

# Self-normalized Classification of Parkinson's Disease DaTscan Images

Yuan Zhou

Department of Radiology and Biomedical Imaging  
Yale University  
New Haven, CT, USA  
zhouyuanzxcv@gmail.com

Hemant D. Tagare

Department of Radiology and Biomedical Imaging  
Yale University  
New Haven, CT, USA  
hemant.tagare@yale.edu

**Abstract**—Classifying SPECT images requires a preprocessing step which normalizes the images using a normalization region. The choice of the normalization region is not standard, and using different normalization regions introduces normalization region-dependent variability. This paper mathematically analyzes the effect of the normalization region to show that normalized-classification is exactly equivalent to a subspace separation of the half rays of the images under multiplicative equivalence. Using this geometry, a new self-normalized classification strategy is proposed. This strategy eliminates the normalizing region altogether. The theory is used to classify DaTscan images of 365 Parkinson's disease (PD) subjects and 208 healthy control (HC) subjects from the Parkinson's Progression Marker Initiative (PPMI). The theory is also used to understand PD progression from baseline to year 4.

**Index Terms**—Image Classification, Machine Learning, PET/SPECT, DaTscan, Parkinson's Disease.

## I. INTRODUCTION

Clinical SPECT (and PET) images are often normalized before classification [1], [2]. Typically, a *normalization region* with nonspecific tracer binding is chosen, and its mean  $\mu$  is used to calculate the *binding potential* (BP), defined as  $BP(v) = (I(v) - \mu)/\mu = (I(v)/\mu) - 1$  for every voxel  $v$  in an image  $I$  [3]. All subsequent image classification is done using BP rather than the original image. Many different normalization regions are used in practice, however, they are not all equivalent [4], [5], [6], [7], [8]; changing the normalization region changes the downstream results.

Given this dependence on the normalization region, one may ask whether SPECT/PET images can be classified without choosing a normalization region. The goal of this paper is to show that this is possible, without sacrificing accuracy. We begin in Section III by mathematically analyzing how normalization affects classification. Such a theory has not yet appeared in the literature. Based on this analysis, in Section III-D we propose a new classification strategy which does not require a normalization region. Instead, the classification is *self-normalizing*. A potential pitfall of not using a normalization region is the possible loss of classification accuracy. In Section IV, we show that there is no loss of classification accuracy when self-normalizing classification is used with real-world data.

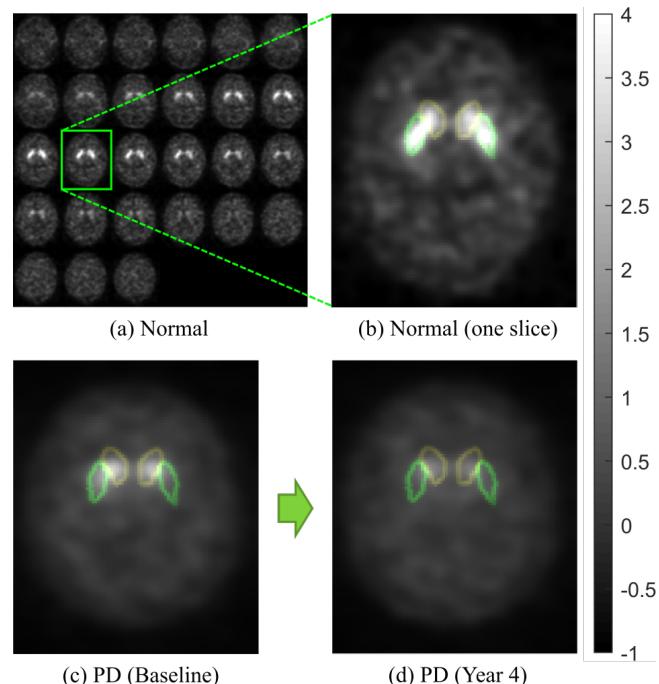


Fig. 1. DaTscan images of a normal subject (a,b) and a PD patient (c,d). Boundaries of caudate and putamen are delineated in yellow and green respectively. Compared to healthy controls, PD patients have reduced signal in the caudate and putamen, and the reduction becomes worse over time. The images were normalized by the mean in the occipital lobe.

For real-world data, we use SPECT images of Parkinson's disease (PD). Imaging PD with [ $^{123}\text{I}$ ]-Ioflupane, commonly called DaTscan imaging, measures the concentration of the dopamine transporter (DaT) protein. Dopaminergic neuronal loss in PD is visible as loss of signal in the putamen and the caudate in DaTscan images (see Fig. 1). The occipital lobe usually serves as the normalization region [9], [10], [11], although the cerebellum [12], and the whole brain except the striatum [10] are also used. As mentioned above, different normalization regions influence the ability of BP (which is called the *striatal binding ratio* (SBR) in PD DaTscans) to classify PD vs. healthy controls (HC) [5].

We carried out two classification experiments on PD DaTscan images to validate our theory. First, we classify

DaTscan images of PD from HC subjects. Second, we classify longitudinal DaTscan images of PD patients obtained 4 years apart. As explained in Section IV, the latter provides insight into how PD progresses during this interval. For each classification problem, we compare the classification accuracy between the classical normalization and our proposed self-normalization, using both linear and non-linear classifiers.

## II. BACKGROUND AND LITERATURE REVIEW

### A. PET/SPECT Normalization

PET/SPECT imaging can be categorized into dynamic and static imaging. Dynamic imaging acquires images at multiple time points after tracer injection and is mainly used in research. Static imaging acquires a single image over a fixed period hence is suitable for clinical applications. Analyzing clinical (static) PET/SPECT images typically requires a preprocessing step which normalizes its intensity [1], [2]. Normalization is required because the amount of radioligand reaching the brain depends on multiple factors, including age, sex, and medication. This results in an unknown scaling factor for each image. Standard normalization selects a normalization region to calculate the BP, thereby eliminating this scaling factor, as mentioned before. Note that in PD DaTscans, the BP is called the *striatal binding ratio* (SBR) since dopamine transporters mainly exist in the striatum.

In PD DaTscans, the occipital lobe is usually selected as the normalization region since it contains few dopamine transporters [9], [10], [11] and most of the binding of the radiotracer in this region is non-specific. Other choices for the normalization region include the cerebellum [12] and the whole brain except the striatum [10]. Different normalization regions alter the predictive power of BP to classify PD, up to a difference of 0.147 in terms of area under the curve [5].

For PET/SPECT imaging using other tracers, e.g. [<sup>18</sup>F]fluorodeoxyglucose for visualizing glucose (FDG-PET), the impact of the normalization regions is also significant [7], [8].

Besides the BP normalization mentioned above, other normalization techniques have also been proposed. For DaTscans, these include analyzing the distribution of intensity values in the whole brain except the striatum [13], and minimizing the squared error within a selected region between a template and the linear transformed image [1].

### B. DaTscan Classification

Classification of DaTscan images into PD and HC has been achieved using standard machine learning techniques [14], [15], [10], [11] as well as deep learning methods [16]. Early classification work tends to use a small proprietary datasets and perform their own registration [17], [14], while later work has shifted to using a public dataset: the Parkinson's Progression Marker Initiative (PPMI) dataset which provides already registered DaTscans [15], [11], [18].

Support vector machine (SVM) and logistic regression are probably the most popular PD vs. HC classifiers, appearing in most of the non-deep-learning studies [14], [15], [10], [11], [18], [19]. Graph-based transductive learning was introduced

for classifying multi-modality neurodegenerative image data in [20]. Recently, convolutional neural networks (CNN) have been applied to classifying DaTscan images for PD diagnosis [21], [16], [22]. Most of these methods calculate the SBR using a normalization region as a pre-processing step before classification [15], [11], [18].

Other studies use geometric image features including the length and volume of the segmented striatum [23], shape fitting coefficients [24], [25], isosurfaces [16], and intensity summary statistics [22].

## III. NORMALIZATION AND CLASSIFICATION

We now turn to explaining the effect of normalization on classification. The effect is most easily explained using linear classification, and we stick to linear classification for most of this section. However, non-linear classification is also addressed.

To begin, note that images differing by a multiplicative factor, such as  $I_2 = \alpha I_1$  for  $\alpha > 0$ , give the same BP. The  $-1$  term in the BP simply adds a fixed constant to every voxel, and has no effect on classification. We ignore this term.

### A. Multiplicative Equivalence

Let  $\Omega$  be the set of voxels in an image, with the total number of voxels being  $d$ . Any nonzero image  $I$  defined on  $\Omega$  is an element of  $\mathbb{R}^d$ . SPECT/PET images have non-negative voxels, i.e. these images lie in the non-negative orthant of  $\mathbb{R}^d$  with the origin removed. Geometrically speaking, the set of all images related by positive scalar multiples is a half-ray passing through the origin of  $\mathbb{R}^d$  (see Fig. 2(a)). We denote the half-ray of the image  $I$  by  $[I]$ . We also denote the set of all half-rays passing through the non-negative orthant of  $\mathbb{R}^d$  (with the origin removed) as  $\mathcal{I}$ . Classifying images under multiplicative equivalence means partitioning  $\mathcal{I}$  into disjoint subsets.

### B. Normalized-Classification

A normalization region is a set of voxels  $N \subset \Omega$ . All images restricted to  $N$  form a subspace of  $\mathbb{R}^d$ , which we denote as  $\Pi_N$ . The projection operator  $\pi_N : \mathbb{R}^d \rightarrow \Pi_N$  projects an image onto this subspace by setting values outside  $N$  to zero.

Let  $\mathbf{1} = [1, 1, \dots, 1]^T \in \mathbb{R}^d$  denote an image with all 1's. Then  $\pi_N(\mathbf{1})$  is the image containing all 1's in the voxels  $N$  and 0's everywhere else. For any image  $I$ , the mean value in the normalization region is  $\frac{1}{|N|}(\pi_N(\mathbf{1}))^T I$ , where  $|N|$  is the number of voxels in  $N$ . Denoting  $\frac{1}{|N|}\pi_N(\mathbf{1})$  by  $\mathbf{1}_N$ , the mean value in the normalization region is simply  $\mathbf{1}_N^T I$  and the normalized image is  $\tilde{I} = I / \mathbf{1}_N^T I$ . Normalization can be interpreted geometrically after noting that the mean value of  $\tilde{I}$  in the normalization region is always 1, since  $\mathbf{1}_N^T \tilde{I} = \mathbf{1}_N^T (I / \mathbf{1}_N^T I) = 1$ . Thus the geometry of normalization is explained as follows (see Fig. 2(b)): From the image  $I$  construct the half-ray  $[I]$ . Then the normalized image  $\tilde{I}$  is the intersection of the half ray  $[I]$  with  $\mathcal{A}_N$ , the  $d-1$  dimensional affine subspace defined by  $\mathcal{A}_N = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{1}_N^T \mathbf{x} = 1\}$ .

Once the image is normalized, it is classified using voxels in another region  $C \subset \Omega$ . We assume that  $N \cap C = \emptyset$ , i.e. the

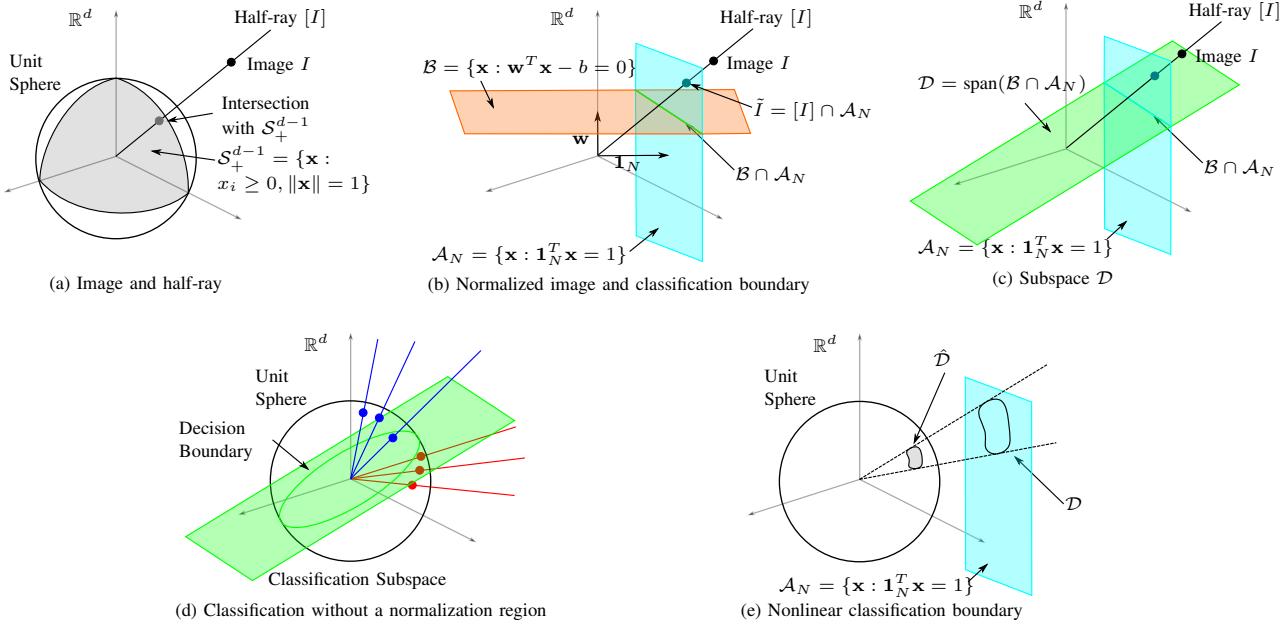


Fig. 2. The geometry of classical normalized-classification and our self-normalized classification. (a) Standard normalization for clinical PET/SPECT images means that images differing by a scaling factor are equivalent. All such images form a half-ray  $[I]$ . (b) Normalization is equivalent to finding the intersection between  $[I]$  and affine subspace  $\mathcal{A}_N$ . Linear classification of normalized images (normalized-classification) has the classification boundary  $\mathcal{B} \cap \mathcal{A}_N$ . (c) Normalized-classification using  $\mathcal{B} \cap \mathcal{A}_N$  is equivalent to classifying the half-rays using a subspace  $\mathcal{D}$ . (d) Proposed self-normalized classification: project the voxels onto the unit sphere and use an affine subspace to separate them. (e) The idea extends to nonlinear classifiers since nonlinear boundaries in  $\mathcal{A}_N$  also correspond to nonlinear boundaries on the unit sphere.

normalization and classification voxels are disjoint. Similar to  $N$  above, images restricted to  $C$  form the subspace  $\Pi_C$  of  $\mathbb{R}^d$ . Because  $N$  and  $C$  are disjoint, we have  $\Pi_C \perp \Pi_N$ .

A linear classifier chooses a unit norm vector  $w \in \Pi_C$  and a real number  $b$  to define a decision boundary  $w^T \tilde{I} - b = 0$ . We refer to linear classification using the normalized image  $\tilde{I}$  as *normalized-classification*, and its boundary as the *normalized-classification boundary*. This boundary has an alternate description. First consider the equation  $w^T x - b = 0$  for  $x \in \mathbb{R}^d$ . Because  $w \neq 0$ , this equation describes a  $d-1$  dimensional affine subspace of  $\mathbb{R}^d$ . Call it  $\mathcal{B}$  (see Fig. 2(b)). Then the normalized-classification boundary  $w^T \tilde{I} - b = 0$  is the intersection of  $\mathcal{B}$  and  $\mathcal{A}_N$  (see Fig. 2(b)), which is the set of all  $x \in \mathbb{R}^d$  satisfying

$$\begin{pmatrix} w^T \\ 1_N^T \end{pmatrix} x = \begin{pmatrix} b \\ 1 \end{pmatrix}. \quad (1)$$

Because  $w$  and  $1_N$  are linearly independent,  $\mathcal{B} \cap \mathcal{A}_N$  is a  $d-2$  dimensional affine subspace of  $\mathbb{R}^d$ .

### C. From Normalized-Classification to Subspace Classification

A slight shift in point-of-view shows that normalized-classification is really just a classification of half-rays by subspaces. The key idea here is to take  $\mathcal{D} = \text{span}(\mathcal{B} \cap \mathcal{A}_N)$  (see Fig. 2(c)). Because  $\mathcal{D}$  is a span of a set of points,  $\mathcal{D}$  is a subspace of  $\mathbb{R}^d$ . It has properties that are described below. Proofs of all properties can be found in the Appendix.

**Claim 1:** The subspace  $\mathcal{D}$  has dimension  $d-1$  and is the set of all  $x \in \mathbb{R}^d$  satisfying  $(w - b1_N)^T x = 0$ .

Because  $\mathcal{D}$  is a subspace, any half-ray through the origin of  $\mathbb{R}^d$  stays exactly on one side of it, or is contained in it. Thus:

**Claim 2:** Every normalized-classification of images is completely equivalent to a classification of the half-rays by the subspace  $(w - b1_N)^T x = 0$ .

In other words, the classification of normalized images  $\tilde{I}$  is exactly the same as the classification of half-rays  $[I]$  by  $\mathcal{D}$  (see Fig. 2(c)). We call the subspace  $\mathcal{D}$ , a *classification subspace*, and classification of half-rays it achieves *subspace classification*. Recalling that a normalized-classification boundary is determined by the pair  $w, b$ , further analysis shows that

**Claim 3:** Every distinct normalized-classification boundary pair  $w, b$  gives a unique classification subspace  $\mathcal{D}$ .

In other words, normalized-classification of images is completely equivalent to subspace classification of half-rays.

Next, we turn to ask: What happens to normalized-classification if the normalization region is changed? Specifically, suppose that the classification region  $C$  remains fixed, but we have two different normalization regions  $N \neq N'$ , with corresponding normalized-classification boundaries given by  $w, b$  and  $w', b'$  respectively. Because normalized-classification is equivalent to subspace classification of  $\mathcal{I}$ , we ask when the

corresponding classification subspaces  $\mathcal{D}, \mathcal{D}'$  are the same? The result is:

**Claim 4:**  $\mathcal{D} = \mathcal{D}'$  if and only if  $\mathbf{w} = \mathbf{w}'$  and  $b = b' = 0$ .

Claim 4 shows that changing the normalization region can give the same classification subspace only in very special cases where  $b = b' = 0$ . Classification with  $b = 0$  rarely happens in real-world classification. Claim 4 explains why changing the normalization region changes the classification.

#### D. Self-normalized Classification

The above analysis strongly suggests how the dependence on the normalization region can be eliminated altogether: Take the intersection of each half-ray and the unit sphere in  $\mathbb{R}^d$  (see Fig. 2(d)) and then classify the intersection points with a subspace. There is no normalization region involved in this; the data normalizes itself — it is *self-normalizing*.

This idea can be pushed further in two ways: First, we need not classify the intersection points on the sphere with only a subspace. We can use any  $d - 1$  dimensional affine subspace. Second, we can apply the same idea to the image restricted to the classification voxels  $C$ . That is, we take only intensities of voxels in  $C$ , project them to the unit sphere, and classify them using an affine subspace. This too eliminates the normalization region.

#### E. Non-linear Classification

The above ideas extend easily to nonlinear classification. To see how, first note that the normalization step is the same as before. It corresponds to moving the image  $I$  to the intersection of the half-ray  $[I]$  and  $\mathcal{A}_N$ . However, the classification boundary, which was the affine subspace  $\mathcal{B} \cap \mathcal{A}_N$  in Fig. 2, is no longer linear. Suppose it is a  $d - 2$  dimensional sub-manifold of  $\mathcal{A}_N$ , which we will call  $\mathcal{D}$  (see Fig. 2(e)). Every  $\mathbf{x} \in \mathcal{D}$  gives a half-ray  $[\mathbf{x}]$  and this half ray intersects the unit sphere at a single point. Because  $\mathcal{D}$  is a  $d - 2$  dimensional submanifold, the set of all half rays of points in  $\mathcal{D}$  is a  $d - 1$  dimensional submanifold of  $\mathbb{R}^d$ , and this manifold intersects the unit sphere transversely to give a  $d - 2$  dimensional submanifold of the unit sphere. Denote this submanifold as  $\hat{\mathcal{D}}$  (see Fig. 2(e)). It is straightforward to see that the partition (i.e. classification) of  $\mathcal{I}$  by  $\mathcal{D}$  is identical to the partition of  $\mathcal{I}$  by  $\hat{\mathcal{D}}$ . The converse is also straightforward: By simply reversing the argument it is easy to see that any classification boundary, which is a  $d - 2$  dimensional submanifold of the part of the unit sphere in the non-negative orthant of  $\mathbb{R}^d$ , induces a  $d - 2$  dimensional submanifold as a classification boundary in  $\mathcal{A}_N$ . Thus, we have

**Claim 5:** Every non-linear classification boundary in  $\mathcal{A}_N$  which is a  $d - 2$  dimensional submanifold is equivalent to a classification boundary on the unit sphere which is also a  $d - 2$  dimensional submanifold. The converse is also true for all decision boundaries that are  $d - 2$  dimensional submanifolds in the non-negative part of the unit sphere.

The theory developed so far suggests that normalized classification and self-normalized classification are equivalent. In other words, there should be no loss of classification accuracy with self-normalized classification. Whether this holds in practice is addressed in the next Section using the PPMI DaTscan dataset. Because this is a PD DaTscan dataset, we refer to BP as SBR from now on.

## IV. NUMERICAL RESULTS

### A. PPMI Data

The PPMI dataset contains 449 early-stage PD subjects and 210 HC subjects. The PD subjects have scans at baseline, and at approximately 1, 2, 4, and 5 years from baseline, with missing scans. Most of the HC subjects have only a single scan. The images have a size of  $109 \times 91 \times 91$  voxels, with  $2 \text{ mm}^3$  voxels. The images are already registered by PPMI to the Montreal Neurological Institute (MNI) atlas. However, following the procedure in [26], we found and removed some misregistered images. This left us with 365 PD subjects (ages:  $62.6 \pm 9.8$  years, male/female: 237/128) and 208 HC subjects (ages:  $60.6 \pm 11.2$  years, male/female: 135/73) to analyze. All PD subjects had baseline images, but only 136 PD subjects had images at year 4.

Because PD affects the two brain hemispheres asymmetrically, the DaTscan images were flipped around the mid-plane so that the more affected side was on the right. For normalized-classification, we used the occipital lobe as the normalization region ( $N$ ) and the striatum as the classification region ( $C$ ). The striatum mask was derived by applying Otsu's threshold [27] on the mean HC image to remove the background and then applying it again on the remaining voxels to remove the nonspecific binding voxels. The occipital lobe mask was taken from [26]. All masks were restricted to the 29–55th slices.

### B. Experimental Setup

The ultimate goal of our experiments is to verify the classification accuracy of the self-normalized classification of Section III-D. However, an important subgoal is to determine whether self-normalized classification provided useful information about voxels that are important to classification. This can be achieved by analyzing the weights for linear classifiers and the saliency maps [28] for non-linear classifiers.

We carried out two classifications: 1) HC vs. PD using only the baseline images, 2) Baseline PDs vs. PDs at year 4. The PD vs. HC classification has obvious clinical importance. The baseline vs year 4 classification is not clinically useful on its own, but because it identifies voxels where the disease progresses from baseline to year 4, it provides a simple disease progression footprint.

In each classification problem, we used three normalization strategies:

- 1) classical normalization using SBR with occipital lobe normalization (referred to as *SBR* from now on),
- 2) self-normalization via a projection of all voxels from the striatum and the occipital lobe onto the unit sphere (referred to as *S + O*),

- 3) self-normalization via a projection of voxels only from the striatum to the unit sphere (referred to as  $S$ ).

To avoid numerical underflow problems, the radius of the sphere used in self-normalization was set to  $\sqrt{d}$ , where  $d$  is the dimension (number of voxels) of the region being projected. The dimension  $d$  for the above three strategies are  $d = 9948, 43077$ , and  $9948$  respectively.

Along with each normalization, we used three classifiers, two of which were linear: logistic regression (LR) and SVM, and one of which was nonlinear: convolutional neural network (CNN). The linear classifiers had a sparsity constraint as implemented in the *fitclm* function in Matlab with a *sparsa* optimizer [29]. The nonlinear classifier was implemented in PyTorch with 2 convolutional/pooling layers (kernel size  $5 \times 5 \times 5$ , 6 and 16 feature maps respectively) followed by 2 linear layers (120 hidden nodes). We used the ReLU activation function after the convolutional/linear layers. For the CNN, voxels were inserted into a 3D image cube padded by zeros. The combination of normalization method with the classifier led to  $3 \times 3 = 9$  classification experiments, whose results are reported below.

For every classification task, we randomly split the data set 100 times into a training set (80%) and a test set (20%). The sparsity parameter for the linear classifiers was chosen using 10-fold cross validation on the 100 training sets. Fig. 3 shows the cross validation results of PD vs. HC. Similar curves (not shown) were obtained for PD baseline vs. PD year 4 classification. Because CNN training is computationally expensive, we did not use cross validation to set the hyperparameters for the nonlinear classifier. Instead, the learning rate was set manually to 0.01 and the number of epochs was set to 500. We used a scheduler which reduced the learning rate by half after half of the epochs. Furthermore, 10% of the training set was kept as the validation set and the remaining was used to train the neural network. We pick the network parameters with the highest accuracy (over the epochs) on the validation set for test.

### C. Classification Accuracy

The classification accuracies of the 100 training/test sets for PD vs. HC and PD baseline vs. PD year 4 are shown in Fig. 4. For PD vs. HC, classical SBR (*SBR*), the self-normalized striatum plus occipital lobe voxels (*S + O*), and the self-normalized striatum voxels (*S*) lead to almost identical results for the test set (see Fig. 4(a) and Table I). The classification accuracies for the test set are quite high (the classification accuracies for the training set are similar). In fact, the classification accuracies in Table I are noticeably higher than those reported in the literature that use the same PPMI data (typical reported accuracies are in the range 95.1–97.9%) [15], [11], [19], [23], [16].

To compare the classification accuracies of standard SBR vs. self-normalized methods, we calculated  $p$ -values from a  $t$ -test comparing the mean classification accuracy of the self-normalizing methods to that of standard SBR. The  $p$ -values

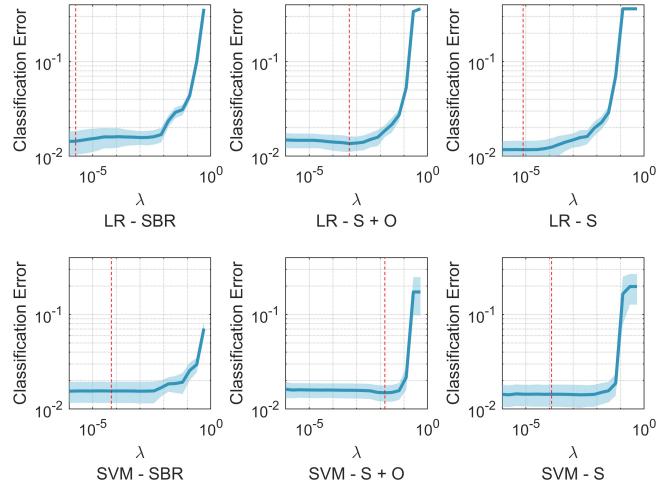


Fig. 3. Cross validation of the linear classifiers for PD vs. HC. The central blue curve shows mean classification error over 100 runs for sparsity parameter  $\lambda$  ranging from  $2^{-20}$  to  $2^{-1}$ . The shaded area shows the corresponding standard deviation. The vertical red dashed lines indicate the chosen  $\lambda$ 's. LR: Logistic Regression, SVM: Support Vector Machine, SBR: Striatum Binding Ratio, S + O: Striatum + Occipital lobe, S: Striatum.

TABLE I  
TEST ACCURACIES OVER 100 RUNS FOR HC VS. PD AND  $p$ -VALUES FROM  $t$ -TEST FOR DIFFERENCE SIGNIFICANCE

	Mean (std) of accuracies (%)			$p$ -value (vs. SBR)	
	SBR	S + O	S	S + O	S
LR	98.77 (0.97)	98.81 (0.91)	98.94 (0.86)	0.744	0.181
SVM	98.67 (1.02)	98.77 (0.86)	98.69 (0.98)	0.475	0.902
CNN	98.68 (1.07)	98.59 (0.94)	98.41 (1.08)	0.543	0.079

(see Table I) are all significantly above 0.05, showing that there is no significant performance difference between the methods.

Classification accuracies for PD baseline vs. PD year 4 show a similar pattern (see Fig. 4(b) and Table II). While the classification accuracies are not as high as PD vs. HC (this is discussed further in Section V), the differences in the performance of the classifiers are again not statistically significant, except for one case (LR on *S + O*).

### D. Salient Voxels

Recall that part of our goal is to identify salient voxels (voxels which contribute significantly to classification). As mentioned before, classification weights of linear classifiers indicate these voxels. For nonlinear classifiers, the saliency map [28], which calculates the derivative of the logit with respect to the input, can be used for similar purposes. Because

TABLE II  
TEST ACCURACIES OVER 100 RUNS FOR PD BASELINE VS. PD YEAR 4 AND  $p$ -VALUES FROM  $t$ -TEST FOR DIFFERENCE SIGNIFICANCE

	Mean (std) of accuracies (%)			$p$ -value (vs. SBR)	
	SBR	S + O	S	S + O	S
LR	77.74 (3.92)	76.47 (4.17)	77.50 (3.94)	0.028	0.666
SVM	75.29 (4.42)	74.63 (4.69)	76.51 (4.50)	0.307	0.055
CNN	73.02 (4.33)	71.74 (5.38)	71.90 (6.08)	0.066	0.135

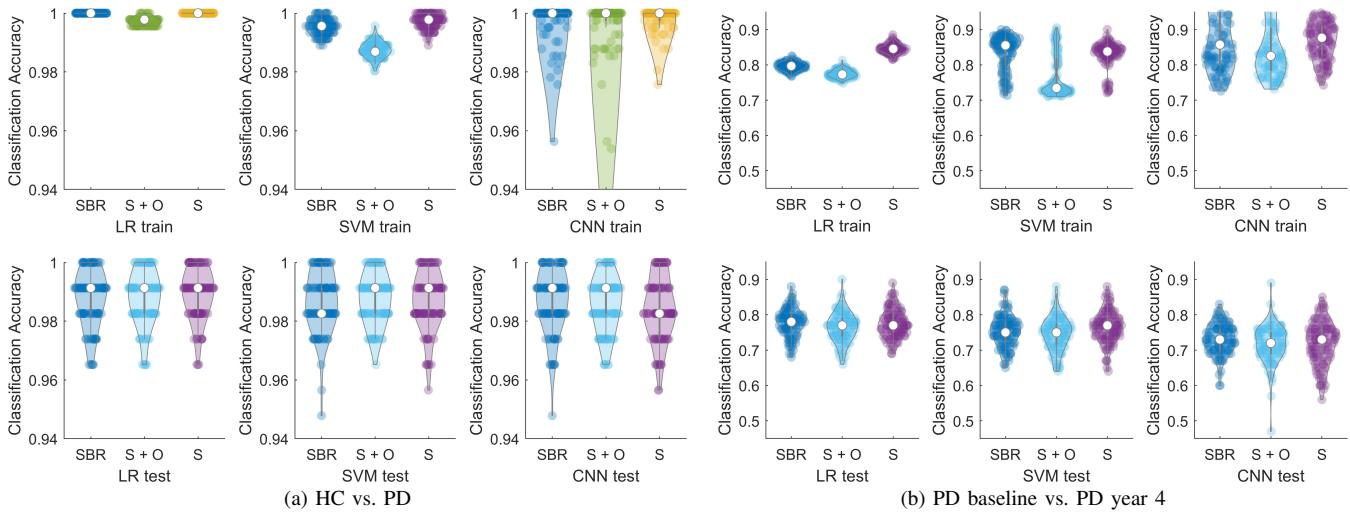


Fig. 4. Violin plots of classification accuracies over 100 runs for PD vs. HC (a) and PD baseline vs. PD year 4 (b). In each subplot, the first row corresponds to the training error and the second corresponds to the test error. LR: Logistic Regression, SVM: Support Vector Machine, CNN: convolutional neural network, SBR: Striatum Binding Ratio, S + O: Striatum + Occipital lobe (projected to the unit sphere), S: Striatum (projected to the unit sphere).

the saliency map is calculated for each test example, we took one train-test split and averaged the saliency maps over all the test examples.

The linear weights and the averaged saliency maps are shown in Fig. 5. Over the 9 combinations of normalizing strategies and classifiers used for each classification task, the results are surprisingly consistent. For HC vs. PD, the negative classifier weights and negative saliency (both rendered in blue) for PD vs. HC are mostly in the putamen on the right side (the more affected putamen). The positive weights and positive saliency is mostly in the left caudate (the least affected side). Thus reduced values in the right putamen relative to the values in the left caudate are significant in classifying PD vs. HC.

PD baseline vs. PD year 4 shows negative coefficients and negative saliency in the putamen on the left side, showing that decreasing values in this putamen corresponds to progression from baseline to year 4.

To make the above observations more quantitative, we evaluated the contribution of each region to the classification. Noting that all classifiers give almost identical salient voxels, we focused on the sparse logistic linear classifier. We restricted  $w$  to each of the four caudate and putamen regions from the MNI atlas, and evaluated  $w^T \hat{I}$  using the restricted  $w$  and self-normalized  $\hat{I}$  for PD vs. HC and PD baseline vs. PD year 4. For PD vs. HC, the mean  $w^T \hat{I}$  has the largest difference when  $w$  is restricted to the right putamen, indicating that this is the region where PD has the largest effect (see Fig. 6). Similarly, for PD baseline vs. PD year 4, we evaluated the mean for PD baseline and year 4. The largest change in the mean is in the left putamen (see Fig. 6).

## V. DISCUSSION AND CONCLUSION

The experiments clearly show that self-normalized classification does not lead to any loss of classification accuracy. In fact, as noted above, the classification accuracies for PD

vs. HC are higher than those reported in the literature. One contributing factor to the increased accuracy is the fact that the images were flipped around the mid-plane so that the more affected side appears to the right.

The lower accuracy of PD baseline vs. year 4 is also understandable. It is estimated that the onset of PD occurs almost 10 years before PD is diagnosed [30]. Thus PD vs. HC images, even at baseline, have about 10 years of accumulated disease evidence for classification. In contrast, baseline vs year 4 has evidence accumulated from less than half of that time. Moreover, the PD images at baseline may correspond to different stages of the disease. Therefore, the two classes are not as easily separable as PD vs. HC.

Finally, the salient voxels identified by the classifiers are quite meaningful. PD is known to start asymmetrically, affecting the putamen in one hemisphere before affecting the caudate. Then, as the disease progresses, it affects the other hemisphere. Consistent with this, our results show a decrease in DaTscan intensity in the right putamen (compared with the left caudate) is significant in classifying baseline PD vs. HC. And subsequent decrease in the left putamen is indicative of longitudinal progression from baseline to year 4.

In summary, the geometry of multiplicative equivalence shows that PET/SPECT images can be classified by self-normalization without any loss of accuracy. This method is effective with linear and non-linear classifiers, and it provides an understanding of disease affecting voxels in the image.

## ACKNOWLEDGMENT

This research was supported by the grant R01NS107328 from the NIH (NINDS). Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database ([www.ppmi-info.org/data](http://www.ppmi-info.org/data)). For up-to-date information on the study, visit [www.ppmi-info.org](http://www.ppmi-info.org). PPMI – a public-private partnership – is funded

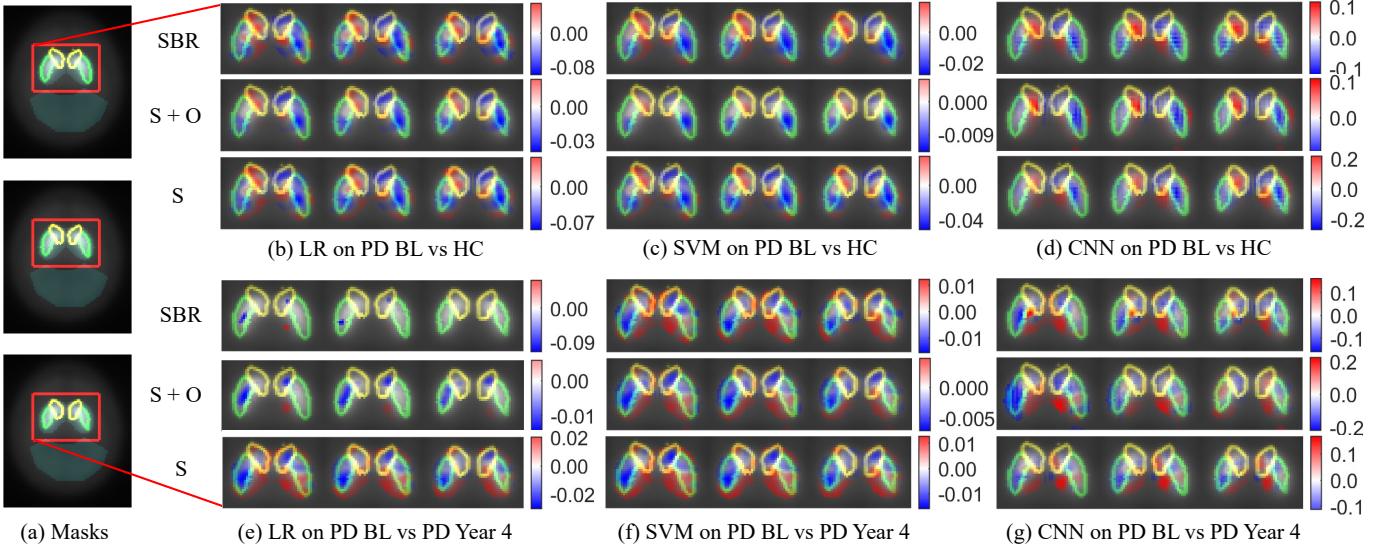


Fig. 5. Linear classifier weights and saliency maps. (a) Masks of caudate (yellow), putamen (green) from the MNI atlas (dilated by 1 voxel), and striatum and occipital lobe (cyan area) for the 40th to 42nd slices. The background shows the mean HC image. (b-d) Weights of LR, SVM and saliency maps of CNN on SBR and sphere projected voxels for PD vs. HC. (e-g) Weights and saliency maps for PD baseline vs. PD year 4. The values of weights/saliency maps are rendered in red and blue colors indicated by the colorbar on the right. Only the 40th to 42nd slices are shown.

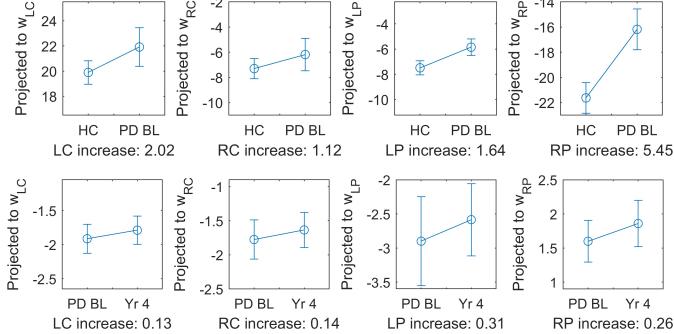


Fig. 6. Mean  $\pm$  standard deviation of  $\mathbf{w}^T \hat{I}$  with  $\hat{I}$  being  $S$  and  $\mathbf{w}$  being the LR weights restricted to LC, RC, LP, RP for PD vs. HC (first row) and PDD baseline vs. PD year 4 (second row). The number below each plot indicates the net change in mean value of  $\mathbf{w}^T \hat{I}$ . It measures the contribution of this region to the decision boundary. LC/LP: Left (less affected) Caudate/Putamen, RC/RP: Right (more affected) Caudate/Putamen.

by the Michael J. Fox Foundation for Parkinson's Research and funding partners. The full list of funding partners can be found at <https://www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/>.

#### APPENDIX: PROOFS OF CLAIMS

**Proof of Claim 1:**  $\mathcal{B}$  and  $\mathcal{A}_N$  are distinct  $d - 1$  dimensional subspaces of  $\mathbb{R}^d$ . Thus their intersection  $\mathcal{B} \cap \mathcal{A}_N$  is non-empty, and is an affine subspace of  $\mathbb{R}^d$ . Moreover,  $\mathcal{B} \cap \mathcal{A}_N$  does not contain the origin (because  $\mathcal{A}_N$  does not) and has dimension  $d - 2$ . Thus its span has dimension  $d - 2 + 1 = d - 1$ , i.e.  $\dim \mathcal{D} = d - 1$ .

Denote the set of all  $\mathbf{x} \in \mathbb{R}^d$  satisfying  $(\mathbf{w} - b\mathbf{1}_N)^T \mathbf{x} = 0$  by  $\mathcal{D}'$ . Because the unit norm vector  $\mathbf{w} \in \Pi_C$ , and the nonzero vector  $\mathbf{1}_N \in \Pi_N$ , and  $\Pi_C \perp \Pi_N$ ,  $\mathbf{w}$  is linearly independent

of  $\mathbf{1}_N$ . Hence the vector  $(\mathbf{w} - b\mathbf{1}_N)$  cannot be 0 for any  $b$ . Thus  $\mathcal{D}'$  is a  $d - 1$  dimensional subspace of  $\mathbb{R}^d$ . We have to show that  $\mathcal{D}'$  is equal to  $\mathcal{D}$ . Any  $\mathbf{x} \in \mathcal{B} \cap \mathcal{A}_N$  satisfies

$$\begin{pmatrix} \mathbf{w}^T \\ \mathbf{1}_N^T \end{pmatrix} \mathbf{x} = \begin{pmatrix} b \\ 1 \end{pmatrix}. \quad (2)$$

Subtracting  $b$  times the second row of the equation from the first shows that  $\mathbf{x}$  also satisfies  $(\mathbf{w} - b\mathbf{1}_N)^T \mathbf{x} = 0$ . It follows that if  $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B} \cap \mathcal{A}_N$ , then  $(\mathbf{w} - b\mathbf{1}_N)^T (\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2) = 0$ , showing that  $\mathcal{D}$ , which is the span of  $\mathcal{B} \cap \mathcal{A}_N$ , is contained in  $\mathcal{D}'$ . But  $\dim \mathcal{D} = \dim \mathcal{D}' = d - 1$ , hence  $\mathcal{D} = \mathcal{D}'$ .  $\square$

**Proof of Claim 2:** All images whose normalized versions lie on one side of  $\mathcal{B}$  have half-rays that lie on one side of  $\mathcal{D}$ .  $\square$

**Proof of Claim 3:** The proof is by contradiction. Suppose we have unit norm vectors  $\mathbf{w}_1, \mathbf{w}_2 \in \Pi_C$  and real numbers  $b_1, b_2$ , such that  $(\mathbf{w}_1, b_1) \neq (\mathbf{w}_2, b_2)$ . And suppose that the classification subspaces for the two are equal. The classification subspaces are given by  $(\mathbf{w}_1 - b_1 \mathbf{1}_N)^T \mathbf{x} = 0$  and  $(\mathbf{w}_2 - b_2 \mathbf{1}_N)^T \mathbf{x} = 0$ . For these subspaces to be the same, there must exist a  $\lambda$  such that  $\lambda(\mathbf{w}_1 - b_1 \mathbf{1}_N) = (\mathbf{w}_2 - b_2 \mathbf{1}_N)$ . Rearranging this equation gives  $(\lambda \mathbf{w}_1 - \mathbf{w}_2) = (\lambda b_1 - b_2) \mathbf{1}_N$ . The term on the left hand side is a linear combination of vectors in  $\Pi_C$ , hence is a vector in  $\Pi_C$ . The term on the right hand side is a vector in  $\Pi_N$ . Since  $\Pi_C \perp \Pi_N$  the equation can hold only if each term is 0. Setting the left hand side to 0 gives  $\lambda \mathbf{w}_1 = \mathbf{w}_2$ . Since  $\mathbf{w}_1$  and  $\mathbf{w}_2$  are unit norm vectors, this implies  $\lambda = 1$  and  $\mathbf{w}_1 = \mathbf{w}_2$ . Similarly, setting the right hand side equal to zero, and using  $\lambda = 1$ , gives  $b_1 = b_2$ . Thus the two classification subspaces are the same if and only if  $\mathbf{w}_1 = \mathbf{w}_2$  and  $b_1 = b_2$ , which contradicts the assumption. Therefore, the subspace classification boundaries

are different, showing that each pair  $w, b$  gives a unique subspace classification boundary.  $\square$

**Proof of Claim 4:** Since the normalization regions  $N \neq N'$ , the vectors  $\mathbf{1}_N$  and  $\mathbf{1}_{N'}$  are linearly independent. Also since  $N$  and  $N'$  are both normalization regions,  $\Pi_N, \Pi_{N'} \perp \Pi_C$ . Thus their direct sum  $\Pi_N \oplus \Pi_{N'}$  is also orthogonal to  $\Pi_C$ .

Suppose  $\mathcal{D} = \mathcal{D}'$ . Then, there must exist a  $\lambda$  such that  $\lambda(\mathbf{w} - b\mathbf{1}_N) = (\mathbf{w}' - b'\mathbf{1}_{N'})$ , i.e.  $\lambda\mathbf{w} - \mathbf{w}' = \lambda b\mathbf{1}_N - b'\mathbf{1}_{N'}$ . As before, the left hand side of this equation is a vector in  $\Pi_C$ . The right hand side is a vector in  $\Pi_N \oplus \Pi_{N'}$  and thus orthogonal to  $\Pi_C$ . The two vectors can be equal only if they are 0. Equating the left hand side to 0 gives  $\mathbf{w} = \mathbf{w}'$  and  $\lambda = 1$ . Equating the right hand side to zero, and using  $\lambda = 1$ , gives  $b\mathbf{1}_N = b'\mathbf{1}_{N'}$ . Since  $\mathbf{1}_N$  and  $\mathbf{1}_{N'}$  are linearly independent, this is possible only if  $b = b' = 0$ , which proves the Claim in one direction.

For the opposite direction, assuming  $\mathbf{w} = \mathbf{w}'$  and  $b = b' = 0$  gives  $\mathbf{w}^T \mathbf{x} = 0$  as the equation for both  $\mathcal{D}$  and  $\mathcal{D}'$ , showing that they are identical.  $\square$

## REFERENCES

- [1] A. Brahim, J. Ramírez, J. M. Górriz, L. Khedher, and D. Salas-Gonzalez, "Comparison between Different Intensity Normalization Methods in  $^{123}\text{I}$ -Ioflupane Imaging for the Automatic Detection of Parkinsonism," *PLOS ONE*, vol. 10, no. 6, p. e0130274, Jun. 2015.
- [2] P. Borghammer, J. Aanerud, and A. Gjedde, "Data-driven intensity normalization of PET group comparison studies is superior to global mean normalization," *NeuroImage*, vol. 46, no. 4, pp. 981–988, Jul. 2009.
- [3] R. B. Innis, V. J. Cunningham, J. Delforge, M. Fujita, A. Gjedde, R. N. Gunn *et al.*, "Consensus nomenclature for in vivo imaging of reversibly binding radioligands," *Journal of Cerebral Blood Flow & Metabolism*, vol. 27, no. 9, pp. 1533–1539, 2007.
- [4] J. Dukart, K. Mueller, A. Horstmann, B. Vogt, S. Frisch, H. Barthel *et al.*, "Differential effects of global and cerebellar normalization on detection and differentiation of dementia in FDG-PET studies," *Neuroimage*, vol. 49, no. 2, pp. 1490–1495, 2010.
- [5] S. Ortega Lozano, M. Martinez del Valle Torres, E. Ramos Moreno, S. Sanz Viedma, T. Amrani Raissouni, and J. Jiménez-Hoyuela, "Quantitative evaluation of SPECT with  $^{113}\text{I}$ -CIT. Importance of the reference area," *Revista Española de Medicina Nuclear (English Edition)*, vol. 29, no. 5, pp. 246–250, Jan. 2010.
- [6] S. Shokouhi, J. W. McKay, S. L. Baker, H. Kang, A. B. Brill, H. E. Gwirtsman *et al.*, "Reference tissue normalization in longitudinal  $^{18}\text{F}$ -florbetapir positron emission tomography of late mild cognitive impairment," *Alzheimer's research & therapy*, vol. 8, no. 1, pp. 1–12, 2016.
- [7] S. Nugent, E. Croteau, O. Potvin, C.-A. Castellano, L. Dieumegarde, S. C. Cunnane *et al.*, "Selection of the optimal intensity normalization region for FDG-PET studies of normal aging and Alzheimer's disease," *Scientific Reports*, vol. 10, no. 1, p. 9261, Jun. 2020.
- [8] F. J. López-González, J. Silva-Rodríguez, J. Paredes-Pacheco, A. Niferola-Baizán, N. Efthimiou, C. Martín-Martín *et al.*, "Intensity normalization methods in brain FDG-PET quantification," *NeuroImage*, vol. 222, p. 117229, Nov. 2020.
- [9] C. Messa, M. A. Volonté, F. Fazio, F. Zito, A. Carpinelli, A. d'Amico *et al.*, "Differential distribution of striatal  $^{123}\text{I}$ - $\beta$ -CIT in Parkinson's disease and progressive supranuclear palsy, evaluated with single-photon emission tomography," *European Journal of Nuclear Medicine*, vol. 25, no. 9, pp. 1270–1276, Sep. 1998.
- [10] F. P. M. Oliveira and M. Castelo-Branco, "Computer-aided diagnosis of Parkinson's disease based on  $^{123}\text{I}$ -FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines," *Journal of Neural Engineering*, vol. 12, no. 2, p. 026008, Feb. 2015.
- [11] H. D. Tagare, C. DeLorenzo, S. Chelikani, L. Saperstein, and R. K. Fulbright, "Voxel-based logistic analysis of PPMI control and Parkinson's disease DaTscans," *NeuroImage*, vol. 152, pp. 299–311, May 2017.
- [12] S. Happe, W. Pirker, G. Klösch, C. Sauter, and J. Zeithofer, "Periodic leg movements in patients with Parkinson's disease are associated with reduced striatal dopamine transporter binding," *Journal of Neurology*, vol. 250, no. 1, pp. 83–86, Jan. 2003.
- [13] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, I. A. Illán, and E. W. Lang, "Linear intensity normalization of FP-CIT SPECT brain images using the  $\alpha$ -stable distribution," *NeuroImage*, vol. 65, pp. 449–455, 2013.
- [14] I. A. Illán, J. M. Górriz, J. Ramírez, F. Segovia, J. M. Jiménez-Hoyuela, and S. J. O. Lozano, "Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging," *Medical Physics*, vol. 39, no. 10, pp. 5971–5980, 2012.
- [15] R. Prashanth, S. Dutta Roy, P. K. Mandal, and S. Ghosh, "Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3333–3342, Jun. 2014.
- [16] A. Ortiz, J. Munilla, M. Martínez-Ibañez, J. M. Górriz, J. Ramírez, and D. Salas-Gonzalez, "Parkinson's Disease Detection Using Isosurfaces-Based Features and Convolutional Neural Networks," *Frontiers in Neuroinformatics*, vol. 13, 2019.
- [17] W. Koch, P. E. Radau, C. Hamann, and K. Tatsch, "Clinical Testing of an Optimized Software Solution for an Automated, Observer-Independent Evaluation of Dopamine Transporter SPECT Studies," *Journal of Nuclear Medicine*, vol. 46, no. 7, pp. 1109–1118, Jul. 2005.
- [18] J. C. Taylor and J. W. Fenner, "Comparison of machine learning and semi-quantification algorithms for  $^{113}\text{I}$ -CIT classification: The beginning of the end for semi-quantification?" *EJNMMI Physics*, vol. 4, no. 1, p. 29, Nov. 2017.
- [19] E. Adeli, G. Wu, B. Saghafi, L. An, F. Shi, and D. Shen, "Kernel-based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease," *Scientific Reports*, vol. 7, no. 1, p. 41069, Jan. 2017.
- [20] Z. Wang, X. Zhu, E. Adeli, Y. Zhu, F. Nie, B. Mansell *et al.*, "Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning," *Medical Image Analysis*, vol. 39, pp. 218–230, Jul. 2017.
- [21] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, "Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging," *NeuroImage: Clinical*, vol. 16, pp. 586–594, Jan. 2017.
- [22] G.-H. Huang, C.-H. Lin, Y.-R. Cai, T.-B. Chen, S.-Y. Hsu, N.-H. Lu *et al.*, "Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, no. 5, pp. 508–523, 2020.
- [23] F. P. M. Oliveira, D. B. Faria, D. C. Costa, M. Castelo-Branco, and J. M. R. S. Tavares, "Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on  $^{123}\text{I}$ -FP-CIT SPECT images," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 45, no. 6, pp. 1052–1062, Jun. 2018.
- [24] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "High-Accuracy Classification of Parkinson's Disease Through Shape Analysis and Surface Fitting in  $^{123}\text{I}$ -Ioflupane SPECT Imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 794–802, May 2017.
- [25] S.-F. Huang, Y.-H. Wen, C.-H. Chu, and C.-C. Hsu, "A Shape Approximation for Medical Imaging Data," *Sensors*, vol. 20, no. 20, p. 5879, Jan. 2020.
- [26] Y. Zhou, S. Tinaz, and H. D. Tagare, "Robust Bayesian analysis of early-stage Parkinson's disease progression using DaTscan images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 549–561, 2021.
- [27] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *arXiv:1312.6034 [Cs]*, Apr. 2014.
- [29] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse Reconstruction by Separable Approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.
- [30] C. Gaig and E. Tolosa, "When does Parkinson's disease begin?" *Movement Disorders*, vol. 24, no. S2, pp. S656–S664, 2009.