NeuroImage: Clinical 2 (2013) 883–893

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical parkinsonian syndromes☆,☆☆

Gaëtan Garraux [a,c,*,1], Christophe Phillips [a,b,1], Jessica Schrouff [a], Alexandre Kreisler [f,g,h,i], Christian Lemaire [a], Christian Degueldre [a], Christian Delcour [e], Roland Hustinx [d], André Luxen [a], Alain Destée [f,g,h,i], Eric Salmon [a,c]

[a] Cyclotron Research Centre, Sart Tilman B30, University of Liège, 4000 Liège, Belgium
[b] Department of Electrical Engineering and Computer Science, Sart Tilman B28, University of Liège, 4000 Liège, Belgium
[c] Department of Neurology, University Hospital Centre, Sart Tilman B35, 4000 Liège, Belgium
[d] Department of Nuclear Medicine, University Hospital Centre, Sart Tilman B35, 4000 Liège, Belgium
[e] Department of Neurology, Centre Hospitalier Peltzer-La Tourelle, 4800 Verviers, Belgium
[f] Université Lille Nord de France, 59000 Lille, France
[g] UDSL (Université Droit et Santé de Lille), Lille, France
[h] CHU de Lille, France
[i] Movement Disorders Unit and EA 2683 MENRT, France

ARTICLE INFO

Article history:
Received 5 August 2012
Received in revised form 6 June 2013
Accepted 7 June 2013
Available online 14 June 2013

Keywords:
Computer-aided diagnosis
Data mining
Pattern recognition
Boostrap resampling
Bagging
Error-Correcting Output Code
Multiclass classification
Relevance vector machine
FDG PET
Parkinson's disease
Multiple system atrophy
Progressive supranuclear palsy
Corticobasal syndrome

ABSTRACT

Most available pattern recognition methods in neuroimaging address binary classification problems. Here, we used relevance vector machine (RVM) in combination with booststrap resampling ('bagging') for non-hierarchical multiclass classification. The method was tested on 120 cerebral ^{18}fluorodeoxyglucose (FDG) positron emission tomography (PET) scans performed in patients who exhibited parkinsonian clinical features for 3.5 years on average but that were outside the prevailing perception for Parkinson's disease (PD). A radiological diagnosis of PD was suggested for 30 patients at the time of PET imaging. However, at follow-up several years after PET imaging, 42 of them finally received a clinical diagnosis of PD. The remaining 78 APS patients were diagnosed with multiple system atrophy (MSA, N = 31), progressive supranuclear palsy (PSP, N = 26) and corticobasal syndrome (CBS, N = 21), respectively. With respect to this standard of truth, classification sensitivity, specificity, positive and negative predictive values for PD were 93% 83% 75% and 96%, respectively using binary RVM (PD vs. APS) and 90%, 87%, 79% and 94%, respectively, using multiclass RVM (PD vs. MSA vs. PSP vs. CBS). Multiclass RVM achieved 45%, 55% and 62% classification accuracy for, MSA, PSP and CBS, respectively. Finally, a majority confidence ratio was computed for each scan on the basis of class pairs that were the most frequently assigned by RVM. Altogether, the results suggest that automatic multiclass RVM classification of FDG PET scans achieves adequate performance for the early differentiation between PD and APS on the basis of cerebral FDG uptake patterns when the clinical diagnosis is felt uncertain. This approach cannot be recommended yet as an aid for distinction between the three APS classes under consideration.

1. Introduction

Computer-aided diagnosis (CAD) integrates data processing, mathematics and statistics into computerized techniques to maximize the information that may be extracted from medical imaging datasets. One of the goals of CAD is to assist the clinicians in the differential diagnosis between several conditions with overlapping clinical features. This problem is commonly encountered in patients with a presumed progressive adult-onset chronic neurodegenerative disorder, in which the clinical phenotype only fully expressed several years after the onset of brain damage. Most CAD in this context addressed a binary classification problem i.e., involving the distinction between two diagnostic classes. One of the challenges of CAD is multiclass classification (Kloppel et al., 2012), which better reflects a situation encountered in routine clinical practice. As compared with binary classification, multiclass classification is a more complex problem and their performances are difficult to compare directly. Here, we present simple binary and new multiclass classification methods and test their performance for the distinction between different forms of degenerative parkinsonism.

☆☆ *Financial support*: This research was supported by FRS-FNRS and a grant from the Rahier Foundation, University of Liège, Belgium.

* Corresponding author at: MoVeRe Group, Cyclotron Research Centre, University of Liège, Sart Tilman B30, 4000 Liège, Belgium. Tel.: +32 4 366 23 16; fax: +32 4 366 29 46.
E-mail address: ggarraux@ulg.ac.be (G. Garraux).

[1] Both authors equally contributed to this work.

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.06.004

Parkinsonism is clinically defined by the association of motor slowness, with muscle rigidity and/or tremor and/or a postural instability (Gibb, 1988). The most common cause of degenerative parkinsonism in adults is Parkinson's disease (PD). Much of the difficulty in the early diagnosis of PD is differentiating it from other forms of degenerative parkinsonism. A common source of misdiagnosis of PD is atypical parkinsonian syndromes (APS) that have a much poorer long-term prognosis such as multiple system atrophy (MSA), progressive supranuclear palsy (PSP) and corticobasal syndrome (CBS). In a clinico-pathological study conducted in a specialist movement disorder service, more than 60% of cases with a final clinical diagnosis of a parkinsonian syndrome other than PD had their diagnosis changed during the course of their illness. Of these, 60% were changed from an initial clinical diagnosis of PD (Hughes et al., 2002; Rajput et al., 1991).

Resting-state cerebral ^{18}fluorodeoxyglucose (FDG) uptake patterns measured using positron emission tomography (PET) has been recommended by the European Association of Nuclear Medicine Neuroimaging Committee for the differentiation between degenerative parkinsonisms (Varrone et al., 2009) under the assumption that FDG PET can capture specific functional and anatomical consequences of neuropathologic abnormalities specific of each condition. This is supported by the demonstration of group differences in regional FDG uptake between PD, MSA, PSP and CBS (Antonini et al., 1998; Eckert et al., 2005; Eidelberg et al., 1993; Feng et al., 2008; Ghaemi et al., 2002; Juh et al., 2004; Klein et al., 2005; Laureys et al., 1999; Otsuka et al., 1997; Teune et al., 2010). One of the most consistent abnormalities at visual inspection and semi-quantitative analyses is a relative decrease in striatal and frontal lobe tracer uptakes in APS as compared with PD or normal control populations. While this has been very informative at the group level, its diagnostic yield has been lower than expected in the early stages of these disorders because of overlapping individual *regional* FDG uptake across groups, which were often composed of small series of either well established cases studied with PET after a relatively long disease duration or early cases but without information on clinical follow-up to ascertain the initial clinical diagnosis (Garraux et al., 2000; Ghaemi et al., 2002; Juh et al., 2004; Laureys et al., 1999; Otsuka et al., 1997).

Here, we examined the value of CAD for the distinction between PD, MSA, PSP and CBS on the basis of cerebral FDG PET. The present study differs from previous ones by several methodological aspects with respect to both the population characteristics and analysis methods. First, to maximize the clinical significance of cerebral FDG PET for distinction between the diagnostic classes under consideration, we included scans performed in the first years after symptom onset (Table 1) at a time when clinical features were outside the prevailing perceptions for PD. Diagnostic classes were then defined later by the retrospective application of clinical diagnostic criteria for PD and APS at follow-up, on average ~8.0 and ~2.8 years after PET assessment (i.e., standard of truth). Second, a crucial difference with previous studies is the analysis of neuroimaging data using an automatic voxel-based multivariate supervised machine learning method, "Relevance Vector Machine" (RVM) (Tipping, 2001), that we have previously applied on a binary case for the distinction between patients with or without an altered state of consciousness on the basis of cerebral FDG uptake patterns (Phillips et al., 2011). We profoundly modified this method to be suitable for multiclass classification.

Classification was both performed in a binary sense, PD versus all the APS subcategories pooled into a single class, and in a multiclass sense, PD and the 3 APS categories considered separately. For multiclass classification, pairwise coupling is a popular approach that combines all comparisons for each pair of classes (Fürnkranz, 2002). Here, we used a one-versus-one approach involving six binary RVM classifiers from which a single prediction was obtained using an Error-Correcting Output Code (ECOC) approach (Dietterich and Bakiri, 1995) (see Section 2.4.3). For cross-validation and assessment of prediction accuracy, RVM was combined with bootstrap aggregation (also known as "bagging") (Breiman, 1996; Efron and Tibshirani, 1993) and the final RVM class assigned to each FDG PET scan was defined by the prediction that received the most votes (see Section 2.4.4).

The final class assigned to each FDG PET scan was then compared with the clinical diagnosis at follow-up to estimate prediction accuracy, sensitivity, specificity, positive (PPV) and negative (NPV) predictive values. The statistical significance of RVM classification accuracy was assessed using a permutation testing (see Section 2.4.4). We also compared binary RVM classification with the radiological diagnosis of the nuclear medicine specialist at the time of PET imaging (i.e., for the distinction between PD and APS).

Finally, from the vote counting in the bootstrap procedure, a "majority confidence ratio" was estimated for each scan on the basis of class pairs that were the most frequently assigned by RVM. This level of confidence was further linked to the PPV (see Section 2.4.5). We believe that this qualification of the classification outcome may provide clinically relevant information at the individual level for physicians who usually request FDG PET scans as an aid to solve a multiclass diagnosis problem.

2. Methods

2.1. Subjects

Patients were all referred for cerebral FDG PET at the Cyclotron Research Centre (CRC), University of Liège, or the University Hospital Center (CHU) of Liège by neurologists because clinical features were outside the prevailing perceptions for PD. In many cases, no other specific diagnosis was mentioned in the PET order form and no standardized clinical assessment was available in this retrospective study. The most frequent atypical features at referral were an equivocal clinical response to scheduled L-DOPA administration, prominent axial symptoms, greater than expected asymmetry of parkinsonian signs, early falls, or the co-occurrence of other features such as a pyramidal and/or cerebellar syndrome, limb dystonic posturing, oculomotor abnormalities, or severe dysautonomic dysfunction. All subjects included in this research protocol gave their written informed consent to participate in the study; the study protocol was approved by the Ethical Committee of the University of Liège.

Table 1
Demographic and clinical data.

	N	Gender (F/M)	Data at the time of PET assessment			Last available follow-up
			Mean age (years)	Mean disease duration (years)	Mean LEDD (mg)	Mean disease duration (years)
PD	42	17/25	56.9 ± 10.3	3.6 ± 3.1	442 ± 239	11.6 ± 5.1
MSA	31	18/13	66.0 ± 8.8	3.4 ± 2.9	559 ± 298	6.4 ± 3.9
PSP	26	9/17	69.4 ± 7.3	3.1 ± 2.4	281 ± 250	5.9 ± 4
CBS	21	15/6	67.8 ± 7	3.3 ± 2	164 ± 189	5.9 ± 2.9
All classes	120	59/61	63.9 ± 10.2	3.4 ± 2.7	386 ± 284	8.0 ± 5.0

LEDD = L-DOPA equivalent daily dose (Tomlinson et al., 2010).

In this retrospective analysis, one hundred and twenty scans from individuals diagnosed with PD, MSA, PSP and CBS at follow-up were selected for inclusion (Table 1). Patients were included on the basis of the United Kingdom Parkinson's Disease Society Brain Bank (UKPDSBB) clinical criteria for PD (Hughes et al., 1992), or the Neuroprotection and Natural History in Parkinson Plus Syndromes (NNIPPS) criteria (Bensimon et al., 2009) for MSA or PSP, or the Lang criteria for CBS (Lang et al., 1994) as reviewed from the medical records by two movement disorders specialists (GG and AD). In the MSA group, one, nineteen and eleven patients were clinically classified as MSA-A, MSA-P and MSA-C, respectively. Detailed clinical motor and neuropsychological assessments were not available for all patients. We considered as an exclusion criterion a clinical follow-up of less than 12 months after PET imaging in order to reduce the risk of clinical misdiagnosis. Other exclusion criteria include any significant structural brain abnormalities on CT scan or MRI, and exposure to drugs that could have caused the clinical findings. Because the PET scanner employed for data acquisition at the CRC had a limited field of view in the axial direction (10.31 cm), we also excluded FDG PET scans when brain coverage in the axial direction was judged inadequate at visual inspection.

2.2. Imaging data acquisition

Image acquisition was performed between 1993 and 2009 either at the CRC (N = 87) using a CTI 951 R 16/31 tomograph (CTI, Knoxville, TN, USA) or at the CHU (N = 33) on a Gemini PET/CT scan (Philips Medical Systems) after an intravenous bolus injection of FDG. The proportion of images acquired on the 2 scanners is the following: overall 72/28%; 74/26% and 72/28% for the PD and APS, respectively; 74/26%, 68/32%, 62/38% and 90/10% for the PD, MSA, PSP and CBS respectively. Globally, the scans from different categories are thus similarly distributed across scanners, except for the CBS, which also counts the fewer scans over all. Participants were studied on their usual medications in a quiet wakeful resting-state, with eyes closed in dimmed ambient light. Head movements were reduced using foam padding and a restraining strap.

2.3. Imaging data processing

After gross manual image reorientation and approximate definition of the image center point, the PET images were spatially processed using the Statistical Parametric Mapping toolbox (SPM8, Wellcome Trust Centre for Neuroimaging, 2008) implemented within Matlab 7.4.0 (MathWorks, Natick, MA, USA).

All images were spatially normalized onto a population-specific FDG PET template created in MNI space, as previously reported (Phillips et al., 2011), and then spatially smoothed using a 12 mm FWHM Gaussian kernel (Garraux et al., 2000). To minimize any bias in the RVM analyses due to differences in brain coverage in the axial direction, we next created a binary mask image representing the brain voxels common to all scans. Furthermore, within this mask, we only considered voxels that had a probability of being grey matter higher than .33, according to the tissue probability map provided in SPM8. To account for the variability associated with various sources of physiological and non-physiological noise inherent to PET data, intensity normalization of regional tracer uptake to the global mean activity was applied to each scan prior to their analysis using a proportional scaling procedure (Friston et al., 1990).

2.4. PET data analysis

2.4.1. Radiological diagnosis

We retrospectively examined the radiological reports of the nuclear medicine specialists who reviewed FDG scans at the time of PET imaging and computed the number of scans considered as suggestive of PD and APS.

2.4.2. Relevance vector machine (RVM) classification

Voxel-based multivariate analysis of FDG PET scans was performed using a RVM (Tipping, 2001) and lead to 'pattern recognition' in the data. RVM relies on the same principle as "Support Vector Machine" (SVM) but is framed within a Bayesian framework contrary to SVM (which is expressed as "maximal margin" problem).

The linear RVM method applied here on FDG-PET data was similar to that described in Phillips et al. (2011) and allowed the classification of data points, i.e. PET scans, into two classes. This is a so-called supervised learning approach since the machine is trained on a training dataset where the true class membership of each data point is provided. Briefly, the RVM belongs to a set of sparse machine-learning approaches that builds a classification/regression function from a weighted linear combination of kernel functions, in which the weights are tuned during the learning phase to produce an optimal classification of the training data (Krishnapuram et al., 2005). Sparse means that the weight estimates are encouraged during the learning process to be either high or exactly zero, to make the model more parsimonious, efficient to run, to avoid over-fitting, and to improve generalization capacity. Based on the output weights, a posterior class-probability of a new test image can be estimated (Tipping, 2001). This posterior class-probability can eventually be thresholded (usually at 0.5 for balanced data set) to provide a class-prediction on the test instance.

Here, RVM was used for the distinction between PD and APS in binary and multiclass situations, when the 3 subcategories (MSA, PSP and CBS) of APS are considered. Practically, this multiclass RVM relied on a set of 6 pairwise RVM's for the one-to-one classification of each pair of classes (PD vs. MSA, PD vs. PSP, PD vs. CBS, MSA vs. PSP, MSA vs. CBS, and PSP vs. CBS) and the output of the 6 classifications were recombined afterwards (see Section 2.4.3). We assessed the performance of both multiclass RVM and conventional binary RVM.

2.4.3. Bootstrap resampling ("bagging")

For cross-validation and assessment of prediction accuracy, RVM was combined with bootstrap aggregation (also known as "bagging") (Breiman, 1996; Efron and Tibshirani, 1993). In statistics, resampling techniques are used to validate models and to assess their statistical accuracy by using random subsets (bootstrapping cross-validation) (Efron and Tibshirani, 1986; Efron and Tibshirani, 1995). Bootstrap resampling consists in uniformly sampling objects from a dataset, with replacement. "Bags" of data were created by random sampling, with replacement, from the original pool of training data. Here, we performed 100 iterations, involving a new bootstrap sample per iteration. At each iteration, a fixed number of PET images were randomly sampled (with replacement) from each category to form the training set, used to build the RVM models (Fig. 1). This fixed number was computed as the number of images in the class with the smallest sample size and therefore depended on whether the RVM analysis was binary (i.e., PD versus APS, 42 images of each class selected) or multiclass (i.e., PD, MSA, PSP and CBS classes considered, 21 images of each category selected). The remaining images formed the test set, which was used to test the built RVM models. Note that, because of the replacement step in bagging, even if N images are randomly sampled from a group of N images, some images will be selected multiple times in the training set and there will remain on average .37*N not-selected images to form the test set. As a result of this division between training and test sets, the training set is balanced between classes while the test set shows similar proportions of each class as in the whole data set. The bootstrap sample considered here can therefore be considered as stratified (Efron and Tibshirani, 1993).

For the binary RVM analysis (PD vs. APS), at each bootstrap resampling, an RVM is trained with the training set. Then the trained RVM classifier is applied on the scans in the test set and the class assigned to each test-scan is obtained by thresholding its posterior class-probability at 0.5 (since the training set is balanced).

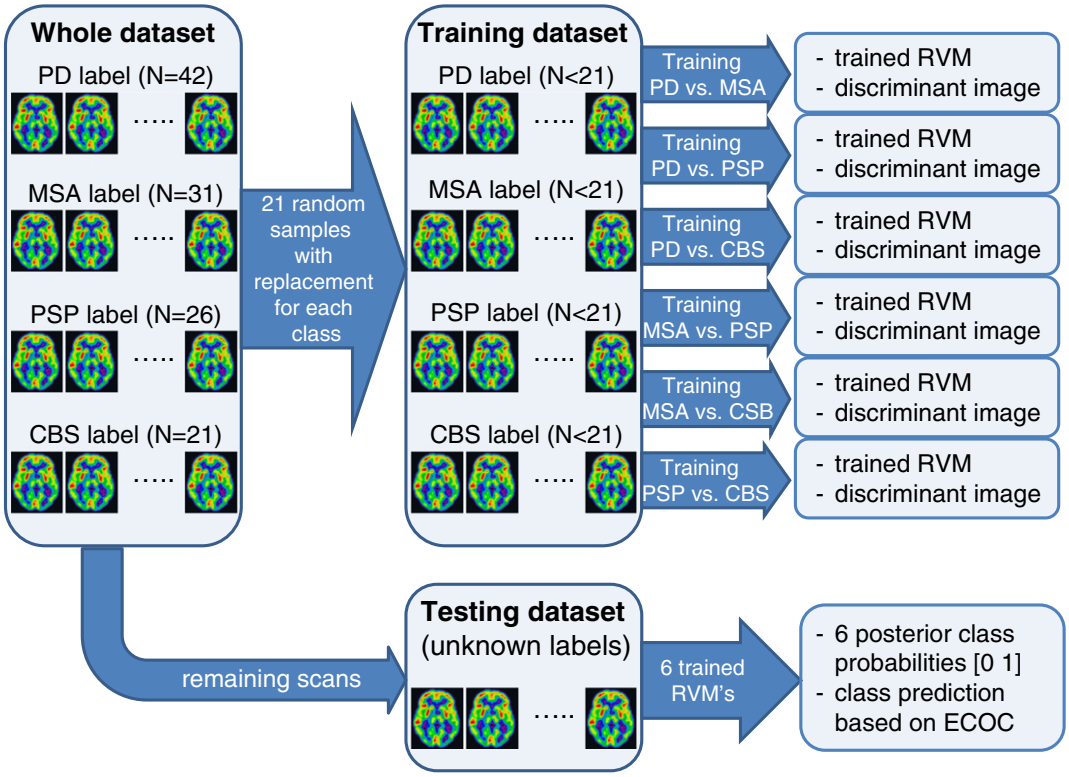

Fig. 1. Bootstrap resampling with replacement ("bagging"). At each iteration, the whole FDG-PET dataset was split into training and test sets. A prediction was assigned to each test instance by each of the six trained RVM models. A single prediction (PD, MSA, PSP or CBS) was obtained from the six RVM models using an Error-Correcting Output Code (ECOC) approach (Dietterich and Bakiri, 1995).

Class assignment of a test instance is more complex in the multiclass classification problem as class prediction relies on the output of 6 pairwise RVM classifiers (Fürnkranz, 2002). Here, this issue was addressed using an Error-Correcting Output Code approach (ECOC) (Dietterich and Bakiri, 1995). In the ECOC scheme, each class is represented by a code-word of length *n*, *n* being the number of pairwise classifications performed, and each character of the code-word is the expected output of the corresponding binary classifier for the specific class (Inline Supplementary Table S1). Given a test image, the six binary RVM's (PD vs. MSA, PD vs. PSP, PD vs. CBS, MSA vs. PSP, MSA vs. CBS, PSP vs. CBS) return 6 probabilistic values forming a test-word that is then compared to the four code-words. Eventually, the class whose code-word leads to the smallest distance is picked as the predicted class for the test image (Hassabis et al., 2009; Mourao-Miranda et al., 2006). Given the probabilistic nature of an RVM output, the distance between the test-word and each code-words is estimated as the sum of absolute difference between the code-word characters and output from each RVM, i.e. using an L1 norm (Schrouff et al., 2012).

At each resampling of the bootstrap procedure of both binary and multiclass RVM, each test instance is thus assigned to a single class. So, after the 100 bagging iterations, we were able to rank the predictions for each scan according to the proportion of the respective votes received. The final RVM class assigned to each FDG PET scan was defined by the prediction that received the most votes, i.e. the most often prediction assigned over the iterations where the scan appeared in the test set (Table 2).

Inline Supplementary Table S1 can be found online at http://dx.doi.org/10.1016/j.nicl.2013.06.004.

2.4.4. RVM classification accuracy

Multiclass and binary RVM were both considered in all analyses detailed in this section. Classification accuracy estimates were assessed by comparing the class assigned to each FDG PET scan with the standard of truth given by the diagnosis at the last available follow-up (Table 2).

We computed prediction accuracy estimates. Overall accuracy was defined as the number of scans correctly classified over the total number of scans, while balanced accuracy was the mean of the class-specific accuracies. By definition, class accuracy is the proportion of images pertaining to each class that are correctly classified. Here, class accuracy estimates were summarized in a confusion matrix where each row represents the instances in a predicted class and each column the instances in a diagnostic class. From the confusion matrix, we computed discrimination measures: sensitivity, specificity, positive and negative predictive values (PPV and NPV) for PD as compared with APS. In multiclass RVM, the PPV was estimated, for each class, as the ratio between the number of scans correctly classified in a class and the total number of scans classified in that class. The NPV was estimated as the ratio between the number of scans correctly not classified in a class and the total number of scans not classified in that class.

We used a permutation approach to make statistical inferences on prediction accuracy under the null hypothesis of classification at chance level. The following three steps were repeated 1000 times: i) class labels (PD/APS or PD/MSA/PSP/CBS for the binary or multiclass RVM, respectively) were randomly permuted between all scans, ii) the (binary or multiclass) classifier was trained, with bootstrap sampling, on the basis of these random labels, and iii) the resulting accuracy was calculated for current label permutation. The observed classification accuracy obtained with the original diagnostic labels was then compared to the histogram of accuracy values over permutations. A p-value was derived as the ratio between the number of permutations with accuracy higher or equal than the original accuracy, and the number of permutations. This p-value is thus an estimation of the probability that a random permutation of the labels leads to higher classification accuracy than the

Table 2
Bootstrap resampling procedure (multiclass RVM).

	Bag #1	Bag #2	Bag #3	Bag #n-2	Bag #n-1	Bag #n	Vote counting	Standard of truth (SOT)	RVM prediction accuracy	Majority confidence ratio
	Vote #1	Vote #2	Vote # 3	Vote #n-2	Vote #n-1	Vote #n	Majority vote?	Clinical diagnosis	Majority vote = SOT?	
Scan #1	PD	–	MSA	PD	PD	–	$N_{PD} = 32$ $N_{MSA} = 2$ $N_{PSP} = 1$ $N_{CBS} = 0$	PD	1	$(32 - 2) / 35 * 100 = 85\%$
Scan #2	PD	PSP	–	MSA	MSA	PD	$N_{PD} = 13$ $N_{MSA} = 9$ $N_{PSP} = 5$ $N_{CBS} = 4$	MSA	0	$(13 - 9) / 31 * 100 = 13\%$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Scan #120	PD	PSP	CBS	–	–	PSP	$N_{PD} = 4$ $N_{MSA} = 3$ $N_{PSP} = 13$ $N_{CBS} = 10$	CBS	0	$(13 - 10) / 30 * 100 = 10\%$

– = scan included in the training set and not in the test set in this bootstrap sample.

true diagnostic labels. Here, the "chance level" was estimated as the mean "classification accuracy" over the 1000 permutations.

Finally, RVM classification accuracy was also compared with the radiological diagnosis made by the nuclear medicine specialist at the time of PET on the basis of clinical and imaging features. This analysis was restricted to the differentiation between PD and APS, since this best matched the clinical question under consideration. The RVM classification and radiological diagnostic could be in agreement (correctly or incorrectly) or disagreement, with one correct and the other incorrect (4 possible scenarios). We estimated the proportion of scans in each scenario when PD and APS patients are considered separately and pooled together.

2.4.5. Diagnostic reliability

In addition to the methods presented in the previous section investigating RVM prediction accuracy, we estimated a classification 'majority confidence ratio' measure that could also be delivered to the clinicians to assist them in their diagnostic process. This additional measure takes advantage of the bagging procedure and is computed on the basis of the two classes that were the most frequently assigned by RVM over baggings.

For each scan, a classification 'majority confidence ratio' was computed as the ratio between the difference in the number of votes between the two classes receiving the largest number of votes, and the number of times the scan was picked in the test set (i.e., total number of votes), expressed in percent (Table 2). In other words, a value of 100% indicates that each time the scan was picked in the test set it was classified in the same class. Conversely, the smaller the value, the more variables were the votes across the bootstrap samplings.

Then, we examined how the PPV varied according to this majority confidence ratio. For a given threshold on confidence level *tc*, we estimated the corresponding PPV(*tc*) by counting the number of scans – total and correctly classified in a class – with a confidence value above *tc*. *tc* was varied from 0% to 90% by steps of 10%. One would expect that the higher confidence level, the higher is the PPV value (i.e., fewer false positives) and the lower is the number of scans considered for computing the corresponding PPV. Finally a correlation coefficient (with its associated p-value) was calculated between the *tc* level considered and the observed PPV(*tc*). This correlation coefficient is helpful in assessing the information conveyed by the 'majority confidence ratio' with respect to the PPV of RVM classification.

2.4.6. Relevance maps

With voxel-based multivariate image classification methods, all voxels potentially contribute to the classification but their respective contribution is not equal. Given a trained RVM with a linear kernel, this relevance value varies from voxel to voxel and can be summarized as a discriminant image or "relevance map". A relevance map thus represents the joint pattern of all voxels whose relative level of activity allows the discrimination between the two classes of data under consideration. Two networks can generally be identified when class A is compared to class B: an excess (deficit) network corresponding to regions displaying positive (negative) relevance, indicating that a relative increase in FDG uptake in these regions increases the likelihood of classification in class A (class B).

A relevance map was created at each resampling of the bootstrap procedure giving in total 100 images for each binary RVM model (Fig. 1). Consistency in the discriminant patterns over resamplings was assessed by normalizing the weight obtained across the 100 bags. This was done in a standardized 1st moment sense, i.e. the mean of the 100 weights (at each voxel) divided by its standard deviation.

By convention here, the excess and deficit networks in discriminant standardized maps were represented by positive and negative Z values, respectively. The largest (absolute) values in standardized maps highlight brain regions where FDG uptake levels contribute the most consistently (over the 100 bootstrap samplings) to the overall distinction between the two classes under consideration. Conversely, voxels with a Z standardized value close to zero have a relatively variable (across the bootstrap samplings) contribution to the distinction between the two diagnostic classes under consideration.

Note that these standardized maps cannot be thresholded as is usually done in univariate analysis (statistical parametric maps) because they reflect the distributed nature of a multivariate analysis. Nevertheless, in a proper cross-validation scheme, they could be used for feature selection such as "recursive feature elimination" (De Martino et al., 2008).

3. Results

An estimation of the computational cost of testing, training and validating the machines is provided in the supplementary material.

3.1. RVM classification accuracy

3.1.1. Binary RVM analysis

On average, scans from PD and APS classes were incorporated in the training dataset, as expected (Breiman, 1996), in 64% and 42%

Table 3
Confusion matrix derived from bootstrap aggregation (bagging) in binary RVM.

RVM classification	Diagnostic classes (SOT)		PPV & NPV
	PD	APS	
PD	**39**	13	.75
APS	3	**65**	.96
Class accuracy (p-value)	.93 (0.0)	.83 (0.0)	

The table shows class accuracies (with the associate p-value) and positive/negative predictive values (PPV and NPV). SOT = standard of truth. The number of scans correctly classified in each class is indicated in bold.

of bootstrap samples, respectively. Overall and balanced accuracies are 87% and 88% respectively. The confusion matrix obtained is shown in Table 3, which also includes the class accuracies as well as PPV and NPV. Classification accuracy estimates are significantly ($p < 0.05$) above chance levels (0.46 and 0.53 for the PD and APS classes respectively).

The comparison between RVM classification and radiological diagnosis accuracy is summarized in Table 4.

In comparison with the final clinical diagnosis at the last follow-up (i.e. standard of truth), RVM and clinical diagnoses were correctly in agreement for most (74%) of the patients and both of them are jointly incorrect for only a few scans (3 out of 120). They disagreed in 24% of the scans (28 patients out of 120). Strikingly, at visual inspection, 36% of the PET scans from patients who received a final clinical diagnosis of PD at the last follow-up were considered not suggestive of PD by the nuclear medicine specialist. On the other hand, RVM tended to slightly underdiagnose APS (16%) as compared with the radiological evaluation.

3.1.2. Multiclass RVM analysis

On average, scans from PD, MSA, PSP and CBS classes were incorporated in the training dataset in 39%, 50%, 56%, and 64% of bootstrap samples, respectively. Overall and balanced accuracies are 66% and 63% respectively. The confusion matrix obtained is shown in Table 5, which also includes the class accuracies as well as PPV and NPV. Classification accuracy was significantly ($p < 0.05$) above chance levels (0.26, 0.25, 0.25 and 0.23 for the PD, MSA, PSP and CBS classes respectively) for the PD and CBS classes only. On the one hand, PD scans are accurately classified and those misclassified seem evenly distributed between the three APS classes (MSA, PSP and CBS). On the other hand, accuracy is lower for the individual MSA/PSP/CBS classification but most misclassified APS scans are distributed among themselves: only 10 out of 37 misclassified MSA/PSP/CBS scans were classified into the PD class.

Further information related to the scanner used to acquire the FDG PET images (at CRC or CHU) can be found in the supplementary material.

3.2. Diagnostic reliability

Fig. 2 shows the PPV as a function of the majority confidence ratio, for the binary and multiclass cases. The correlation coefficients between the PPV and confidence level are .98 ($p < 10^{-6}$) and .93 ($p < 10^{-4}$) for the binary case (PD and APS classes respectively) and .97 ($p < 10^{-5}$), .95 ($p < 10^{-4}$),−.53 ($p = .12$) and .71 ($p = .02$) for the multiclass case (PD, MSA, PSP and CBS classes respectively).

Table 4
Accuracy of RVM classification and the radiological diagnosis at the time of PET.

	42 PD patients	78 APS patients	Total 120 patients
Correct agreement	26 (62%)	63 (81%)	89 (74%)
Incorrect agreement	2 (5%)	1 (1%)	3 (3%)
Correct RVM, incorrect radiological	13 (31%)	2 (3%)	15 (13%)
Correct radiological, incorrect RVM	1 (2%)	12 (15%)	13 (11%)

The table summarizes the accuracy of binary RVM classification and radiological diagnosis, for the two diagnostic classes (PD and APS) together (last column) or separately. RVM classification and the radiological diagnostic could be in agreement (correctly or incorrectly) or disagreement, with one correct and the other incorrect with respect to the standard of truth (SOT) given by the clinical diagnosis at the last available follow-up several years after PET assessment (Table 1).

Table 5
Confusion matrix derived from bootstrap aggregation (bagging) in multiclass RVM.

RVM classification	Diagnostic classes (SOT)				PPV/NPV
	PD	MSA	PSP	CBS	
PD	**38**	6	2	2	.79/.94
MSA	1	**14**	5	1	.67/.83
PSP	1	7	**14**	5	.52/.87
CBS	2	4	5	**13**	.54/.92
Class accuracy (p-value)	.90 (0.0)	.45 (.149)	.55 (.067)	.62 (.025)	

The table shows class accuracies (with the associate p-value) and positive/negative predictive values (PPV and NPV). SOT = standard of truth. The number of scans correctly classified in each class is indicated in bold.

For the binary case, the PPV of both PD and APS (almost) monotonically increase as a function of the confidence level. Given the curves and significant correlation coefficients, the classification confidence estimated for each scan appeared to be a good indicator of PPV. Nevertheless this majority confidence ratio is not a direct estimator of the PPV value as both sets of values varies over relatively different scales, for example a majority confidence ratio of 50% corresponds to an 80% PPV for the PD class.

For the multiclass case, the PPV also increases regularly with the level of confidence for the PD, MSA and CBS classes. As in the binary case, the classification confidence seems a good indicator of PPV for these three classes. For PSP, several scans were misclassified as PSP with a relatively high confidence level (>60%). This leads to a maximum of PPV (67%) above 40% confidence and a drop of the PPV value for higher confidence level (down to 33% for confidence above 80%).

In both binary and more particularly multiclass cases, the 'dips' in the PPV curves are due to the low number of scans correctly/incorrectly classified over the whole scale of confidence values (from 0 to 100%). With fewer bins (confidence levels above 0, 20, 40, 60 and 80%) the curves have similar profile but are smoother.

3.3. Relevance maps

3.3.1. Binary RVM analysis

RVM identified two types of discriminating patterns between PD and APS, as seen on the standardized map shown on Fig. 3A. The excess network (EN) mainly encompassed the ventral part of upper brain stem, medial thalami, ventral striatum, head of caudate nuclei, medial temporal areas, middle and anterior cingulate areas, medial frontal cortex including the pre-supplementary motor area (SMA), insula cortex, superior and caudal aspects of dorsal frontal cortices. The deficit network (DN) included the lateral aspects of both thalami, posterior associative areas mainly in medial parietal areas and posterior cingulate gyri, lateral temporal and occipital areas, as well as the inferior part of the frontal lobe including subgenual, orbitofrontal and inferior lateral prefrontal cortices.

3.3.2. Multiclass RVM analysis

Discriminant standardized maps for the multiclass RVM analysis are shown in Fig. 3B (PD vs. MSA, PSP, and CBS) and Inline Supplementary Fig. S1 (comparison between APS subtypes).

Inline Supplementary Fig. S1 can be found online at http://dx.doi.org/10.1016/j.nicl.2013.06.004.

- *PD* vs *MSA*: the main constituting areas of the EN were the cerebellum (both vermis and cerebellar hemispheres), medial thalami, posterior putamen, caudate nuclei, the hypothalamic region, limbic areas (anterior and middle cingulate regions and insula cortices), caudal and lateral aspects of frontal lobes. The DN network

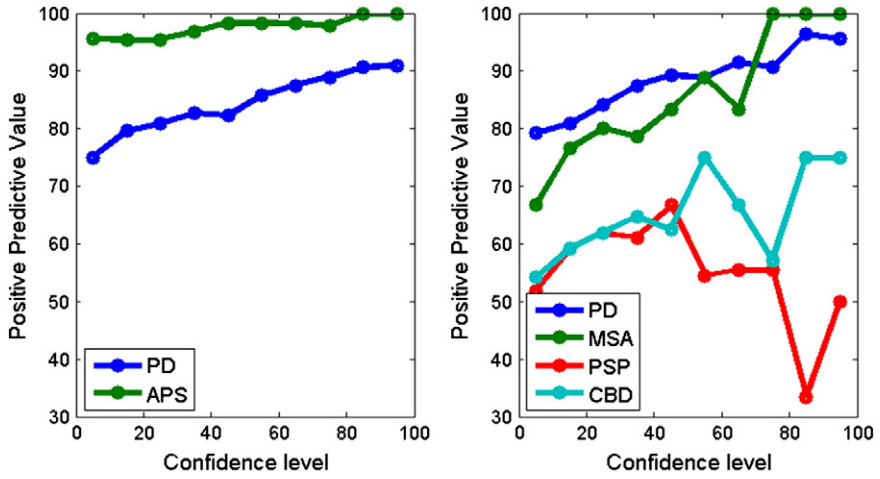

Fig. 2. Positive predictive values as a function of the majority confidence ratio. Variation of positive predictive value (PPV) when only scans above a classification confidence threshold are counted (total number and number of correctly classified): the threshold *tc* of confidence level is varied from 0 to 90% by steps of 10%.

mainly involved the lateral thalamic areas and posterior associative cortices and inferior frontal lobes.

- *PD* vs *PSP*: the EN mainly comprised upper brain stem, medial thalami, caudate nuclei, ventral striatum, insula cortex, medial frontal areas (including the subgenual and anterior cingulate cortices) and both lateral and medial aspects of caudal frontal lobes. The DN included the lateral thalami, posterior associative cortices; the DN also encompassed middle and inferior frontal cortex.
- *PD* vs *CBS*: the EN mainly inxcluded upper brain stem, medial thalami, putamen, caudate nuclei, insula cortices, medial frontal cortex, and caudal lateral frontal areas. As in the comparison with MSA and PSP, the DN strongly involved posterior associative cortices and middle and inferior frontal areas.
- *MSA* vs *PSP*: the EN encompassed medial temporal areas and rostral medial frontal areas including the presupplementary motor area and anterior cingulate cortex. Bilateral caudal lateral frontal cortices and lateral parietal cortices were also part of the EN. Bilateral thalami, posterior putamen and cerebellum were the main constituting parts of the DN along with perirolandic regions, and posterior associative cortices.
- *MSA* vs *CBS*: the EN encompassed lenticular nuclei, insula cortex, and frontal areas mainly in their medial aspects. The DN mainly encompassed the cerebellar vermis and globi pallidi, bilaterally.
- *PSP* vs *CBS*: the EN mainly included cerebellum, thalamus and bilateral lenticular nuclei; at the cortical level, the EN was mostly composed of caudal frontal areas and primary sensori-motor cortices, bilaterally; the DN encompassed upper brainstem, medial thalamus and globus pallidum, bilaterally.

4. Discussion

We have presented here an original supervised machine learning method for both binary and multiclass classification of neuroimaging datasets of a single modality by using RVM in combination with booststrap resampling (bagging). The method is fast, requires little user intervention and could be easily extended to a clinical setting. Generalizability and accuracy were investigated on the early distinction between PD and three other forms of degenerative parkinsonism on the basis of cerebral FDG uptake pattern measured with PET. The clinical question under consideration is not trivial since PD is associated with a much better long-term prognosis than APS. At the time of PET imaging, all 120 participants exhibited parkinsonian features that were outside the prevailing perception for PD. The radiological diagnosis based on visual inspection of FDG data by a nuclear medicine specialist at the time of PET was PD and APS for 30 and 90 scans, respectively. However, at follow-up on average 8 years after PET imaging, 42 patients finally received a clinical diagnosis of PD. The remaining 78 APS patients were diagnosed with MSA (N = 31), PSP (N = 26) and CBS (N = 21), respectively. In comparison with this standard of truth, the sensitivity (accuracy), specificity, PPV and NPV of the radiological diagnosis for PD as compared with APS were 64%, 96%, 90% and 83%, respectively.

RVM prediction sensitivity, specificity, PPV and NPV for PD were 93% 83% 75% and 96%, respectively using binary RVM (PD vs. APS; Table 3) and 90%, 87%, 79% and 94%, respectively, using multiclass RVM (PD vs. MSA vs. PSP vs. CBS; Table 5). Multiclass RVM achieved 45%, 55% and 62% classification accuracy for, MSA, PSP and CBS, respectively. Prediction accuracy for PD was significantly better than prediction at chance level estimated from random permutations in both binary and multiclass RVM. Classification accuracy was also significantly above chance level for the APS class and CBS using binary and multiclass RVM, respectively. Altogether, the results suggest that automatic multiclass RVM classification of FDG PET scans is suitable for CAD between PD and APS but cannot be recommended yet as an aid for distinction between the three APS classes under consideration. RVM classification performance should improve by feeding the classifier with qualitative (i.e., motor symptom asymmetry, oculo-motor disturbances…) and quantitative (i.e., L-DOPA responsiveness…) clinical features (Warr and Walker, 2012). However, the combination of different modalities in a single model is not immediate and still requires important methodological developments (using for example a "Multi-kernel learning" technique as in Gonen and Alpaydin, 2011).

We compared binary RVM classification with the radiological diagnosis (PD or APS) made by nuclear medicine specialists who reviewed imaging features at the time of PET (Table 4). RVM classification and radiological diagnosis were frequently in agreement (74% and 3% for correctly and incorrectly classification) but when they disagreed they did so differently: PD and APS misclassification occurred more frequently using radiological and RVM classification, respectively. This suggests that there may have been a tendency amongst nuclear medicine specialists to place too much emphasis on subtle atypical FDG PET imaging features for PD and to overdiagnose APS. This is perhaps not unreasonable since patients were referred for clinical features that were felt atypical for PD. Alternatively, this may also suggest a more variable FDG uptake pattern in PD than formerly recognized in our specific group of patients. This variability is therefore used by RVM to build more sensitive (albeit less specific) classification models for PD.

RVM prediction accuracy estimates for PD as compared with the three other classes were similar to those reported by Tang et al.

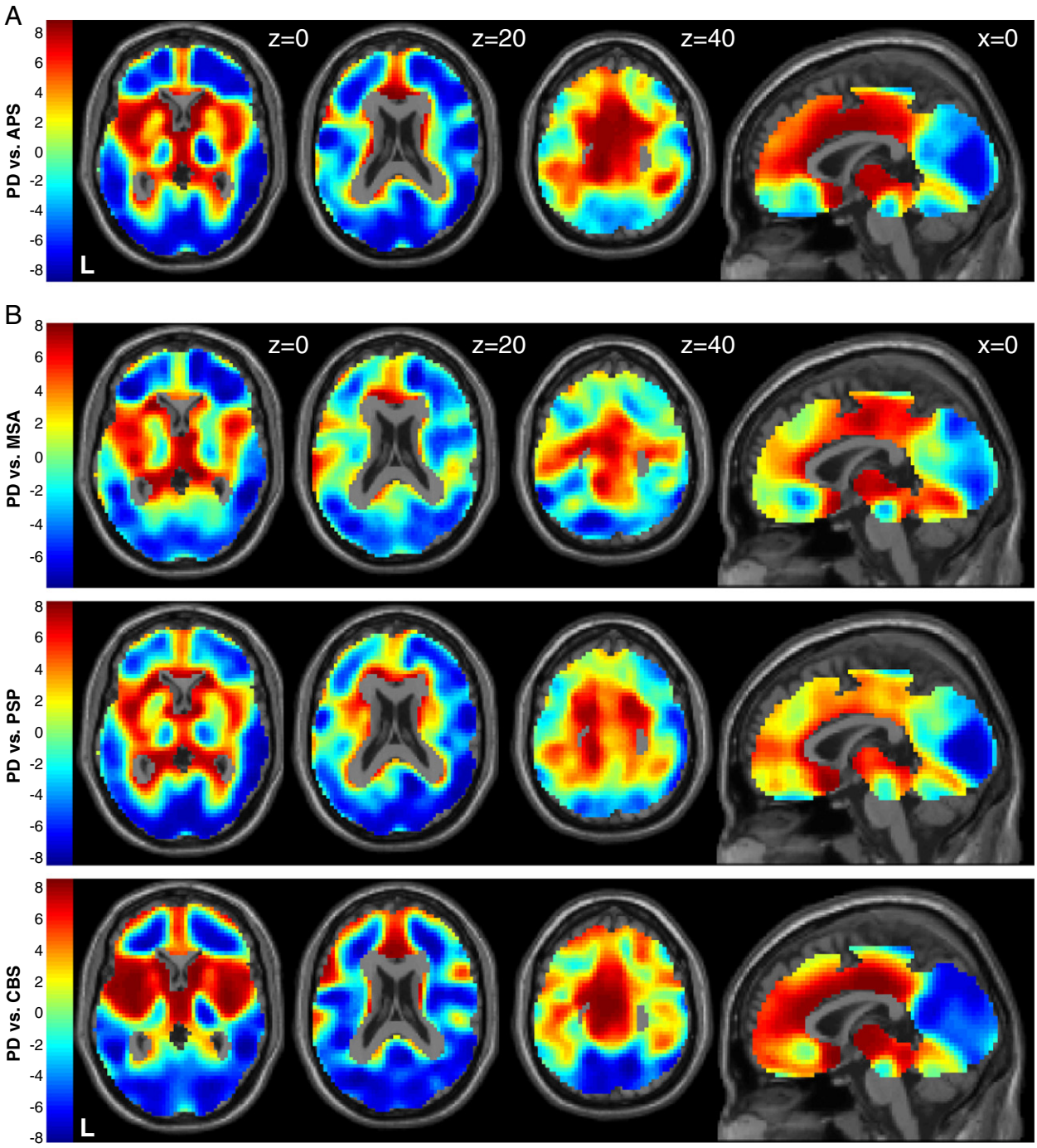

Fig. 3. Discriminant standardized maps between PD and APS. Unthresholded discriminant standardized map computed from the binary (A) and multiclass (B) RVM analyses comparing PD and APS classes. The color scale represents the standardized values computed on the basis of the 100 discriminant images created during bootstrap resampling (Fig. 1). By convention here, the excess network (EN) where FDG uptake is relatively increased in the PD class as compared with the APS class is represented by positive standardized values while relative deficits are represented by negative standardized values. Most voxels have a standardized value close to zero and therefore their contribution to the distinction between the two diagnostic classes under consideration is rather variable across the bootstrap samplings. The discriminant maps are displayed on representative axial (Z = 0, 20 and 40 mm), and sagittal (X = 0 mm) slices through a standard T1-weighted MRI in stereotactic space. Z and X values at the bottom indicate the distance (in mm) of the image from the axial plane through the anterior and posterior commissures and from the parasagittal plane through the midline, respectively. L = Left.

(2010) in the only comparable FDG PET study despite profound methodological differences (Tang et al., 2010). In that study, class prediction was based on a logistic regression model involving three predictor variables that quantified the individual expression of metabolic covariance patterns related to PD, MSA, and PSP classes. These patterns were defined by voxel-based principal component analyses (PCA) that, by definition, seek the directions of greatest variation in the FDG PET datasets (Spetsieris et al., 2009). Another important methodological difference is the hierarchical (two-level) binary classification approach used by Tang et al (2010): the distinction between PD and APS was addressed at the first level while the second level involved the distinction between the two APS classes under consideration, namely MSA and PSP. Furthermore, in that study, ~14% of the scans that were classified as indeterminate at the first level (13/96 and 11/71 in the PD and APS classes, respectively) on the basis of criteria defined on the same dataset were discarded from the accuracy analysis.

This double dipping procedure employed by Tang et al. (2010) (i.e., using non-independent samples for defining and applying a given threshold) is questionable because it introduces some circularity (Kriegeskorte et al., 2009). Here, all scans were considered for computing prediction accuracy estimates making difficult a strict comparison of classification performance between studies. Both the present approach and that proposed by Tang et al. (2010) should be validated on an independent sample. Furthermore a potential confound common to both studies is the difference in mean age at the time of PET assessment, which was significantly lower in PD than in APS patients (Table 1). As in the study of Tang et al. (2010), the present data did not allow classification accuracy to be statistically adjusted for age effects (Miller and Chapman, 2001). A younger age at PET assessment in the PD class than in the three other classes clearly represents a limitation of the RVM classifiers built here.

Here, we computed an additional quantitative measure that could also be delivered to the clinicians to assist them in their diagnostic process. This classification majority confidence ratio measure takes advantage of the bagging procedure and was computed on the basis of the two most frequent votes assigned to each scan over bagging (Table 2). The procedure used to compute this majority confidence ratio in multiclass RVM provided a qualitative outcome measure that could be delivered to the clinicians, as for the binary case. Specifically in the multiclass case, the majority confidence ratio will reflect how the votes are distributed among the runner-up classes. For example, consider two scenarios where the votes are distributed as 70–10–10–10% and 70–30–0–0%. In both cases, the winning class has the same majority of votes (70%) but the majority confidence ratio is 60% and 40%, respectively. With the 70–30–0–0% votes, the runner-up category is "close" (30% compared to 70%) and could potentially be a valid alternative, while for the 70–10–10–10% votes none of the other classes accumulate relatively large number of votes, making the other classes less plausible. This feeling is reflected by the lower majority confidence ratio (40%) compared to that (60%) of the other voting scenario. The majority confidence ratio could also be interpreted as an estimate of the dilution of the votes among the non-winning classes.

One could also look at the class pairs that were the most frequently assigned to each scan over bagging. In our data, PD and MSA were the most frequent class pairs assigned to the PD patient' scans (data not shown). This suggests that PD and MSA are frequently considered as rival classes by multiclass RVM, an observation consistent with results of clinico-pathological correlations (Hughes et al., 2001; Wenning et al., 1995). In the three APS classes, the closest rival class was another APS class. In other words, APS misclassification is more likely to occur between APS classes than between PD and APS.

The majority vote ratio was linked to the PPV: a higher ratio value was generally associated with a higher PPV (Fig. 2). The majority confidence ratio is nevertheless not an estimate of the PPV and to be truly useful in a practical clinical setting, one would still have to estimate the relationship between these two values. This could probably be achieved empirically from the data, as shown in Fig. 2.

Methodological issues inherent to the application of automated classification algorithms in a clinical setting have been reviewed recently (Kloppel et al., 2012). Several sources of misclassification should be specifically considered here. Despite the relatively large number of PET scans used to build the RVM classifiers, we speculate that the sample size might still be too low to capture the full variability in FDG uptake pattern underlying the clinical heterogeneity of PD and APS. Imbalanced datasets might be associated with lower classification accuracy in the minority classes if this problem is not adequately handled. The approach privileged here was to use stratified bootstrap samples in which the number of scans between classes in the learning sets is balanced across classes (Fig. 1) (Efron and Tibshirani, 1993). While parkinsonism often manifests asymmetrically, the laterality of image features was not taken into account for the image-based classification. For instance, one could speculate that better classification accuracy might be achieved by left–right flipping individual FDG PET scans so that the cerebral hemisphere contralateral to the clinically most affected body side is represented on the same image side in all subjects. This was not performed here because defining an objective criterion for flipping is not applicable. On the one hand, asymmetry of parkinsonism is not always clinically obvious in some APS patients especially in PSP and on the other hand FDG uptake features can be expressed bilaterally, without apparent asymmetry such as in PD.

A clinical misdiagnosis at follow-up might be another source of RVM misclassification but this risk is not equal across the classes under consideration. A definite diagnosis of PD, MSA, PSP and CBS can only made post-mortem on the basis of a neuropathological examination, which was not available here. When the pathological diagnosis is lacking, the application of clinical diagnostic criteria up to the latest available clinical follow-up as in the present study considerably increases diagnosis accuracy. Indeed, crude clinical diagnosis accuracy estimates for PD, MSA and PSP are about 65%, 22% and 17% in the first years after symptom onset, respectively (Osaki et al., 2002; Osaki et al., 2004; Rajput et al., 1991) while, at follow-up, the highest accuracy that can be achieved using clinical criteria is ~90% for PD, 70–75% for MSA and PSP and much lower for CBS (Hughes et al., 2001; Ling et al., 2010). In this retrospective study, we acknowledge that the accuracy of the clinical diagnosis depended heavily on the reliability and completeness of the supplied clinical picture by the treating neurologists. We have no doubt that this methodology is appropriate in PD for whom clinical information was still available on average more than 8 years after PET scanning. In APS who had poorer prognosis and thus shorter follow-up, we acknowledge that some subtle clinical details may have been missed and we cannot exclude the possibility that this may have contributed to some extent to misdiagnosis between APS. Altogether, these data suggests that a clinical misdiagnosis at follow-up (i.e., incorrect standard of truth) probably plays a greater role in RVM misclassification of APS than PD scans.

RVM predictions were based on the one-to-one comparison between the distinctive patterns of resting-state cerebral FDG uptake (Fig. 1) and a bootstrap resampling procedure from which we derived standardized maps identifying the most consistent voxel weights the excess and deficit networks (EN and DN). The relevance maps shown in Fig. 3 are thus unlikely driven by outliers. By convention in the comparisons between PD and APS classes, the EN encompasses brain areas where FDG uptake levels are relatively preserved in the former as compared with the latter.

These ENs are consistent with strictly localized decreases in the level of FDG uptake previously identified using mass univariate analyses in MSA (Antonini et al., 1998; Eckert et al., 2005; Eidelberg et al., 1993; Feng et al., 2008; Ghaemi et al., 2002; Juh et al., 2004; Otsuka et al., 1997), PSP (Eckert et al., 2005; Juh et al., 2004; Klein et al., 2005) and CBS (Eckert et al., 2005; Laureys et al., 1999) with respect to PD or an age-matched normal control population. The ENs in APS also partially overlap with the topographical distribution of neuropathological changes identified using magnetic resonance imaging (MRI) and histopathological studies. For instance, the strong contribution of frontal cortical areas found by multiclass RVM in the EN of PSP and CBS classes but not MSA class is in agreement with the finding that the brains of patients with PSP and CBS have relatively greater pathology in the frontal cortex early in the disease course, while this is marginally present in MSA (Dickson et al., 2010; Schrag et al., 2000; Wenning et al., 1997). Altogether, this provides support to the biological relevance of the EN identified in APS by RVM without any a priori assumption.

The DNs identified by RVM analyses comparing PD and APS subclasses consistently encompass many cortical areas with the notable exception of the dorsal and medial aspects of caudal frontal lobes, which are part of the EN (see above). The most consistent areas of the DN over bootstrap resamplings are the ventral aspects of frontal

lobes and posterior associative cortices including the precuneus, posterior cingulate, occipital and lateral aspects of parietal and temporal cortices (Fig. 3). Decreased FDG uptake in these cortical areas is a consistent finding in studies comparing non-demented PD patients with a normal control population using either univariate analyses (Eckert et al., 2005; Garraux et al., 2011; Hu et al., 2000) or the PCA method (see above) (Ma et al., 2007). As shown in Fig. 3b, RVM consistently identified the occipital cortex in the deficit network of the PD class as compared with MSA, PSP and CBS classes. This is in agreement with results obtained using univariate analysis methods that reported decreased FDG uptake in the occipital cortex as a supportive feature for PD as compared with APS (Hellwig et al., 2012). However, the primary site of supraspinal pathology in PD is in the brainstem while cortical areas are usually not affected by neuropathological abnormalities at least in the first years after the initial diagnosis of PD (Halliday et al., 2008). The pathophysiological basis of this widespread cortical decrease in FDG uptake in non-demented PD patients is currently unknown.

All patients were scanned under their usual medications and there were group differences in L-DOPA equivalent daily doses (LEDD) (Tomlinson et al., 2010) at the time of FDG PET assessment. The effects of chronic administration of antiparkinsonian drugs on regional cerebral FDG uptake are currently unknown. In one study performed in PD patients at an advanced disease stage, decreased FDG uptake in the bilateral ventral/orbital frontal cortex and the thalamus was reported 1 h after an acute challenge of orally administered L-DOPA (Berding et al., 2001). In another study, L-DOPA administered IV was shown to reduce FDG uptake in the putamen, thalamus, and cerebellum in patients with PD (Feigin et al., 2001). To the best of our knowledge, there are no comparable studies in APS who typically show a poor clinical response to antiparkinsonian drugs. To the best of our knowledge, there are no comparable studies on the *chronic* effects of antiparkinsonian medications on resting-state FDG uptake pattern assessed using a *multivariate* approach. The underlying effects, if any, on RVM classification performance and relevance maps are unclear.

In conclusion, these novel results show that a multiclass classification problem, mimicking a situation encountered in a clinical setting, can be adequately addressed by an automatic, single step, one-to-one comparison cerebral FDG uptake patterns using RVM in combination with bootstrap resampling. The method is fast, fully automatic and can be easily implemented in a clinical setting. However, additional methodological milestones should be achieved before the present methods could be fully transferred to the clinic including pattern validation on larger independent cohorts of incident cases involving a standardized, prospective data acquisition, optimization of image preprocessing methods (Merhof et al., 2011) to allow multicenter studies, and optimization of the RVM analysis on an independent sample including a refinement of image-based cut off values used for classification (Kriegeskorte et al., 2009).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2013.06.004.

Disclosure/conflict of interest

The authors have nothing to disclose.

Acknowledgments

CP and GG are respectively Research Associate and Senior Research Associate at the Fonds National de la Recherche Scientifique de Belgique (FRS-FNRS), respectively. JS is a doctoral researcher at the Fonds de la Recherche en Industrie et Agriculture (FRIA), Belgium. The authors are grateful to Prof. Georges Franck for patient referrals and Mrs Aurélie Dessoullières and Annick Claes for their help in data management. This research was supported by FRS-FNRS and a grant from the Rahier Foundation, University of Liège.

References

Antonini, A., Kazumata, K., Feigin, A., Mandel, F., Dhawan, V., Margouleff, C., Eidelberg, D., 1998. Differential diagnosis of parkinsonism with [^{18}F]fluorodeoxyglucose and PET. Movement Disorders 13, 268–274.
Bensimon, G., Ludolph, A., Agid, Y., Vidailhet, M., Payan, C., Leigh, P.N., 2009. Riluzole treatment, survival and diagnostic criteria in Parkinson plus disorders: the NNIPPS study. Brain 132, 156–171.
Berding, G., Odin, P., Brooks, D.J., Nikkhah, G., Matthies, C., Peschel, T., Shing, M., Kolbe, H., van Den, H.J., Fricke, H., Dengler, R., Samii, M., Knapp, W.H., 2001. Resting regional cerebral glucose metabolism in advanced Parkinson's disease studied in the off and on conditions with [(18)F]FDG-PET. Movement Disorders 16, 1014–1022.
Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.
De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.
Dickson, D.W., Ahmed, Z., Algom, A.A., Tsuboi, Y., Josephs, K.A., 2010. Neuropathology of variants of progressive supranuclear palsy. Current Opinion in Neurology 23, 394–400.
Dietterich, T.G., Bakiri, G., 1995. Solving multiclass learning problems via Error-Correcting Output Codes. JAIR 2, 263–286.
Eckert, T., Barnes, A., Dhawan, V., Frucht, S., Gordon, M.F., Feigin, A.S., Eidelberg, D., 2005. FDG PET in the differential diagnosis of parkinsonian disorders. NeuroImage 26, 912–921.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statical accuracy. Statistical Sciences 1, 54–77.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.
Efron, B., Tibshirani, R., 1995. Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical Report of the Division of Biostatistics. Stanford University, p. 176.
Eidelberg, D., Takikawa, S., Moeller, J.R., Dhawan, V., Redington, K., Chaly, T., Robeson, W., Dahl, J.R., Margouleff, D., Fazzini, E., 1993. Striatal hypometabolism distinguishes striatonigral degeneration from Parkinson's disease. Annals of Neurology 33, 518–527.
Feigin, A., Fukuda, M., Dhawan, V., Przedborski, S., Jackson-Lewis, V., Mentis, M.J., Moeller, J.R., Eidelberg, D., 2001. Metabolic correlates of levodopa response in Parkinson's disease. Neurology 57, 2083–2088.
Feng, T., Wang, Y., Ouyang, Q., Duan, Z., Li, W., Lu, L., Xiang, W., 2008. Comparison of cerebral glucose metabolism between multiple system atrophy Parkinsonian type and Parkinson's disease. Neurological Research 30, 377–382.
Friston, K.J., Frith, C.D., Liddle, P.F., Dolan, R.J., Lammertsma, A.A., Frackowiak, R.S., 1990. The relationship between global and local changes in PET scans. Journal of Cerebral Blood Flow and Metabolism 10, 458–466.
Fürnkranz, J., 2002. Round Robin classification. Journal of Machine Learning Research 2, 747.
Garraux, G., Salmon, E., Peigneux, P., Kreisler, A., Degueldre, C., Lemaire, C., Destee, A., Franck, G., 2000. Voxel-based distribution of metabolic impairment in corticobasal degeneration. Movement Disorders 15, 894–904.
Garraux, G., Bahri, M.A., Lemaire, C., Degueldre, C., Salmon, E., Kaschten, B., 2011. Brain energization in response to deep brain stimulation of subthalamic nuclei in Parkinson's disease. Journal of Cerebral Blood Flow and Metabolism 31, 1612–1622.
Ghaemi, M., Hilker, R., Rudolf, J., Sobesky, J., Heiss, W.D., 2002. Differentiating multiple system atrophy from Parkinson's disease: contribution of striatal and midbrain MRI volumetry and multi-tracer PET imaging. Journal of Neurology, Neurosurgery, and Psychiatry 73, 517–523.
Gibb, W.R., 1988. Accuracy in the clinical diagnosis of parkinsonian syndromes. Postgraduate Medical Journal 64, 345–351.
Gonen, M., Alpaydin, E., 2011. Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268.
Halliday, G., Hely, M., Reid, W., Morris, J., 2008. The progression of pathology in longitudinally followed patients with Parkinson's disease. Acta Neuropathologica 115, 409–415.
Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A., 2009. Decoding neuronal ensembles in the human hippocampus. Current Biology 19, 546–554.
Hellwig, S., Amtage, F., Kreft, A., Buchert, R., Winz, O.H., Vach, W., Spehl, T.S., Rijntjes, M., Hellwig, B., Weiller, C., Winkler, C., Weber, W.A., Tuscher, O., Meyer, P.T., 2012. [(1)(8)F]FDG-PET is superior to [(1)(2)(3)I]IBZM-SPECT for the differential diagnosis of parkinsonism. Neurology 79, 1314–1322.
Hu, M.T.M., Taylor-Robinson, S.D., Chaudhuri, K.R., Bell, J.D., Labbé, C., Cunningham, D.A., Koepp, M.J., Hammers, A., Morris, R.G., Turjanski, N., Brooks, D.J., 2000. Cortical dysfunction in non-demented Parkinson's disease patients: a combined (31)P-MRS and (18)FDG-PET study. Brain 123, 340–352.
Hughes, A.J., Ben Shlomo, Y., Daniel, S.E., Lees, A.J., 1992. What features improve the accuracy of clinical diagnosis in Parkinson's disease: a clinicopathologic study. Neurology 42, 1142–1146.
Hughes, A.J., Daniel, S.E., Lees, A.J., 2001. Improved accuracy of clinical diagnosis of Lewy body Parkinson's disease. Neurology 57, 1497–1499.
Hughes, A.J., Daniel, S.E., Ben Shlomo, Y., Lees, A.J., 2002. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain 125, 861–870.
Juh, R., Kim, J., Moon, D., Choe, B., Suh, T., 2004. Different metabolic patterns analysis of Parkinsonism on the ^{18}F-FDG PET. European Journal of Radiology 51, 223–233.
Klein, R.C., de Jong, B.M., de Vries, J.J., Leenders, K.L., 2005. Direct comparison between regional cerebral metabolism in progressive supranuclear palsy and Parkinson's disease. Movement Disorders 20, 1021–1030.
Kloppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourao-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. NeuroImage 61, 457–463.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 12, 535–540.
Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J., 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 957–968.

Lang, A.E., Riley, D.E., Bergeron, C., 1994. Cortico-basal ganglionic degeneration. In: Calne, D.B. (Ed.), Neurodegenerative Diseases. W.B. Saunders, Philadelphia, pp. 877–894.
Laureys, S., Salmon, E., Garraux, G., Peigneux, P., Lemaire, C., Degueldre, C., Franck, G., 1999. Fluorodopa uptake and glucose metabolism in early stages of corticobasal degeneration. Journal of Neurology 246, 1151–1158.
Ling, H., O'Sullivan, S.S., Holton, J.L., Revesz, T., Massey, L.A., Williams, D.R., Paviour, D.C., Lees, A.J., 2010. Does corticobasal degeneration exist? A clinicopathological re-evaluation. Brain 133, 2045–2057.
Ma, Y., Tang, C., Spetsieris, P.G., Dhawan, V., Eidelberg, D., 2007. Abnormal metabolic network activity in Parkinson's disease: test–retest reproducibility. Journal of Cerebral Blood Flow and Metabolism 27, 597–605.
Merhof, D., Markiewicz, P.J., Platsch, G., Declerck, J., Weih, M., Kornhuber, J., Kuwert, T., Matthews, J.C., Herholz, K., 2011. Optimized data preprocessing for multivariate analysis applied to 99mTc-ECD SPECT data sets of Alzheimer's patients and asymptomatic controls. Journal of Cerebral Blood Flow and Metabolism 31, 371–383.
Miller, G.A., Chapman, J.P., 2001. Misunderstanding analysis of covariance. Journal of Abnormal Psychology 110, 40–48.
Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage 33, 1055–1065.
Osaki, Y., Ben Shlomo, Y., Wenning, G.K., Daniel, S.E., Hughes, A., Lees, A.J., Mathias, C.J., Quinn, N., 2002. Do published criteria improve clinical diagnostic accuracy in multiple system atrophy? Neurology 59, 1486–1491.
Osaki, Y., Ben Shlomo, Y., Lees, A.J., Daniel, S.E., Colosimo, C., Wenning, G.K., Quinn, N., 2004. Accuracy of clinical diagnosis of progressive supranuclear palsy. Movement Disorders 19, 181–189.
Otsuka, M., Kuwabara, Y., Ichiya, Y., Hosokawa, S., Sasaki, M., Yoshida, T., Fukumura, T., Kato, M., Masuda, K., 1997. Differentiating between multiple system atrophy and Parkinson's disease by positron emission tomography with ^{18}F-dopa and ^{18}F-FDG. Annals of Nuclear Medicine 11, 251–257.
Phillips, C.L., Bruno, M.A., Maquet, P., Boly, M., Noirhomme, Q., Schnakers, C., Vanhaudenhuyse, A., Bonjean, M., Hustinx, R., Moonen, G., Luxen, A., Laureys, S., 2011. "Relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. NeuroImage 56, 797–808.
Rajput, A.H., Rozdilsky, B., Rajput, A., 1991. Accuracy of clinical diagnosis in parkinsonism—a prospective study. Canadian Journal of Neurological Sciences 18, 275–278.
Schrag, A., Good, C.D., Miszkiel, K., Morris, H.R., Mathias, C.J., Lees, A.J., Quinn, N.P., 2000. Differentiation of atypical parkinsonism with routine MRI. Neurology 54, 697–702.
Schrouff, J., Kusse, C., Wehenkel, L., Maquet, P., Phillips, C., 2012. Decoding semi-constrained brain activity from FMRI using support vector machines and gaussian processes. PloS One 7, e35860.
Spetsieris, P.G., Ma, Y., Dhawan, V., Eidelberg, D., 2009. Differential diagnosis of parkinsonian syndromes using PCA-based functional imaging features. NeuroImage 45, 1241–1252.
Tang, C.C., Poston, K.L., Eckert, T., Feigin, A., Frucht, S., Gudesblatt, M., Dhawan, V., Lesser, M., Vonsattel, J.P., Fahn, S., Eidelberg, D., 2010. Differential diagnosis of parkinsonism: a metabolic imaging study using pattern analysis. Lancet Neurology 9, 149–158.
Teune, L.K., Bartels, A.L., de Jong, B.M., Willemsen, A.T.M., Eshuis, S.A., de Vries, J.J., van Oostrom, J.C.H., Leenders, K.L., 2010. Typical cerebral metabolic patterns in neurodegenerative brain diseases. Movement Disorders 25, 2395–2404.
Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 211–244.
Tomlinson, C.L., Stowe, R., Patel, S., Rick, C., Gray, R., Clarke, C.E., 2010. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. Movement Disorders 25, 2649–2653.
Varrone, A., Asenbaum, S., Vander Borght, T., Booij, J., Nobili, F., Nagren, K., Darcourt, J., Kapucu, O.L., Tatsch, K., Bartenstein, P., Van Laere, K., 2009. EANM procedure guidelines for PET brain imaging using [F-18]FDG, version 2. European Journal of Nuclear Medicine and Molecular Imaging 36, 2103–2110.
Warr, L., Walker, Z., 2012. Identification of biomarkers in Lewy-body disorders. The Quarterly Journal of Nuclear Medicine and Molecular Imaging 56, 39–54.
Wellcome Trust Centre for Neuroimaging, 2008. University College London, UK. http://www.fil.ion.ucl.ac.uk/spm/.
Wenning, G.K., Ben-Shlomo, Y., Magalhaes, M., Daniel, S.E., Quinn, N.P., 1995. Clinicopathological study of 35 cases of multiple system atrophy. Journal of Neurology, Neurosurgery, and Psychiatry 58, 160–166.
Wenning, G.K., Tison, F., Ben Shlomo, Y., Daniel, S.E., Quinn, N.P., 1997. Multiple system atrophy: a review of 203 pathologically proven cases. Movement Disorders 12, 133–147.