J \Omega_k$  and a testing set defined by  $S_{2j} = \Omega_j$ . In the training phase, the model (3) is fitted to subject  $i$  such that the fitted model obtained in the  $j$ th cross-validation iteration can be written as

$$y^{(i)} = \hat{W}^{(i,j,1)} x + \hat{b}^{(i,j,1)} \tag{4}$$

where  $\hat{W}^{(i,j,1)}$  and  $\hat{b}^{(i,j,1)}$  are the estimates of the parameters of the classification model (3) obtained from the data  $x^{(i)}$  belonging to the training set  $S_{1j}$ , say  $x^{(i,j,1)}$  and their corresponding labels  $Y^{(i,j,1)}$ , of subject  $i$ . Note that  $Y^{(i,j,1)}$  refers to the original (non-permuted) labels and it only belongs to the samples used in the training phase into the  $j$ th cross-validation iteration. In the testing phase, the test error is obtained using the estimated model (4) and the data  $x^{(i)}$  and their corresponding labels  $Y^{(i)}$  belonging to the testing set. Note that the parameters of the classifier estimated in the training phase, in this case the weight pair  $(\hat{W}^{(i,j,1)}, \hat{b}^{(i,j,1)})$ , characterizes the estimated classification model (4), which is used to calculate an estimate of the test error in the testing phase. This estimate of the test error is the proportion of the data  $x^{(i)}$  belonging to the testing set incorrectly classified by the model (4). Formally, it represents the difference between the observed labels and the corresponding labels predicted by the estimated model (4), for all the data  $x^{(i)}$  belonging to the testing set. Then, the test error obtained in the  $j$ th iteration of cross-validation, denoted by  $e^{j,i}$ , can be written as

$$e^{j,i} = \frac{1}{n} \sum_{k=1}^n \|\hat{W}^{(i,j,1)} x^{(i,j,2)} + \hat{b}^{(i,j,1)} - Y^{(i,j,2)}\| \quad e^{j,i} = \frac{1}{n} \sum_{k=1}^n \|\hat{W}^{(i,j,1)} x^{(i,j,2)} + \hat{b}^{(i,j,1)} - Y^{(i,j,2)}\| \quad (5)$$

where  $x^{(i,j,2)}$  and  $Y^{(i,j,2)}$  are the testing data and their corresponding labels, respectively, belonging to the testing set  $S_{2j}$  of subject  $i$ .

To measure the performance of the classification model, given in [equation \(3\)](#), for subject  $i$ , the test error, given by the mean of the estimates ([equation \(5\)](#)) over  $J$  iterations, is considered. Therefore, the test error of the classification model (3) for subject  $i$  can be written as

$$e^i = \frac{1}{J} \sum_{j=1}^J e^{j,i} \quad e^i = \frac{1}{J} \sum_{j=1}^J e^{j,i} \quad (6)$$

The expression (6) is a function of the data  $x^{(i)}$  and the corresponding labels  $Y^{(i)}$  and so it is a statistic. Furthermore, [equation \(6\)](#) represents an estimative of the expectation of the misclassification error rate of the classifier (3) which can be written as  $E(\|y^{\wedge} - y\|)$ , where  $y^{\wedge}$  is the label predicted by some classifier that was estimated from the training set. For all the analyses in this work, a measure of classification accuracy  $a^i$  will be used, which is given by

$$a^i = 1 - e^{-a^i} = 1 - \frac{1}{J} \sum_{j=1}^J e^{-a^j}, i=1, \dots, J \quad (7)$$

and for a classifier denoted as  $C$ , let the classification accuracy be denoted as  $a^i(C)$ . Remark that, as observed in Stelzer et al.,<sup>6</sup> under the assumption of independence of the  $J$  cross-validation folds, the accuracy value obtained at the end of the cross-validation process could be represented in terms of the constant  $1/J$  times a binomial random of  $J \times k$  trials, and then, for a large number of these trials, it could be approximated to a normal probability law in accordance to CLT. In practice  $k$  represents the number of samples in the testing set of each cross-validation fold ( $k$  is 18 in this work, as will be shown later).

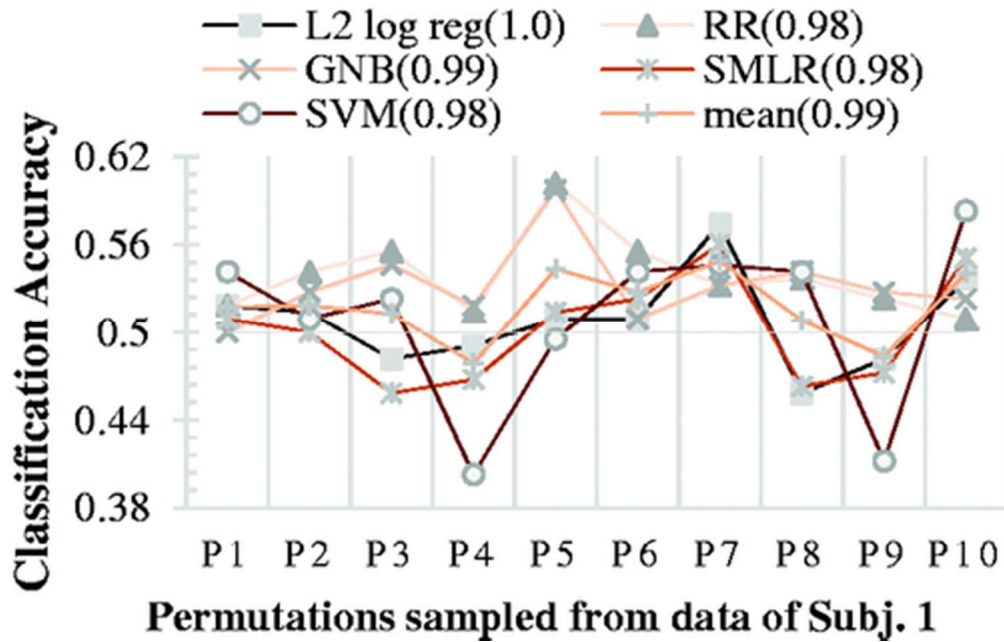
### 2.3 Using a group of classifiers

To fulfill the independence condition required by the CLT, the PD of the mean accuracy of groups of classifiers will be analyzed. The main idea is based on the fact that increasing the number of iterates may help the distribution of means to concord with the CLT. In other words, there is a chance that using more classifiers in a group would increase the number of iterates. Hence, it is more likely that PDs of mean accuracies of a group of classifiers will approximate to normal according to CLT. Adding more classifiers, however, may decrease the true best classification error rate compared to a superior classifier, like support vector machines (SVM) especially when the number of classifiers and sample size are small. Nonetheless, for cases with inferior classifiers, for example Gaussian Naïve Bayes, the classification error rate of using more classifiers may increase. Hence, it is difficult to claim that any given classifier is superior, this is because the classification error rate of using more classifiers would give the median accuracy, as illustrated in [Figure 2](#). That would be more useful than using a single (superior) classifier. Using a group of classifiers may release the user from the confusion of which classifier to choose and may prevent phishing the best classifier. Now, assume having a finite set of  $s$  different classifiers,  $C_1, C_2, \dots, C_s$ , and let  $C_{rs} = (G_1, G_2, \dots, G_L)$  be the set of the all  $L = (rs)$  groups of  $r$  classifiers,  $r=1, 2, \dots, s$ , which can be selected from  $s$  classifiers. The classification accuracy of two classifiers or more can be combined to yield the mean accuracy of the classifier group given by



$$a^i(Ck_1, Ck_2, \dots, Ck_r) = \frac{1}{r} \sum_{q=1}^r a^i(Ck_q) \quad (8)$$

where  $r$  denotes the number of classifiers in the group and  $k_q \in (1, 2, \dots, s)$  for all  $q=1, 2, \dots, r$ . This imitates that the mean is now considered over  $r \times J \times k_r \times J \times k$  values, and if these classification accuracy values are random, then the distribution (PD) will approach normality more than the sum shown in [equations \(6\)](#) and/or [\(7\)](#).



**Figure 2.** Classification accuracy of using each of the five classifiers working at the same permutation sample. Ten permutation samples are shown, P1, P2, ..., P10. The legend shows the name of each classifier and the classification accuracy for the non-permuted labels (in parentheses). These results were obtained from classifying fMRI patterns of face vs. house in Subject 1 without doing motion correction.

#### 2.4 Central limit theorem

Despite the fact that there are many variants of the CLT, one form that has been used most frequently (known as Lindeberg–Lévy theorem) states that the arithmetic mean of an adequately large number of iterates of independent and identically distributed random variables, each with a well-defined expected value and well-defined variance, will approximately be normally distributed. [17.18](#) Variants of the CLT have been established by

relaxing the assumptions of independence or identical distribution or both. This work will evaluate the fit of the normal model to the PD of classification accuracy that is obtained by an arithmetic mean of random variables. In this work, mean is depicted in [equation \(7\)](#) or [\(8\)](#) and each random variable corresponds to the accuracy value obtained from one of the cross-validation folds.

## 2.5 Permutation tests

Permutation testing is based on a simple idea that was proposed by Roland Fisher that he denoted as the exact test.<sup>7</sup> Let there be some effect, one just needs to permute/scramble the data then test if the effect is still there or not. In principle, the scrambling has to be repeated several times to cover all the possible combinations of permutations. This would be computationally impractical in large data; therefore, Monte Carlo technique has to be used. This means that the number of permutations has to be limited to fewer samples than the upper bound that the sample size allows. Since the distribution underlying the classification procedure is unknown, permutation tests have been used in many neuroimaging studies to estimate the classification' statistical significance.<sup>2,4,19</sup> The method has also been used to find the statistical significance in many other fields, for example in genetics.<sup>20</sup>

In permutation testing, the classifier's test error ([equation \(6\)](#)), or equivalently the classification accuracy ([equation \(7\)](#)), can be used as a test statistic for assessing the null hypothesis that the relationship between the data and the labels are not correlated. The alternative hypothesis is that the classifier arisen in the training step is associated with a small misclassification error rate (or, equivalently, a high classification accuracy rate). Mostly, the null hypothesis can be rejected if the test error is low enough compared to theoretical chance-level. The null distribution of the test error can be used to find the statistical significance of the classifier and one might think of it as estimating the probability, under the null hypothesis, of finding a result as extreme as, or more extreme than, the test error value actually observed (i.e. the  $p$ -value). Analog idea can be applied using the classification accuracy. Thus, the estimated  $p$ -value represents the statistical significance of the observed classification results that are likely to be obtained by random

chance. Performing the permutation is straightforward, let  $\mathbb{P}$  be the set of all possible label permutations denoted as operators, and let one sampled permutation operator be denoted as  $\mathcal{P}$ , i.e.  $\mathcal{P} \in \mathbb{P}$ . Each permutation operator works on the label set; thus, the permuted labels can be written as

$$Y'(i,j,1)=Y(i,j,1)(\mathcal{P}(1))Y'(i,j,1)=Y(i,j,1)(P(1)) \quad (9)$$

$$Y'(i,j,2)=Y(i,j,2)(\mathcal{P}(2))Y'(i,j,2)=Y(i,j,2)(P(2)) \quad (10)$$

where  $\mathcal{P}(1)$  is a permutation operator that works on the training set to give the scrambled training set, and  $\mathcal{P}(2)$  works on the testing set to give the scrambled testing set  $Y'(i,j,2)Y(i,j,2)$ . Subsequently, to find the classifier's permuted test error, one can use  $Y'(i,j,1)Y(i,j,1)$  and  $Y'(i,j,2)Y(i,j,2)$  instead of  $Y(i,j,1)Y(i,j,1)$  and  $Y(i,j,2)Y(i,j,2)$  (scrambling both the training set and the testing set). To perform scrambling only the testing set, one can use  $Y(i,j,1)Y(i,j,1)$  to train the classifier and then use  $Y'(i,j,2)Y(i,j,2)$  to test the classifier. Both of these permuting schemes will be tested in this work.

## 2.6 The probability of normality and the accordance to CLT

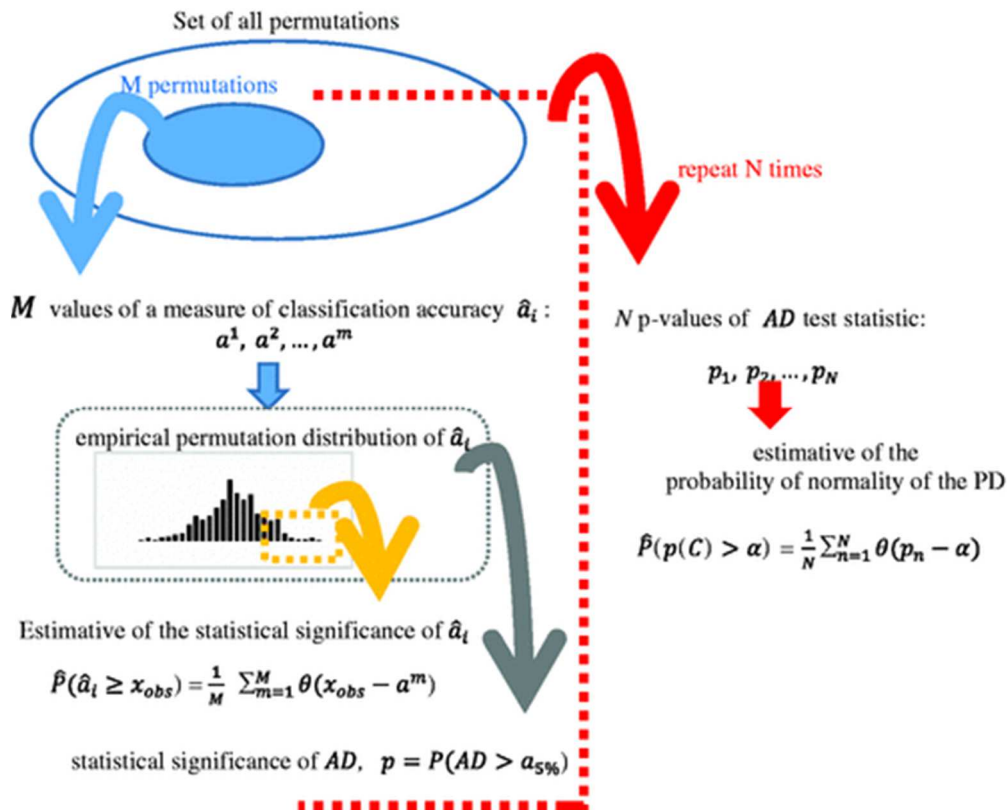
AD-test will be used for inspecting the fitting of the normality to the PD which has been empirically generated using  $M$  permutations (or iterations). The ADAD-test statistic yields a  $p$ -value that represents an estimate of how likely the PD is close to normal, let it be denoted as  $p\text{-value(PD)}$ , or just  $p$ . The decision will be made at a level of significance of  $\alpha\%$ . Therefore, the normality of the PD is rejected if  $p\text{-value(PD)} < \alpha$ .

Due to the stochastic character of  $p$ -values (or, equivalently, in generating PDs), using AD-test based on one generated PD may not be adequate to showing accordance or discordance to CLT, i.e., the normality of the PD generated by any finite set of permutations. Consequently, finite repetitions in generating a set of PDs for each experiment are vital to estimate the probability of normality. According to the law of large numbers, the proportion of occurrence of an event observed in a large number of trials will be close to the probability of the event and will tend to become closer as more trials are performed in the same conditions. Therefore, to find the estimate of the

probability of normality of the PD, we will calculate the proportion of the accordance of normality of a large number of generated PDs. There is a need to repeatedly find PDs up to  $N$  times, thus, yielding the set of observed  $p$  value (PD):  $(p_1, p_2, \dots, p_N)$   $p$  value (PD):  $(p_1, p_2, \dots, p_N)$  (see [Figure 1](#)). Now, for a set containing  $N$  PDs, one can estimate the probability of normality of a PD calculated via the permuted classification accuracy of classifier  $C$  as follows

$$P(p(C) > \alpha) = \frac{1}{N} \sum_{n=1}^N \theta(p_n - \alpha) \quad (11)$$

where  $\theta$  is a step-function (e.g.  $\theta(\beta) = 1, \text{ if } \beta \geq 0; 0 \text{ otherwise}$ ), and  $\alpha$  is set to 0.05, which is the critical value one may choose to reject the null hypothesis of the normality of the PD. In [equation \(11\)](#),  $p(C)$  represents the random  $p$ -value resulted from applying the AD-test on the permuted classification accuracy values determined by classifier  $C$ . This statistical measure would give a summarization of the random  $p$ -values and thus an estimate of the probability of normality.



**Figure 1.** Schematic diagram of the steps performed for estimating the normality of the PD of the classification accuracy of classifier  $C$  for Subject  $i$ . One value of the classification accuracy of classifier  $C$  for Subject  $i$  is obtained

using a permutation procedure. The permutation procedure is repeated  $M$  times for Subject  $i$ , generating one empirical PD of the classification accuracy. This PD is investigated whether it is well fitted to a normal probability law using the AD-test ( $a_{5\%}$  is the critical value of  $AD^*AD^*$  (formula 2) at a 5% significance level). Note that one generated PD can also be related to finding the classification significance ( $x_{obs}$  is the value of the measure of accuracy  $a^{i,a}$  observed for the classifier when the labels are not permuted), which is out of our interest in this paper. The whole process is repeated  $N$  times, resulting in a pool of  $Np$ -values which are aimed to evaluating the normality of the  $N$  generated PDs and thus, obtaining an estimative of the probability of normality of the PD of the classification accuracy of classifier  $C$  for Subject  $i$ .

AD: Anderson–Darling test; PD: permutation distribution.

To calculate the probability of normality for a group of classifiers  $G$ , one first needs to find the mean accuracy of the group using [equation \(8\)](#) for each generated permutation. Thus, a PD of the mean accuracy of the group  $G$  can be constructed from  $M$  permutations, and a  $p$ -value that estimates the normality level can be obtained by applying AD-test on the PD. To find the probability, we need to find the set of these  $p$ -value(s). Therefore, repeating the procedure  $M$  times, a set containing  $Np$ -value(s) will be generated. Now, using formula (11), the probability of normality of the PD of the mean accuracy of the group  $G$  can be estimated. Assume using  $L$  groups of classifiers and by performing these aforementioned steps for  $L$  groups leads to the following set of  $p$ -values  $p_1(G_1), p_2(G_1), \dots, p_N(G_1), \dots, p_1(G_L), p_2(G_L), \dots, p_N(G_L)$ . By the total probability law<sup>21</sup> and by [equation \(11\)](#), and for a PD estimated via the mean classification accuracy for the group  $C_{sr}$  of  $L$  classifiers, the formula that can be used to estimate the probability of normality is as follows

$$P(p(C_{sr}) > \alpha) = \sum_{l=1}^L P(p(G_l) > \alpha) = \frac{1}{L} \sum_{l=1}^L P(p(G_l) > \alpha) = \frac{1}{L} \sum_{l=1}^L \theta(p_n(G_l) - \alpha) = \sum_{l=1}^L P(G=G_l) P(p(G) > \alpha | G=G_l) = \sum_{l=1}^L P(G=G_l) P(p(G_l) > \alpha) = \frac{1}{L} \sum_{l=1}^L P(p(G_l) > \alpha)$$

where  $\theta$  is the step-function and  $\alpha$  is set to 0.05, or as desired. In [equation \(12\)](#),  $p(G)$  represents the random  $p$ -value regarding the application of the AD-test on the permuted mean classification accuracy values determined by the classifier group  $G$ .

## 2.7 The used classifiers

There are several classifiers that one can choose to perform the classification analysis. The following classifiers have been used in this work:

1. Artificial neural networks (NNs), from Netlab, [22](#) NNs with one hidden neuron were used, and using 50 epochs for training.
2. Support vector machines (SVM), from SVM light. [23,24](#)
3. Logistic regression (L2-LR), K class LR classifier, with optional regularization via L2 norm of weight vector(s). [25,26](#)
4. Gaussian naïve Bayes (GNB). [25,26](#)
5. Sparse multinomial logistic regression (SMLR). [25,26](#)
6. Ridge regression (RR), [25,26](#) this is a linear regression classifier that penalizes small weights (like weight regularization in back-propagation of NNs, as if doing an implicit feature selection). Moreover, this classifier has an analytic solution involving matrix inversion, thus it is deterministic, unlike NNs.

Due to the high classification performance they provide, these classifiers are often used in MVPA. GNB, on the other hand, has the lowest performance but maintains very efficient execution.

## 2.8 PDs and finding the probability of normality, an example

To have an idea on how to find the probability of normality using a group of classifiers, assume using the following three <sup>aa</sup> classifiers  $C_1, C_2, C_3$ . One can find the classification accuracy for the same permutation sample, i.e. the same relabeling set, for each of the following groups:

One-classifier group:  $C_{13} = ((C_1), (C_2), (C_3))$   $C_{31} = ((C_1), (C_2), (C_3))$ ,

Two-classifiers group:  $C$

$_{23} = ((C_1, C_2), (C_1, C_3), (C_2, C_3))$   $C_{32} = ((C_1, C_2), (C_1, C_3), (C_2, C_3))$ ,

Three-classifiers group:  $C_{33} = ((C_1, C_2, C_3))$

In the above example, using the One-classifier group will yield three PDs, three PDs for the Two-classifiers group, and one PD for the Three-classifiers group. This is because a One-classifier group has three separate accuracy values, one for  $C_1$ , another for  $C_2$ , and another for  $C_3$ . One can literally write these accuracies as:  $a^{(C_1)}$ ,  $a^{(C_2)}$ , and  $a^{(C_3)}$ , hence, a PD for each of the three. The Two-classifiers group will result in three accuracy values; first the mean accuracy of  $(C_1, C_2)$  that can be written as  $a^{(C_1, C_2)}$ , then, the mean accuracy of  $(C_1, C_3)$  so  $a^{(C_1, C_3)}$ , and finally the mean accuracy of  $(C_2, C_3)$  so  $a^{(C_2, C_3)}$ . For the Three-classifiers group only one overall accuracy value exists that is found by taking the mean of all the three classifiers, namely the mean accuracy of  $(C_1, C_2, C_3)$ , yielding  $a^{(C_1, C_2, C_3)}$ .

To elucidate the estimation of the probability, consider a group of two classifiers taken from three available classifiers,  $C_{23} = ((C_1, C_2), (C_1, C_3), (C_2, C_3)) = (G_1, G_2, G_3)$ , to give three distributions  $(PD_1, PD_2, PD_3)$ , such that  $PD_1$ ,  $PD_2$ , and  $PD_3$  are found from the mean classification accuracy,  $a^{(C_1, C_2)}$ ,  $a^{(C_1, C_3)}$ , and  $a^{(C_2, C_3)}$ , respectively. Then, using AD-test gives the following  $p$ -values:  $p_1(G_1), p_2(G_1), \dots, p_N(G_1), p_1(G_2), \dots, p_N(G_2), p_1(G_3), p_2(G_3), \dots, p_N(G_3)$ , and the probability of normality can then be found using [equation \(12\)](#).

To give another example, [Figure 2](#) illustrates individual classification accuracies using five classifiers each working at the same permutation sample as well as the mean classification accuracy given by the five. Calculating the coefficient of variation for each permutation resulted in values between 0.03 and 0.098. These (relatively) low values suggest that the mean classification accuracy may be an adequate global measure for quantifying the joint performance of classifier groups.



Regardless of permutation testing and the CLT, using the mean accuracy of a group of classifiers can be considered as a measure since it is estimated as a combination of several classifiers, and thus, one could move apart from peeking the result from a supreme classifier.

## 2.9 The dataset

Publicly available data that were collected while subjects were viewing eight object categories<sup>1,29</sup> have been used in this work. This dataset is well designed and contains adequate runs and brain volumes and thus has become one of the major data used in several brain analysis works. In the dataset, hemodynamic changes (blood oxygenation level-dependent signals) were measured in each subject with a gradient echo planar imaging on a GE 3 T scanner with a repetition time (TR) 2.5 s, yielding 40 slices of resolution with  $64 \times 64$  in each volume (40 3.5 mm-thick sagittal images). High resolution T1-weighted spoiled-gradient recall images were obtained for each subject, with 124 slices of resolution ( $256 \times 256$ ) (124 1.2 mm-thick sagittal images).

The fMRI dataset contains responses of six human subjects that were collected using an MRI machine while they were performing a one-back repetition detection task and visualizing each of the eight objects/stimuli. In addition to rest condition, stimuli were gray-scale images of faces, houses, cats, bottles, scissors, shoes, chairs, and scrambled pictures (xpic). Twelve time series (i.e. runs) were obtained from each subject responses. Each time series began and ended with 12 s of rest and contained eight stimulus blocks of 24 s duration, one for each class, separated by 12 s intervals of rest. Each stimulus was presented for 0.5 s with an interstimulus interval of 1.5 s. Stimuli for each meaningful class contained four images, and for each class, there were 12 exemplars of meaningful stimuli containing pictures of the same face or objects photographed from different angles. In the end, there were 12 runs for each subject, containing 1452 brain volumes including rest, see<sup>1</sup> and<sup>29</sup> for more details. It is worth to mention that Subject 5 has one missing run in the server of the database<sup>29</sup> and therefore was dropped from the analysis.

## 3 Experimental results

### 3.1 Implementation issues



- By eliminating rest brain volumes, every subject in the data ended encompassing 864 brain volumes, with only eight classes. To use a binary classifier, one can pick data representing a pair-of-stimuli: bottle vs. shoe, bottle vs. scissor, chair vs. scissor, face vs. house, and face vs. xpac, etc. Therefore, there will be 216 volumes for each pair-of-stimuli (e.g. bottle vs. shoe) in the form of 12 Runs x 18 volumes/Run. Furthermore, each run has two blocks such that one block has nine volumes of the first category and the other has nine volumes for the second category. The motive behind selecting these pair-of-stimuli was to obligate high classification power, house and face, and low classification power, shoe, bottle, and scissor. In fact, these pair-of-stimuli(s) have been selected after performing a few exploratory analyses on the dataset. The use of this particular dataset is of particular interest to this work, not only due to the vast number of conditions and the adequate number of trials it has, but due to the high computational complexity that permutation testing and normality analysis demands.
- The analysis will be performed with and without motion correction (aka fMRI alignment). This might be a useful prototype to study the effect of the native space with less possible interpolation effect and then comparing these results to the ones with motion correction. SPM8<sup>27</sup> has been used to perform motion correction. With or without motion correction, all of the fMRI signals will be de-trended and z-scored.
- Labels will be shifted by 5 s (2TRs) to compensate for the hemodynamic lag (this is a common practice in MVPA experiments and studies<sup>2</sup>). Since the BOLD's haemodynamic response has a delay with respect to the stimulus onset, it is common practice to correct the regressors' delay in each condition. This is usually done by convolving the regressors with a set of basis functions chosen to model the haemodynamic response. The haemodynamic response, however, will set the onset/offset of the regressors, and therefore using a shift of 2TRs would be sufficient in most cases. The correctness of this shifting strategy is supported by the literature,<sup>2,4,6,28</sup> as well as the high classification accuracy we got in the non-permuted/true regressors with error rates close to 99% in some classifiers, like LR (see [Figure 2](#)). Furthermore, our tests have shown that TR shifting and haemodynamic model convolution yield the same

classification accuracy. Nevertheless, correcting the haemodynamic lag through shifting the regressors by a value of 2 is not a rule of thumb to be used with other datasets. This is supported by the fact that the haemodynamic lag is exceedingly known to vary across the brain, and further investigation may lead to better results in the future.


- Instead of permuting brain-volumes, only the labels are permuted. This technique has exactly the same effect and was used in previous studies.<sup>12</sup>
- Uniformly distributed pseudorandom numbers will be generated; these numbers will then be used to scramble the labels. To preserve independence among runs, only labels within runs will be scrambled.
- Permutation (scrambling) schemes: Although most studies in the literature scramble the whole data labels (e.g. the training and the testing set), there has been a debate whether one might only need to scramble the testing set. The idea behind permutation testing is breaking the correlation, if exists, between patterns of brain activity and the labels. Permuting only the testing set has the advantage of efficient implementing since the classifier will only be trained once, and then, tested over and over using the same scrambled testing set. Scrambling the training set and the testing set implies that training the classifier and testing it should be performed over and over for all the scrambled training and testing sets, respectively. Because no definite answer exists to this dilemma, both approaches, scramble-train-test and scramble-test, will be endeavored in this work. More details on permutation schemes can be found in Etzel and Braver.<sup>28</sup>
- Each classifier runs in a leave one run out manner. Consequently, for the 12 folds cross-validation scheme, 11 runs will be used for training the classifier and one run will be preserved for testing it.
- As demonstrated earlier, generating only one PD is not sufficient to estimate the probability of normality. Thus, each classification and the respective permutation testing have to be repeated  $N$  times to find  $N$  PDs. If, however,  $M$  permutations are used to generate each PD, then a total of  $M \times N \times N$  operations will be needed, where each operation designates training a classifier then testing it. For example, using  $M=1000$ ,  $N=1000$  indicates the need to run  $1E6$

classifier instances (training and testing) and that is a computationally demanding task. To solve this problem, we opted to generate a pool of permutations of size  $v$  (one might think of it as a PD of size  $v$ ).

Afterwards, a total of  $N$  PDs each of size  $M$  (provided that  $M < v < M < v$ ) will be randomly sampled from this pool. In this work, the permutations pool has a size  $v=10,000$ .

- Region-of-Interest (ROI): Two ROIs that came along with the fMRI dataset have been used in the analyses. These ROIs are as follows: VT4 (called mask\_vt4 in the dataset) is a mask that represents voxels in the ventral temporal cortex, H8 (called mask8\_house\_vt) is a mask that contains a set of voxels maximally responsive to house category. The number of voxels each ROI has is shown in [Table 1](#).
- Details on the used classifiers were provided in section 2.7.

Click to view table



**Table 1.** The number of voxels in each ROI.

**Table 1.** The number of voxels in each ROI.

Mask abbreviation	Mask name	Subject 1	Subject 2	Subject 3	Subject 4	Subject 6
VT4	mask4_vt	577	464	307	675	348
H8	mask8_house_vt	163	48	117	148	115

Note: These masks in functional native space were provided by the authors, [1.29](#) “VT” refers to a mask in the ventral temporal, house\_vt is GLM contrast-based localizer maximally responsive to house stimulus.

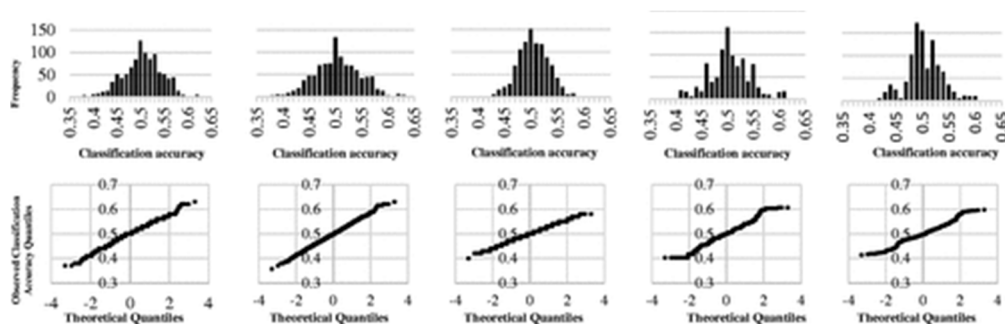
ROI: region-of-interest.

[View larger version](#)

### 3.2 PD and normality

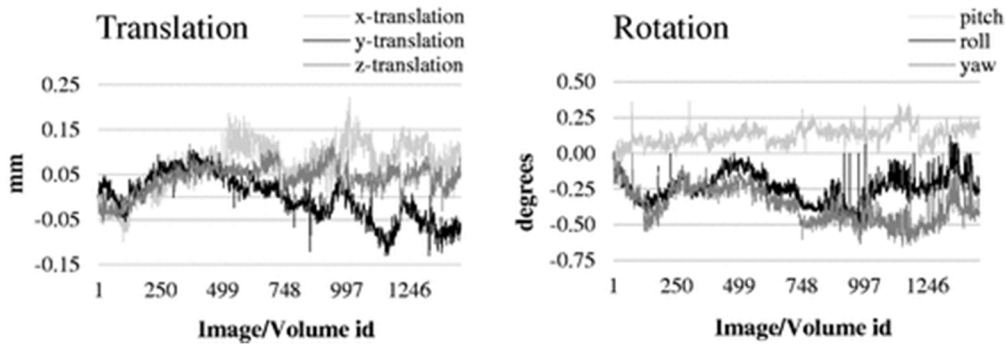
Before demonstrating the detailed results, it is useful to visualize cases where the normal distribution fits the PDs or not, thus exposing if these PDs behave in accord with CLT. Three PDs, which are estimated from the

classification analysis of fMRI data containing patterns of brain activity of face vs. house of Subject 1, are depicted in [Figure 3 \(a\) to \(c\)](#). Two PDs for the same experiment setting after performing motion correction to the MRI volumes of Subject 1 are also shown in the same [Figure 3\(d\) and \(e\)](#). To have an idea about the amount of motion in the subject's fMRI data, motion correction parameters are shown in [Figure 4](#). Among those five cases shown in [Figure 3](#), there are two cases where the normal distribution seems to be adequate to the PD of the classification accuracy, and all the points are closer to a line as the QQ-plots show. Nevertheless, [Figure 3](#) suggests the existence of multimodal distributions; or eventually, mixture of normal distributions fitted to the PDs, this topic is, however, beyond the scope of this paper. What is importantly shown in these figures is that a PD for the same experimental setting, but without motion correction to the fMRI data, may or, more likely, may not behave in accord CLT. For motion-corrected subjects, however, the probability of having a normal PD approaches 0, even if one opts for using a group of classifiers.



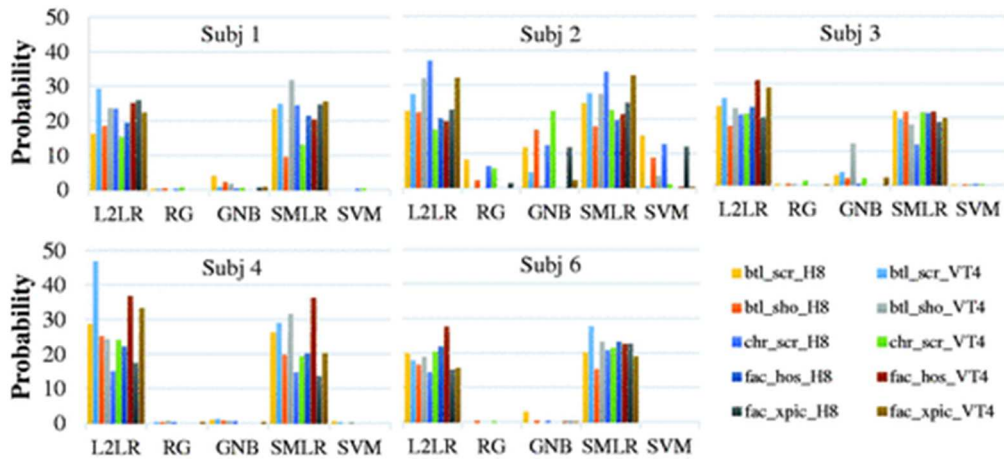
**Figure 3.** Permutation distributions constructed using the mean classification accuracy of L2 LR (a) a distribution that is not normal, with AD-test  $p$ -value = 0.004, b) A permutation distribution that approximates to normal with AD-test  $p$ -value = 0.11. For both a and b, L2 LR classifier has been used and the probability of normality was 25%. (c) This permutation distribution gave AD-test  $p$ -value = 0.7469, and thus approximates to a normal distribution, where the classification accuracy is the mean of using five-classifiers and the probability of normality for this case was 57%, (d) Permutation distribution constructed using the mean classification accuracy of L2 LR, normality testing via AD-test resulted in  $p$ -value = 6.9E-9 and the estimated probability of normality was zero, (e) Permutation distribution obtained using mean classification accuracy of five classifiers with AD-test  $p$ -value = 1.17E-13, and the estimated probability of

normality was also zero. The classification was performed using patterns of face vs. house extracted from VT4 (the ventral temporal cortex) of Subject 1 and the size of each permutation distribution is 1000. Motion correction was used only on the data for d and e. A corresponding QQ plot is shown below each histogram.



**Figure 4.** Motion correction parameters of Subject 1, performed using SPM 8.

For more in-depth analysis, the probability of normality for PDs obtained by the classification accuracy per each classifier has been investigated. Such probability, which was estimated using [equation \(11\)](#) for each five pair-of-stimuli at two different ROIs for each of the five subjects, is exhibited for five classifiers in [Figure 5](#). At a first glance, the results seem quite stimulating due to some classifiers (like L2LR and SMLR) that reveal a greater tendency to not reject normal fitting to the PDs. This tendency, however, is relatively small and far from the expected 95% confidence level. Indeed, under the null hypothesis of normality, the  $p$ -value of AD-test will be uniformly distributed and, consequently, the estimated probability (11) should approximate close to 95% for  $\alpha = 0.05$ . In the results shown in [Figure 5](#), the probability was observed to vary between 9% (Subject 1) and 44% (Subject 4) for L2LR and SMLR classifiers, 22% (Subject 2) for GNB, SVM and RG classifiers, and between 0% (Subjects 3 and 6). In terms of average probability over all stimuli, the estimated probabilities were (23%, 1%, 3%, 22%, 1%) for L2LR, RR, GNB, SMLR, and SVM respectively. Due to these low (averaged) confidence levels, it becomes difficult to foresee a classifier type giving a PD that concurs with CLT. Moreover, the estimated probabilities vary from one classifier to another, L2LR gave a low probability value of accordance to CLT (23%) and SVM gave extremely low probability value (1%), and that is even harder to construe.



**Figure 5.** Probability of normality of the permutation distribution obtained via the classification accuracy for each of the five classifiers considered in this work. Each binary classifier worked on data extracted either from H8 or VT4 brain ROIs for the five pair-of-stimuli. Invisible bars designate zero values.

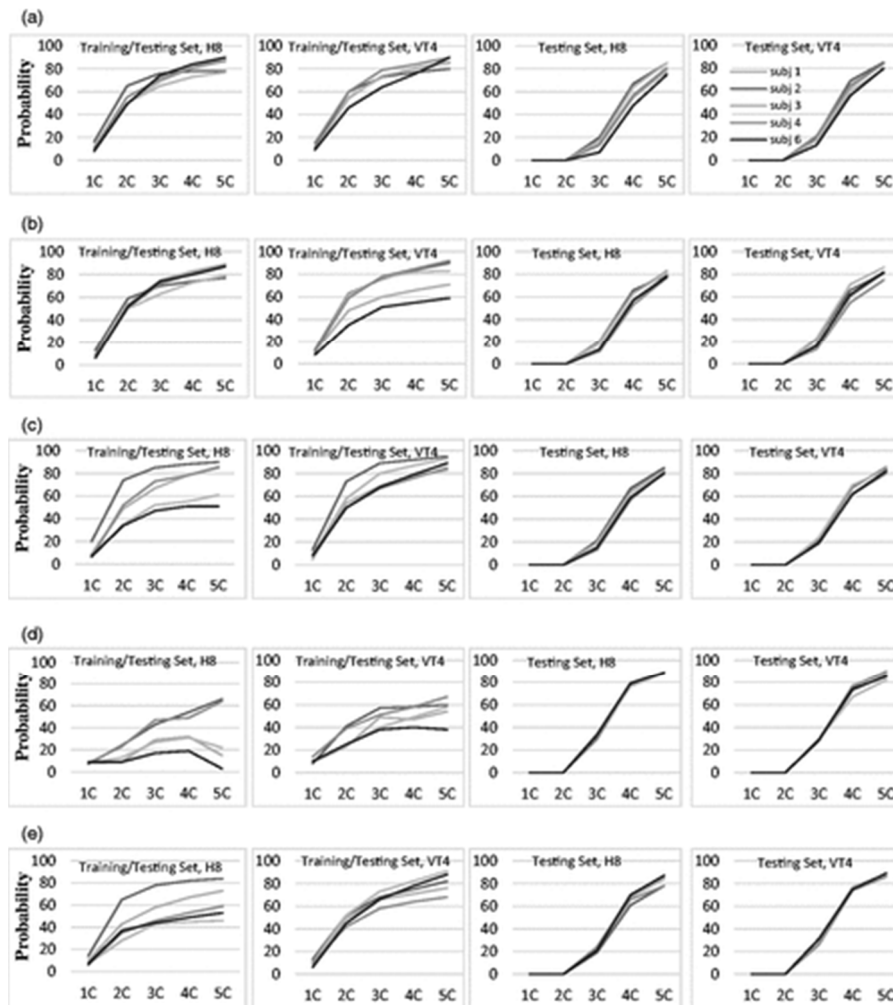
Assessing the existence of a significant correlation between classification accuracy and probability of normality is crucial to this study. To do this, conditions yielding superior classification accuracy values have been scrutinized separately from the others yielding inferior classification accuracy. Thus, mustering utterly the results obtained via the stimuli that gave superior accuracy, e.g., face vs. house, face vs. xpic in both H8 and VT4 regions from all subjects, a correlation value equal to  $-0.06$  ( $p$ -value=0.8) was found (mean  $\pm$  standard deviation of the classification accuracy over all these stimuli and subjects was  $0.91 \pm 0.10$ ). However, mustering utterly the results obtained via the stimuli that gave inferior classification accuracy, e.g. bottle vs. scissor, bottle vs. shoe, chair vs. scissor in both H8 and VT4 regions from all subjects, a correlation value equal to  $0.08$  ( $p$ -value = 0.65) was found (mean  $\pm$  standard deviation of the classification accuracy over all these stimuli and subjects was  $0.67 \pm 0.09$ ). Therefore, both conditions revealed that the probability of normality is not significantly associated with the classification accuracy.

### 3.3 PD via the mean accuracy of a group of classifiers

The purpose of the approach presented in this section is to investigate the probability of accordance to CLT for PDs generated using averages of classification accuracies of sets of  $r_n$  classifiers ( $r=1,2,3,4,5$ ;  $n=1,2,3,4,5$ ),

selected from five different types of classifiers (L2LR, RR, GNB, SMLR, and SVM) to work as test statistic for permutation testing. For the five subjects and the five pair-of-stimuli that were selected for this study, we performed several experiments using two scrambling schemes: scramble-train-test and scramble-test. For this experimental setting, we focused on data without motion correction and we performed de-trending, to remove linear trends due to scanner drifts, and z-scoring. It is convenient to divulge that the average of classification accuracies for each set of  $r$  classifiers has been calculated using formula (8) and this has been done for each of the two ROIs and for each subject. The probabilities of normality of the PD of the mean classification accuracy by a group of classifiers were estimated using formula (12) and are depicted in [Figure 6](#). The results shown in [Figure 6](#) indicate clearly that a PD obtained via the mean classification accuracy of one classifier does not approximate to normal, and that the more classifiers used, the more likely that the PDs will behave in accord with the CLT. Using a group of classifiers, however, the accordance of PDs to CLT was not eminent as it can be seen in [Figure 6\(d\)](#) (training/testing set, H8, face vs. house). In fact, [Figure 6\(d\)](#) reveals that using a group of five-classifiers, the probability of having a normal distribution decreased in three subjects out of five, having even achieved a value close to 3% for Subject 6. This small probability is lower than all the other pair-of-stimuli over all subjects as well. This is an interesting issue since it occurs when classifying face vs. house in H8 (note that H8 ROI is maximally responsive to houses) which has a high classification accuracy 96% in the non-permuted data. Nevertheless, reason behind this drop in probability of normality for this case is not clear hitherto.





**Figure 6.** Probability of normality (the y-axis) of various permutation distributions for each of the five subjects, versus groups of classifiers (1C, 2C, 3C, 4C, 5C) that denote using (one-classifier, two-classifiers, three-classifiers, four-classifiers, five-classifiers). All these analyses were performed without head motion correction to fMRI data. Training/testing set refers to scrambling both the training and the testing sets, and testing set refers to scrambling only the testing set. H8 and VT4 are the two regions-of-interest used.

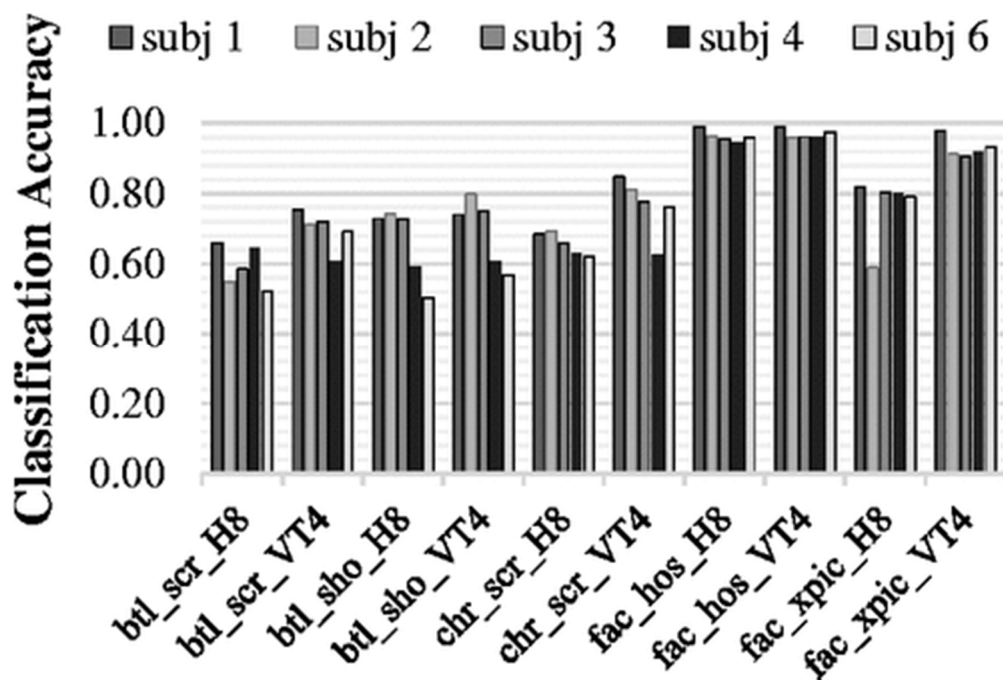
Another interesting issue still stands out in [Figure 6](#) for the case of scramble-train-test vs. scramble-test, which is; independently of the pair-of-stimuli and subject, a more stable probability pattern among the pair-of-stimuli and subjects is detected in scrambling-test scheme. Undeniably, calculating the standard deviations of these probabilities of normality for each group of classifiers and scrambling scheme, values of volatility were obtained when scrambling both training and testing sets than those obtained when scrambling only testing sets. These results seem crucial and suggest that



permutation schemes can lead to different PDs and consequently, different assessment of the statistical significance of classification accuracy.

### 3.4 The classification accuracy and the normality

It is imperative to investigate whether this accordance of PDs with CLT is related to the classification accuracy or not. [Figure 7](#) illustrates the mean classification accuracy (over the five classifiers) for each pair-of-stimuli at the two ROIs. Comparing [figures 6](#) and [7](#) one can conclude that approaching 95% of probability of normality is not correlated to the value of classification accuracy, i.e. whether the classification accuracy is high or low. For instance, one case with low classification result, as in the bottle vs. scissor situation at H8 in Subject 6 which is close to theoretical chance-level with value equal to 0.52, and another case with high classification accuracy, as in face vs. house situation at VT4 in Subject 6 with value equal to 0.98, are associated with high estimates of the probability of having a normal distribution, which are given by 90% and 85%, respectively. Analytically, using the training-testing permutation scheme, no statistically significant correlations between the classification accuracy and the probability of normality according of PDs were found (values of Pearson correlations:  $-0.47$ ,  $-0.17$ ,  $-0.27$ ,  $-0.24$ ,  $-0.38$ , with  $p$ -values: 0.17, 0.63, 0.45, 0.51, 0.28, for Subjects 1, 2, 3, 4, and 6, respectively). A similar conclusion was obtained for permuting only the testing scheme. The used labels in [Figure 7](#) mimic the used stimuli and ROIs, e.g. btl\_scr\_H8 means classifying patterns of activity at ROI H8 for bottle vs. scissor.

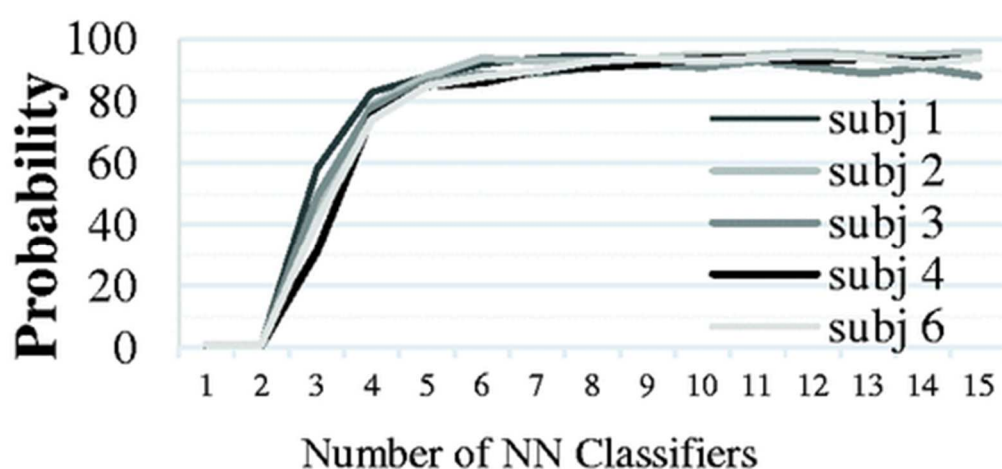


**Figure 7.** Mean classification accuracy over the five classifiers for the used pair-of-stimuli at VT4 and H8 ROIs. (Shows the non-permuted classification accuracy).

3.5 Will the probability of normality approach 95% confidence level when using several artificial neural network classifiers?

The previous analysis with groups of classifiers, covering from one up to five classifiers, compassed a (Mean±StdDev) summary value of  $0.78 \pm 0.15$ .  $0.78 \pm 0.15$  for the probability of normality using Five-classifiers group. It would be interesting to investigate how many NNs classifiers would assist attaining the confidence level of probability of normality having 95% or more. NN classifiers could work in linear and/or nonlinear mode and they have some stochastic degree to ensure a different classifier instance at each execution, the learning is non-deterministic. Basically, NNs are randomly initialized prior to learning with back-propagation and other weight optimization procedures. Therefore, there will not be a unique optimal solution and that means every trained NN will depend on the initial weights and other stochastic processes during learning. Using such classifiers will facilitate one incrementing the number of classifiers used as needed. Up to 15 NNs have been used and the confidence level of accordance to normal law approached a value close to 95% when using only the testing set

permutation. [Figure 8](#) shows the results of bottle vs. shoe for patterns extracted from H8, other pair-of-stimuli and regions have a similar behavior. These experiments have been repeated after using testing-training permutation scheme and surprisingly, the probability of accordance to normal law under this scrambling setting was zero, regardless of the number of NN classifiers used. This indicates that, although different NN classifiers have different random initial-weights and stochastic non-deterministic learning algorithms, they are not adequate for PDs to accord to CLT, and that the mechanism of the permutation scheme can affect the PD of the classification errors/accuracies.



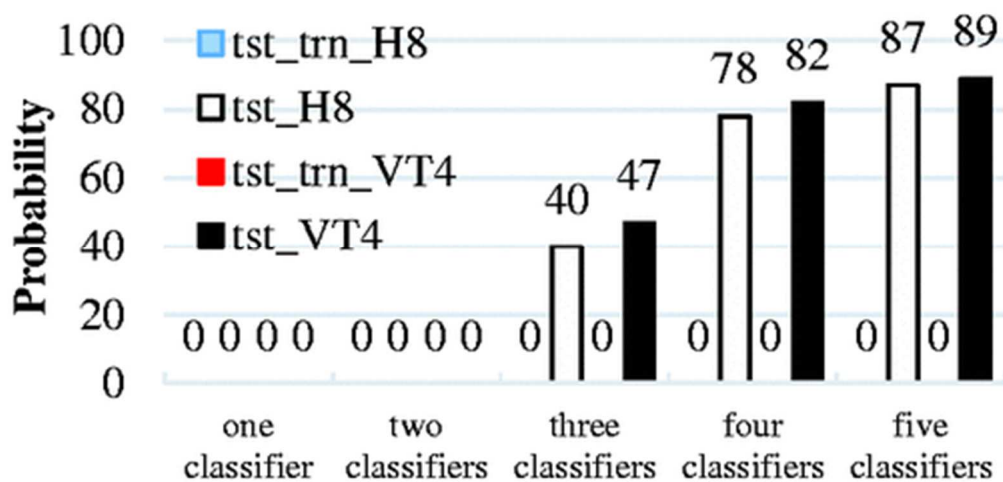
**Figure 8.** Probability of normality versus number of neural networks used to find the classification accuracy. The effect of increasing the number of neural network classifiers to approach the confidence-level in stating accordance with central limit theorem is obvious. Only the testing set has been scrambled here, in classifying bottle vs. shoe brain responses that were extracted from H8 ROI. Other pair-of-stimuli gave similar curves.

### 3.6 The effect of motion correction on the normality

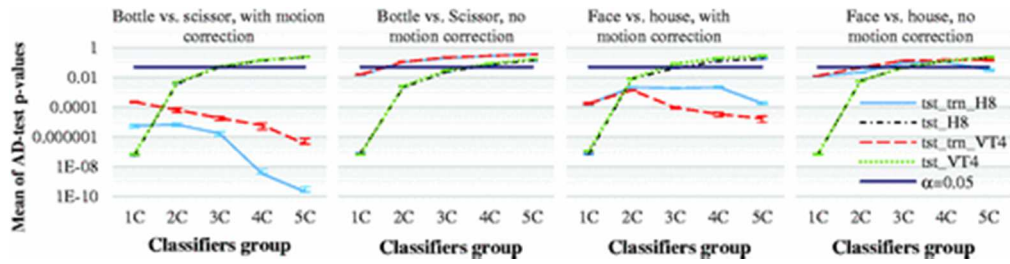
All the previous analyses were performed using the data in the native space and the classification accuracies were high and reached 0.99 in some cases, see [Figure 7](#). Motion correction (which is also known as alignment, registration) is a default preprocessing procedure that is usually performed prior to performing fMRI analysis. The purpose of motion correction is to reduce the distortion in fMRI signal due to subject's head movement during the fMRI acquisition session. To investigate the effect of preprocessing on

the probability of PD to behave in accord with CLT, within- and between-session motion corrections to Subject 1 have been performed. Using training/testing sets scrambling, the obtained results for bottle vs. scissor were surprising that after performing motion correction, the probability of normality was zero for every group of classifiers. Using testing set scrambling, however, the obtained probability reached 95% confidence level and is similar to that of [Figure 7](#) where no motion correction has been performed, especially for a group of three classifiers or more.

We are abutting an astounding result in the case of scrambling both the training and the testing sets that resulted in total non-accordance with CLT when motion correction has been performed to the data. To evaluate the precision of the decision provided by AD-test  $p$ -values, we calculated their mean and standard deviation. [Figure 10](#) illustrates the mean of the AD-test  $p$ -values for PDs obtained from classifying bottle vs. scissor and face vs. house patterns of activity extracted from Subject 1 when motion correction has or has not been applied.



**Figure 9.** The estimated probability of normality vs. the number of classifiers used for classifying bottle vs. scissor of motion-corrected subject 1. Data labels have been placed above each bar to highlight the missing values when the probability value being zero. Analogous results were obtained for the other pair-of-stimuli. *tst\_trn\_H8*: scrambling the testing and the training sets for data extracted from ROI H8. *tst\_H8*: scrambling only the testing set for data extracted from ROI H8. *tst\_VT4* is the other ROI.

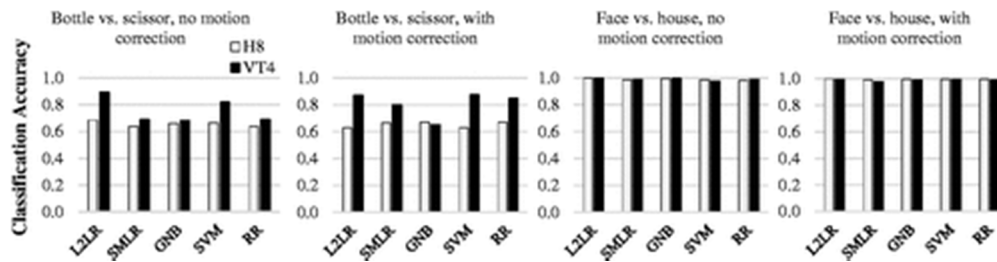


**Figure 10.** Mean of Anderson–Darling test  $p$ -value(s) for permutation distributions obtained from classifying data from Subject 1 when motion correction has or has not been applied to the data. Cases for five groups of classifiers are shown; (1C, 2C, 3C, 4C, 5C) denote using (one-classifier, two-classifiers, three-classifiers, four-classifiers, five-classifiers). Error bars (95% confidence level) are sometimes invisible due to low values.

It is clear from [Figure 10](#) that increasing the number of classifiers pushes down the mean of AD-test  $p$ -values when the training set and the testing set are both scrambled in data that have been corrected for motion. This indicates that approaching a normal distribution will not be possible even if more classifiers are used, and the PD will not limit to normal as expected by CLT. Another interesting issue in this case is that the mean of AD-test  $p$ -values for the two investigated ROIs (H8 and VT4) was different regardless of the used number of classifiers (unlike the case when only testing set scrambling is used). In the other case of scrambling only the testing set, the mean of AD-test  $p$ -values continues to escalate which indicates that the probability of normality increases with incrementing the number of classifiers. In conclusion, these two figures indicate that regardless of the classification accuracy, the non-accordance of PDs to CLT cannot be solved by increasing the number of classifiers when both the training and the testing sets are scrambled for motion-corrected subjects.

To have an idea on the classification accuracy with or without motion correction, [Figure 11](#) exhibits the classification accuracy for Subject 1 with and without performing motion correction. The classification accuracy for bottle vs. scissor without motion correction was 0.75, and after performing motion correction was 0.80, but for face vs. house was 0.99 before and after performing motion correction. This indicates that motion correction has a

marginal effect on the value of classification accuracy, which concurs with earlier findings.<sup>30</sup>



**Figure 11.** Classification accuracy vs. used classifier, at ROIs H8 and VT4 of Subject 1 with and without performing motion correction. These graphs show the (true) non-permuted classification accuracy.

The effect of motion correction on the normality of PDs was noticeable in [Figure 10](#); however, as with previous findings, both [Figures 10](#) and [11](#) indicate that the probability of normality of a PD has no relation to the true (non-permuted) classification accuracy, i.e. whether the classification accuracy is high or low. It is apparent that the mean classification accuracy of classifying patterns of face vs. house is higher than the mean classification accuracy of classifying bottle vs. scissor. Therefore, the results illustrated in [Figures 9](#) to [11](#) indicate that normality testing of PDs may be used to detect the degree of dependence/independence among fMRI runs regardless of the value of classification accuracy.

#### 4 Discussion and conclusions

The main yield of this work is that PDs of fMRI data classification do not approximate to normal. As a result, it is potential to argue that this non-abundance to the CLT indicates resilient dependence in the folds of the cross-validation scheme that is usually used to find the classifier performance. This dependence may swindle unreliable classification significance either due to the overlapping of training samples or the dependency among different magnetic resonance imaging sessions. If that is the case, an alternative to this classification-error measurement dilemma is a classification approach and/or a performance metric that makes the PD approximates normality. One way to achieve this is by using a set of classifiers and then taking the mean accuracy over these classifiers. This approach has been tested in this work

using five different classifiers, which included LR, GNB and three more, and has led to normal PDs with greater confidence levels. Besides these five classifiers, what other classifiers should be included in the test?

Hypothetically, NNs could be used to generate a set of classifiers, since each classifier instance has its own random initial weights, and possibly, different weights optimization approach and/or learning strategies. Unlike the five different classifiers used, the performed analyses have shown that NNs classifiers are highly dependent. This work has only considered binary classifiers using within-subject analysis of visual stimulation data. Our findings are strongly in accordance to empirical results shown in Stelzer et al.<sup>6</sup> which reveal that the higher the correlation of the cross-validation folds, the larger the deviation of the classification accuracy from the exact binomial distribution expected under the assumption of independence between cross validation folds and thus, to the approximated normal distribution expectable from binomial distribution when the numbers of trial is large.

Results using single classifier analyses on subjects without motion correction were appealing. Among the six classifiers studied in this work, LR classifier and SMLR classifier gave the highest probability of accordance to CLT (~ 20%), which means these two classifiers have a strong power to detect across fMRI runs dependence. Using the same subjects that were not corrected for motion, and using the other four classifiers, namely: NNs, RR, GNB, and SVMs, there were no obvious across fMRI runs dependency (the probability of normality was less than 5%). Nevertheless, for each of the used scrambling schemes—scramble the training and the testing set, and scramble only the testing set—a performance metric based on the mean classification accuracy of a group of five (or more) classifiers resulted in PDs that behave in accord to CLT with high confidence level, and the probability of normality reached ~90%.

After performing motion correction for the subjects used in this study, the probability of accordance to CLT for each (single) classifier was 0%, for both scrambling schemes. This is compared to 1% to 49% for results of subjects without motion correction. Consequently, motion correction profoundly affected the shape of the PD, even though it did not increase the

classification accuracy by a momentous margin. This points out that motion correction may increase the across fMRI runs dependency, i.e. intra and inter stimulus pattern dependency. Moreover, using a performance metric based on the mean classification accuracy of a group of classifiers did not improve the accordance to CLT for scrambling both the training and the testing sets of motion-corrected data (which is the usual way permutation testing has been performed in the literature). In fact, using a group of classifiers, the shape unexpectedly departed away from normality and the fact that the average(s) of AD-test  $p$ -values were declining as one increases the number of classifiers has been verified. Unpredictably, scrambling both the training and the testing set did not provoke PDs to approximate to normal law similar to the case where subjects data have not undergone motion correction. Nonetheless, scrambling only the testing set spoke differently (using a group of classifiers again) and the probability of accordance to CLT was nearly similar to that when motion correction procedure was not performed on the subjects. It is quite difficult at this stage to interpret this discrepancy between the two scrambling schemes. It may designate, however, that scrambling both the training and the testing sets have higher inter and intra stimulus dependency. Hence, if the normality of PD is an essential brick to find the significance of the classification, then, one should opt to scramble only the testing set to obtain the  $p$ -values related to the classification significance. Nonetheless, the option of not performing motion correction to fMRI data is only related the multivoxel classification analysis and this has no connection to the decisive need to perform motion correction in univariate/univoxel analysis of fMRI data.

For subjects in the native space, i.e. without motion correction and without geometrical normalization, there is a weak correlation between the probability of normality and the classification of categories when both the training and the testing sets are scrambled. To elucidate on this, it is apparent that the mean probability of normality of PDs obtained from classifying face vs. house is slightly lower than that of classifying bottle vs. shoe (one classifier analyses). One probable cause is that patterns of activity in blocks containing stimuli that may result high classification accuracy, e.g. faces vs. houses, have more across run dependency than blocks containing lower classification accuracy, e.g. bottles vs. shoes. It is difficult to rationalize



this phenomenon in this work and further studies are needed with the anticipation of using it as a diagnostic tool in the near future. It must be mentioned that the accordance of PDs to CLT has not been explored in this work for multi-class problems and other fMRI paradigms, and that goes to inter and intra stimuli dependency across runs in across subjects' studies too.

It is essential to mention that the analyses that have been done in this work have taken non-trivial execution time. The temptation of scrambling only the testing set is quite practical for the reason that it is much faster than the one where both the training and the testing sets are scrambled. In fact, scrambling both sets as well as the relevant classification analyses could be performed efficiently using parallel computing via GPU. Thus, implementing PD normality test as a diagnostic tool to detect the independence between different brain regions may be feasible.

## Acknowledgment

We thank the reviewers for their comments and suggestions.

## Funding

This work was supported by Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology, FCT) and Quadro de Referência Estratégico Nacional (national strategic reference framework, QREN) under the Mais Centro initiative: UID/4539/2013-CNC.IBILI, CENTRO-07-ST24-FEDER-00205, (FCT, PEst/C/SAU/3282/2013), COMPETE FCOMP-01-0124-FEDER-022690, CIDMA (FCT, UID/MAT/04106/2013) (FCT, SFRH/BD/69735/2010).

## Notes

aOne may choose to use the average accuracy of the number of classifiers at hand, especially if using a large number of homogenous classifiers, and not the approach shown above that has more in-depth analysis.

## References

1. Haxby, JV Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 2001; 293: 2425–2430. [Google Scholar](#), [Crossref](#), [Medline](#), [ISI](#)  
[Open URL](#)
2. Etzel, JA, Gazzola, V, Keysers, C. An introduction to anatomical ROI-based fMRI classification analysis. *Brain Res* 2009; 1282: 114–125. [Google Scholar](#), [Crossref](#), [Medline](#)  
[Open URL](#)
3. Carlson, TA, Schrater, P, He, S. Patterns of activity in the categorical representations of objects. *J Cognitive Neurosci* 2003; 15: 704–717. [Google Scholar](#), [Crossref](#), [Medline](#)  
[Open URL](#)
4. Pereira, F, Mitchell, T, Botvinick, M. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 2009; 45: S199–S209. [Google Scholar](#), [Crossref](#), [Medline](#)  
[Open URL](#)
5. Friston, K. Sample size and the fallacies of classical inference. *Neuroimage* 2013; 81: 503–504. [Google Scholar](#), [Crossref](#), [Medline](#) [Open URL](#)
6. Stelzer, J, Chen, Y, Turner, R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage* 2013; 65: 69–82. [Google Scholar](#), [Crossref](#), [Medline](#)  
[Open URL](#)
7. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1954. [Google Scholar](#)
8. Nichols, TE, Holmes, AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 2002; 15: 1–25. [Google Scholar](#), [Crossref](#), [Medline](#), [ISI](#) [Open URL](#)
9. Golland, P Permutation tests for classification. *Learn Theory Proc* 2005; 3559: 501–515. [Google Scholar](#), [Crossref](#) [Open URL](#)

10. Noirhomme, Q Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *Neuroimage Clin* 2014; 4: 687–694. [Google Scholar](#), [Crossref](#), [Medline](#) [Open URL](#)
11. Ojala, M, Garriga, GC. Permutation tests for studying classifier performance. *J Mach Learn Res* 2010; 11: 1833–1863. [Google Scholar](#) [Open URL](#)
12. Al-Rawi, MS, Cunha, JPS. On using permutation tests to estimate the classification significance of functional magnetic resonance imaging data. *Neurocomputing* 2012; 82: 224–233. [Google Scholar](#), [Crossref](#) [Open URL](#)
13. Anderson, TW, Darling, DA. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann Math Stat* 1952; 23: 193–212. [Google Scholar](#), [Crossref](#) [Open URL](#)
14. Anderson, TW, Darling, DA. A test of goodness of fit. *J Am Stat Assoc* 1954; 49: 765–769. [Google Scholar](#), [Crossref](#) [Open URL](#)
15. D'Agostino RB. Tests for the normal distribution. In: Ralph B, Agostino D' and Stephens MA (eds) *Goodness-of-fit techniques*. New York: Marcel Dekker, 1986, pp. 367–413. [Google Scholar](#)
16. Stephens, MA. EDF statistics for goodness of fit and some comparisons. *J Am Stat Assoc* 1974; 69: 730–737. [Google Scholar](#), [Crossref](#) [Open URL](#)
17. Rice JA. *Mathematical statistics and data analysis*. 2nd ed. Belmont, CA: Wadsworth Publishing Co Inc, 1994. [Google Scholar](#)
18. Billingsley, P. *Probability and measure* New York John Wiley & Sons. [Google Scholar](#)
19. Golland, P, Fischl, B Permutation tests for classification: towards statistical significance in image-based studies. In: Taylor, C, Noble, JA (eds). *Information processing in medical imaging, proceedings*, Berlin: Springer-Verlag, 2003, pp. 330–341. [Google Scholar](#), [Crossref](#) [Open URL](#)

20. Pahl, R, Schafer, H. PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics* 2010; 26: 2093–2100. [Google Scholar](#), [Crossref](#), [Medline](#) [Open URL](#)
21. Pfeiffer PE. *Concepts of probability theory*. New York: Dover Publications, 1978. [Google Scholar](#)
22. NETLAB. Netlab neural network software, [www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/](http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/) (accessed 1 August 2015). [Google Scholar](#) [Open URL](#)
23. Joachims T. SVM light, [www.cs.cornell.edu/people/tj/svm\\_light/svm\\_multiclass.html](http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html) (accessed 1 August 2015). [Google Scholar](#) [Open URL](#)
24. Joachims, T. Making large-scale support vector machine learning practical. In: *Bernhard Schölkopf, Christopher JC Burges, Alexander J Smola (eds) Advances in Kernel Methods*, Cambridge: MIT Press, 1999, pp. 169–184. [Google Scholar](#) [Open URL](#)
25. Norman, KA Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 2006; 10: 424–430. [Google Scholar](#), [Crossref](#), [Medline](#) [Open URL](#)
26. Princeton Multi-Voxel Pattern Analysis (MVPA) Toolbox, [www.csbmb.princeton.edu/mvpa/](http://www.csbmb.princeton.edu/mvpa/) (accessed 1 August 2015). [Google Scholar](#) [Open URL](#)
27. SPM8. Statistical parametric mapping, [www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/) (accessed 1 August 2015). [Google Scholar](#) [Open URL](#)
28. Etzel JA and Braver TS. MVPA permutation schemes permutation testing in the land of cross-validation. In: *2013 3rd International workshop on pattern recognition in neuroimaging*. Philadelphia: IEEE publications, 2013, pp.140–143. [Google Scholar](#)

Haxby et al. Faces and objects in ventral temporal cortex  
29. (fMRI), <http://dev.pymvpa.org/datadb/haxby2001.html> (accessed 1 August 2015). [Google Scholar](#)

Etzel, JA, Valchev, N, Keyzers, C. The impact of certain methodological choices on  
30. multivariate analysis of fMRI data with support vector machines. Neuroimage 2011;  
54: 1159–1167. [Google Scholar](#), [Crossref](#), [Medline](#)