Label Aided Deep Ranking for the Automatic Diagnosis of Parkinsonian Syndromes

Andrés Ortiz[a,c], Francisco J. Martínez Murcia[b,c], Jorge Munilla[a], Juan M. Górriz[b,c,*], Javier Ramírez[b,c]

[a]Department of Communications Engineering, University of Málaga
[b]Department of Signal Theory, Communications and Networking. University of Granada
[c]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI)

Abstract

Parkinsonism is the second most common neurodegenerative disease in the world. Its diagnosis usually relies on visual analysis of Emission Computed Tomography (SPECT) images acquired using $^{123}I - ioflupane$ radiotracer. This aims to detect a deficit of dopamine transporters at the striatum. The use of Computer Aided tools for diagnosis based on statistical data processing and machine learning methods have significantly improved the diagnosis accuracy. In this paper we propose a classification method based on Deep Ranking which learns an embedding function that projects the source images into a new space in which samples belonging to the same class are closer to each other, while samples from different classes are moved apart. Moreover, the proposed approach introduces a new cost-sensitive loss function to avoid overfitting due to class imbalance (an usual issue in practical biomedical applications), along with label information to produce sparser embedding spaces. The experiments carried out in this work demonstrate the superiority of the proposed method, improving the diagnosis accuracy achieved by previous methodologies and validate our approach as an efficient way to construct linear classifiers.

Keywords: Parkinsonian Syndromes, Computer Aided Diagnosis, Deep Ranking, cost-sensitive learning, label aided classifier, class imbalance.

[*]Corresponding Author. Tel: +34 958243271

Email addresses: aortiz@ic.uma.es (Andrés Ortiz), fjmartinez@ugr.es (Francisco J. Martínez Murcia), munilla@ic.uma.es (Jorge Munilla), gorriz@ugr.es (Juan M. Górriz), javierrp@ugr.es (Javier Ramírez)

1. Introduction

Parkinsonism (PD) is the second most common neurodegenerative disorder after the Alzheimer's disease, with a prevalence between 1% and 3% in the population over 65 years of age. It is characterized by different motor symtpoms such as tremor, hypokinesia and rigidity [1], mainly related to the progressive loss of dopamine Transporters (DaT) of the nigrostriatal pathway.

Currently, diagnosis is commonly supported by nuclear imaging, specifically by Single Photon Emission Tomography (SPECT) using the $^{123}I - ioflupane$ radiotracer (also known by its tradename DaTSCAN). DaTSCAN binds to the dopaminergic transporters at the striatum, allowing to measure quantitatively the amount of DaT in this region. This way, PD can be differentialy diagnosed by detecting a dopaminergic deficit in PD patients with respect to Controls (CN) or other diseases presenting similar symptoms. Different Computer Aided Diagnosis (CAD) tools based on statistical techniques have been developed [2, 3, 4, 5, 6, 7]. These methods search statistical differences between two groups, PD and CN, using statistical learning. However, the use of methods based on artificial neural networks (ANN) have gained popularity in the last years, due to new frameworks for the development of complex neural architectures and new effective training algorithms. This has enabled to use deep neural networks (namely, deep learning) along with a great variety of layers in a wide range of applications [8]. In fact, different deep learning architectures have been proposed to deal with problems of different nature such as speech recognition [9], drug discovery [10] and genomics [11]. Specifically, in the field of image processing and classiffication, Deep Belief Networks (DBN) based on Restricted Boltzman Machines (RBM) [12] have been applied to different image classification problems such as quantitative analysis of gold immunochromatographic strip [13]. On the other hand, combinations of DBNs have been used in [14] for face image modelling, and [15] shows an ensemble architecture for feature extraction and classification of Magnetic Resonance Images (MRI). Other architectures based on Deep Stacked Autoencoders (DAE) [16, 17] have been used in [18] for facial expression recognition. Nevertheless, one of the fields in which the use of deep learning has drastically improved the classification results obtained by statistical learning or other machine learning methods is in computer vision and image classification outperforming the state-of-the art methods [8]. In particular, architectures based on convolutional neural networks (CNN) have reached human-level performance [19] in object recognition applications. CNN is a biologically-inspired model that resembles the human vision system, computing image features at different abstraction levels by means of the convolution operator, which is subsequently applied to the response of the previous layer [20]. Simple architectures such as LeNet-5 [21] or AlexNet [22] provided good results with respect to statistical learning approaches, and many CNN-based architectures have been proposed to improve their classification accuracy. Thus, deeper architectures such as Inception [23], for example, provide a high number of abstraction levels, allowing to compute complex features. Nevertheless, deeper networks are more difficult to train and tend to be overfitted. On the other hand, accuracy degradation could appear when the number of layer increases, which could not be necessary caused by overfitting but to the limitations of the training algorithm [24]. Thus, the solution proposed in [24] to build and train deeper models efficiently consist in copying layers from the learned shallower model to be added at the output of a convolution step, building the so-called *residual blocks*. This method is used in ResNet [24]. Other recent CNN-based networks can capture the orientation of detected objects [25].

All of the previous approaches aim to classify images into different categories, and outperforming baseline architectures such as LeNet-5 or AlexNet requires the use of more complex and deeper architectures. In [26], a model composed of three multiscale networks trained in parallel is proposed for image retrieval. Unfortunately, this architecture did not provide significantly better results than LeNet-5 or AlexNet [27] and the training process implies a considerably higher computational burden.

In this work we propose a deep ranking-based method that not only allows to classify the images but also to compute a similarity measure among them, outperforming, at the same time, previous deep learning based approaches in PD. Moreover, this approach tackles one usual problem in biomedical applications: the prevalence of one class in imbalanced datasets, by introducing a cost-sensitive loss function. Hence, the main contributions of this work can be summarized in the following four points: 1) a deep learning architecture is proposed to rank 3D DaTSCAN images according to their similarity. Thus, unlike other Deep Learning approaches in which the outcome is the predicted class, a similarity measurement is provided in this case. 2) The proposed method can be used to embed a 3D image into a lower-dimensional subspace in which it is possible the computation of similarity metrics based on distance between vectors (for instance, the Euclidean distance). 3) The Hinge loss function proposed in the original Deep Ranking paper [26] has been modified to include labels information in order to reinforce the learning. 4) A cost-sensitive loss function is used to deal with imbalanced datasets, as it is the case of the PPMI database we are using in this work.

The rest of the paper is organized as follows. Section 2 reviews the related work on neuroimage classificataion. Section 3 describes the database used in this work and the architecture of the proposed model, along with the techniques used to deal with imbalanced datasets. Section 5 presents the classification results using data from the Parkinson Progression Neuroimaging Initiative (PPMI) database and discuses these results. Regions of interests are also computed by means of the deep neural model and shown in this section. Finally, Section 6 shows the conclusions drawn from this work along with its practical applicability.

2. Related Work

Current neuroimaging systems provide high spatial and color resolution, and they have become the least invasive method for the diagnosis of brain disorders. However, the diagnosis of neurodegenerative diseases such as PD by means of visual assessment is a time consuming task and subject to the experience of the expert neurologist. On the other hand, the vast amount of information contained in a 3D image requires the use of computer aided tools to be exploited, allowing to find complex, disease-related patterns and thus increasing the diagnosis accuracy. This way, different computer-based methods have been developed for the diagnosis of PD. Previous approaches use statistical learning techniques along with signal processing methods to reveal disease-related patterns and to classify the images. Thus [2] proposes the use of 2D empical mode decomposition to split DaTSCAN images into different intrinsic mode functions, accounting for different frequency subbands. The components are used to select features related to PD that clearly differentiate them from CN and allowing a easy visual inspection. Other proposals dealing with decomposition into components are [6] and [28]. [4] proposes the extraction of 3D textural-based features for the characterization of the dopamine transporters concentration in the image and [29] decomposes the DaTSCAN images into statistically independent components which reveal patterns associated to PD. Moreover, in this approach, image voxels are ranked by means of their statistical significance in class discrimination. A recent approach also based on multivariate decomposition techniques is proposed in [28], where the use of functional principal component analysis on 3D images is proposed. This is addressed by sampling the 3D images using fractal curves in order to transform the 3D DatSCAN images into 1D signals, preserving the neighbourhood relationship among voxels. In addition, [5] uses partial least squares (PLS) to extract features that are eventually used for classification using linear support vector classifiers. In [7], the authors use univariate (voxel-wise) statistical parametric mapping and multivariate pattern recognition using linear discriminant classifiers to differentiate among different Parkinsonian syndromes.

The previous works use statistical learning methods. As explained in the introduction, methods based on neural networks, especially deep learning-based methods, have paved the way to discover complex patterns and, consequently, to outperform the diagnosis accuracy obtained by classical statistical methodologies [15, 30]. The use of models containing stacks of layers composed of a large number of units that individually perform simple operations allows to compute models containing a large number of parameters. Moreover, these massively parallellize architectures are able to discover very complex patterns in the data by a learning process formulated as an optimization problem. Thus, [15] proposes the combination of Deep Belief Networks (DBN), each learning over data extracted from a different brain region, aiming to search for patterns related to the Alzheimer's Disease onset. Moreover, [30, 27] propose the use of Convolutional Neural Networks (CNN) to discover patterns associated to PD. Increasing the accuracy requires the use of deeper networks. However, the increment in the number of parameters in simple CNN [30, 27] makes the network prone to overfitting and the limitations of the training algorithms arise. Thus, architectures combining more elaborated blocks such as [24] are required to effectively increase the number of layers. In this work, we propose the combined use of simple CNNs composed of 4 layers trained in parallel, to outperform the accuracy provided by both, previous statistical learning approaches and previous deep learning architectures. Moreover, the model used in this work has additional advantages; unlike previous CNN-based approaches, it not only allows to classify the image but also to embed it into a lower dimensional subspace where similarity measures can be computed by means of distance metrics. Moreover, the deep ranking method has been improved including a label-aided loss function by modifying the original Hinge loss proposed in [26] to produce better defined clusters. Additionally, the proposed loss function is cost-sensitive to address the imbalanced database problem. As a result, a deep ranking model with simple CNNs composed of 4 layers outperforms the previous deep learning approaches as well as those based on statistical learning.

3. Materials and Methods

3.1. Database

Data used in the preparation of this article was obtained from the Parkinson's Progression Markers Initiative (PPMI) (`www.ppmi-info.org/data`). For up-to-date information on the study, visit `www.ppmi-info.org`. The images in this database were imaged 4 + 0.5 hours after the injection of between 111 and 185 MBq of DaTSCAN. Raw projection data are acquired into a 128×128 matrix stepping each 3 degrees for a total of 120 projection into two 20$\%$ symmetric photopeak windows centered on 159 KeV and 122 KeV with a total scan duration of approximately 30 - 45 minutes [31].

A total of $N = 642$ DaTSCAN images from this database were used in the preparation of the article. Specifically, the baseline acquisition from 448 subjects suffering from PD and 194 normal controls was used. More details on the demographics of this dataset are given in Table 1.

Table 1: Demographics of the PPMI dataset

Group	Sex	N	Age [STD]
Control	F	65	58.85 [11.95]
	M	129	62.00 [10.74]
PD	F	160	61.49 [9.96]
	M	288	62.89 [9.71]

3.2. Spatial Normalization

Spatial normalization is frequently used in neuroimaging studies. It eliminates differences in shape and size of brain, as well as local inhomogeneities due to individual anatomic particularities. It is particularly key in group analysis, where voxel-wise differences are analysed and quantified[32]. In this procedure, individual images are mapped from their individual subject space (image space) to a common reference space, usually stated using a template. The mapping involves the minimization of a cost function that quantifies the differences between the individual image space and the template. The most frequent template is the Montreal Neurological Institute (MNI), set by the International Consortium for Brain Mapping (ICBM) as its standard template, currently in its version ICBM152[33], an average of 152 normal MRI scans in a common space using a nine-parameter linear transformation. A particular case of affine transformation is the similarity transformation, where only scale, translation and rotation are applied. This is often used for motion correction and reorientation of brain images with respect to a reference, and is frequently performed automatically on many imaging equipment. The DaTSCAN images from the PPMI dataset are roughly re-aligned. We will refer to this as non-normalized (given that it is only a similarity transformation that preserves shape) or ‘original’. We further preprocessed the images using the SPM12[34] New Normalize procedure with default parameters, which applied affine and local deformations to achieve the best warping of the images and a custom DaTSCAN template defined in [35]. Finally, the images were linearly down-resampled to a final size of (57,69,57), the input size of the network.

3.3. Intensity Normalization

Intensity normalization is a technique that changes the global or local intensity values of an image in order to ensure that the same intensity levels correspond to similar physical measures. In nuclear imaging, the use of intensity normalization is key in order to compare brain activity or function between subjects. A similar intensity should indicate a similar drug uptake and therefore, differences in these values may be due to different pathologies[36, 37, 38].

Global intensity normalization in neuroimaging usually follows the expression:

$$\hat{\mathbf{I}}_i = \mathbf{I}_i / I_{n,i} \tag{1}$$

where $\mathbf{I}_i$ is the image of the i^{th} subject in the dataset, $\hat{\mathbf{I}}_i$ is the normalized image, and I_n is an intensity normalization value that is computed independently for each subject. In this work, we used the Integral Normalization[39], which sets I_n to the average of all values in a certain volume of the image, in an approximation of the integral. In Parkinson, this is often set to the average of the brain without the specific areas: the striatum; although the influence of these areas is often small, and it can be approximated by the mean of the whole image.

4. Methods

4.1. Deep Ranking

Deep Ranking [26] is a method originally proposed to find a similar set of images in a given, usually large image database and for reverse image search applications. The method aims to learn a similarity function that allows to compute a distance measurement among images. Unlike classical approaches based on the computation of invariant features [40, 41], DR uses a combination of deep learning architectures to compute features and also to learn an embedding function $f(\cdot)$ that projects the input images into a new space in which images from different classes are clearly separable. Moreover, the value returned by that function can be used as a similarity measure between images, as it is proportional to the distance between images in the embedding space.

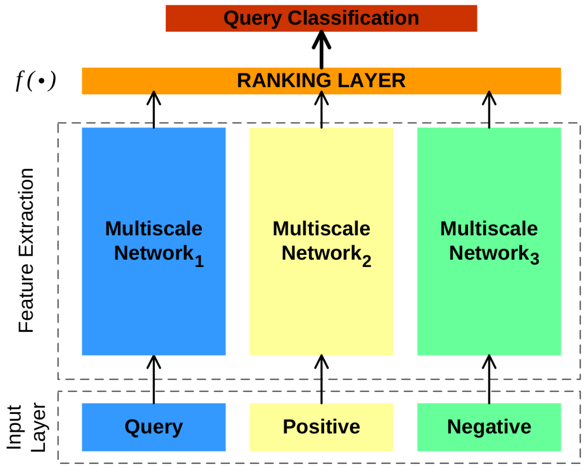

Figure 1: Structure of the Deep Ranking Model proposed in [26]. CNN blocks contains multiscale convolutional networks

DR model consists in training three identical networks in a weight-sharing fashion. This implies that parameters learned during backpropagation are the same in the three neural networks. However, these three networks are fed using images belonging to different classes: the first, named *query* (q) network, uses an image from the class to be retrieved from the database. The other two networks are fed with images from the same class (q^+) and images from a different class (q^-), respectively. Thus, after forward propagation of the inputs, each network will provide a different output, but the weights in the backpropagation are shared during the updating process. Finally, the uppermost level implement a ranking layer that computes the loss function

$$loss = max\{0, \lambda + D(f(q), f(q^+)) - D(f(q), f(q^-))\} \tag{2}$$

which is a form of *Hinge* loss, where $f(x)$ is the embedding function, λ is a regularization parameter and D is the Euclidean defined as:

$$D(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2 \tag{3}$$

.

As a result, the embedding function $f(\cdot)$ is learnt by minimizing the loss function in equation 2. The original DR model is shown in Figure 1, where CNN blocks contain a specific multiscale network developed to be invariant under scale transformations in the input layer.

4.2. Learning non-linear classifiers from DeepRanking

Deep Ranking can be seen as a method to learn a classifier in a space in which samples may not be linearly separable. For the sake of clarity, let (x_i, y_i) be a pair composed of samples ($x_i \in \mathbb{R}^d$) and their corresponding labels ($y_i \in \{-1, +1\}$), where $i = \{1, ...n\}$.

Then, a linear classifier can defined as

$$g(x_i) = \mathbf{W}^\top x_i + b \tag{4}$$

where the classification outcome is based on $sign(g(x_i))$.

The training of the classifier consists in calculating $\mathbf{W}$ and b, which defines the best separating hyperplane as $x_i\mathbf{W}^\top + b = 0$. The computation of the hyperplane can be formulated in different ways. In Support Vector Machines, the hyperplane is computed by maximizing the margin $\frac{2}{\|\mathbf{W}\|}$ to the hyperplane as

$$\max_{\mathbf{W}} \frac{2}{\|\mathbf{W}\|} \quad subject\ to\ \ \mathbf{W}^\top x_i + b = \begin{cases} \geq +1, & \text{if } y_i = +1 \\ \leq -1, & \text{if } y_i = -1 \end{cases} \tag{5}$$

The optimization of this objective function can also be formulated as the cuadratic programing problem

$$\min_{\mathbf{W},\xi,\mathbf{b}} \frac{1}{2}\|W\|_2^2 + \lambda \sum_{i=1}^{N} \xi_i \quad subject\ to\ \ y_i(W^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, ..., N \tag{6}$$

The previous equations define a linear classifier. It means that only linear separable classes can be correctly classified using this method. Therefore, when input data are not linearly separable, it is necessary to compute a non-linear decision boundary in the input space. Basically, it consists in finding an embedding function $F : \mathbb{R} \rightarrow \mathbb{R}^{\mathbb{K}}$ to map the original data into a new space in which $f(x)$ is linearly separable.

Using the distance measurement D, it is possible to define an scoring function F

$$F(q, p^{+}, p^{-}) = D(f(q), f(p^{+})) - D(f(q), f(p^{-})), \quad p^{+} \in \mathcal{P}, p^{-} \in \mathcal{N} \tag{7}$$

where $\mathcal{P}$ and $\mathcal{N}$ are the two classes to be classified. Thus, the *scoring* function F:

$$F(q) = \begin{cases} \geq 0, & \text{if } q \in \mathcal{N} \ (y_i = +1) \\ < 0, & \text{if } q \in \mathcal{P} \ (y_i = -1) \end{cases} \tag{8}$$

since $f(\cdot)$ embeds the samples x_i into a new space in which the distance between similar samples is proportional to their similarity in the original space. Deep Ranking allows using a Deep Learning architecture to learn the embedding function via a loss function that increases whenever dissimilar points present larger distances in the embedding space. This way, the minimization of the loss function ensures that the learnt embedding function F maps similar samples close together in the embedding space.

We propose a different loss function, regarding the original Deep Ranking method [26], which includes label definition (and not only the distance measurement) in the following way:

$$loss\left(F(q, p^{+}, p^{-}), y_i\right) = \max\{0, \lambda - y_i F(q, p^{+}, p^{-})\}, \quad \lambda > 0 \tag{9}$$

where λ is again the regularization parameter.

Thus, the proposed loss function (in equation 9) returns 0 when $F(q, p^{+}, p^{-})$ and y_i have the same sign, while increases linearly with $F(q, p^{+}, p^{-})$ when they have opposite sign. The result of the minimization of this loss function is the learning of the linear classifier $F(\cdot)$.

4.3. Quartet sampling

As explained in Section 4.1, the input to the model consists of triplets containing three samples: the query sample, a sample of the same class that the query sample (also called p^{+}) and a sample of different class that the query sample (also called p^{-}). This way, each input sample can be represented as $t = \{q, p^{+}, p^{-}\}$. Each component of a triplet is processed by the corresponding siamese network as shown in Figure 1 and the distance between the embeddings provided by the last layer is computed as indicated in Equation 7. In our implementation, the original triplets are extended to quartets, sets of four values including the triplets previously described and the corresponding y_i containing label information. Quartet generation from the original samples is addressed in the following way:

Algorithm: Quartet sampling algorithm

```
for ( $q$ in $\mathcal{P}$ ) {
    extract $k$ random samples $p^+ \in \mathcal{P}$, $p^+ != q \rightarrow P^+$;
    for ( $p^+$ in $P^+$ ) {
        extract $k$ random samples $p^- \in \mathcal{N} \rightarrow P^-$;
$quartet_1$={$Q$, $P^+$, $P^-$, -1}
for ( $q$ in $\mathcal{N}$ ) {
    extract $k$ random samples $n^+ \in \mathcal{N}$, $n^+ != q \rightarrow N^+$;
    for ( $n^+$ in $N^+$ ) {
        extract $k$ random samples $n^- \in \mathcal{P} \rightarrow N^-$;
$quartet_2$={$Q$, $N^-$, $N^+$, +1}
Quartet=concatenate{$quartet_1$, $quartet_2$}
where $Q$ is the set of query samples.
```

4.4. Dealing with class imbalance by cost-sensitive loss

Demographic data (Table 1) shows that the PPMI database is highly imbalanced. Since the number of PD subjects is more than twice the number of Controls, this may cause an important bias in the learned model, known as the *accuracy paradox* [42, 43]. There are different ways to mitigate the undesirable effect of class imbalance. The first one consists in oversampling the under-represented class (actually, sampling with replacement). Although this is a common practice, it presents an obvious drawback: the database does not have new samples (which is what we really need), but pre-exisiting samples are simply replicated. The second way consists in undersampling the over-represented class (actually, deleting samples). This approach also has a clear flaw, since fewer samples, and consequently, less information, is available to learn the model. Thus, in practice, over and under-sampling methods can be used in actual problems only when a very large amount of sample data are available. As a result, they cannot be applied in most practical biomedical problems, where that large data sets are not commonly available. A third method compensates the class imbalance by applying cost-sensitive learning [44]. This approach assigns different weights to miss-classification of samples from different classes, making the *loss* function *sensitive* to the class.

$$loss\left(F(q,p^{+},p^{-}),y_i\right) = \max\{0, 1 - y_i F(q,p^{+},p^{-}) \cdot \omega_i\} \tag{10}$$

where ω_i is the weight assigned to class i, computed as

$$\omega_i = \begin{cases} \frac{\#C_i}{\#C_j} & \text{if} \quad \#C_i > \#C_j \\ 1 & otherwise \end{cases} \tag{11}$$

and $\#C_i$ is the number of samples of class i.

4.5. Label Aided Deeep Ranking

In this work, we modify the basic structure of the Deep Ranking model proposed in [26]. More specifically, we define three identical networks containing a Multilayer AlexNet based architecture [22] instead of the multiscale convolutional network used in the original model. The spatial normalization procedure performed on the images (see Section 3.2) avoids dealing with different scales and the cost sensitive hinge loss, described in 10, works better than the original with imbalanced datasets.

Each multimodal network in Figure 2 contains a CNN block, an Alexnet-type convolutional network composed of five 3D convolutional layers with *RelU* activations. Maxpooling is used to reduce the data size throughout the network. Moreover, batch normalization layers are included to ease convergence and dropout in hidden layers regularizes the corresponding outputs, making the network less sensitive to specific weights, which aims to prevent overfitting and thus to improve the generalization performance. Eventually, the network contains three fully-connected layers, ending-up with a 512-neuron dense layer, corresponding to the number of features extracted from

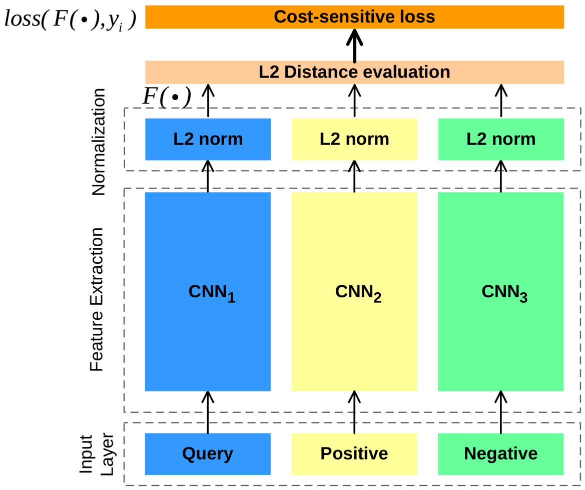

Figure 2: Deep Ranking architecture used in this work. The ranking layer uses a cost-sensitive loss function which is also aware of the labels during the training stage.

the images. The last layer (ℓ_2 normalization) normalizes the outputs to limit the bounds of the output space.

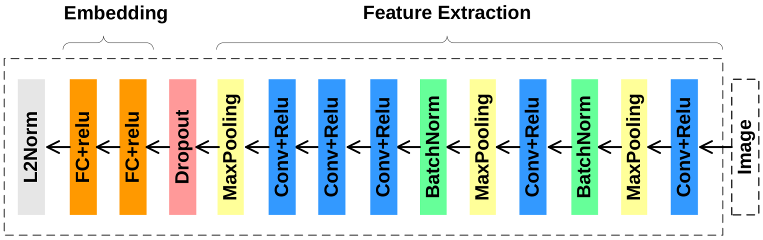

Figure 3: Layers contained inside each block of the Deep Ranking architecture used in this work.

Figure 4 shows the embedding learnt by our deep ranking model when trained using DaTSCAN images from CN and PD subjects, where a non-linear boundary is defined. It shows a 3D projection of the last dense layer of the DR model by means of the t-distributed Stochastic Neighbor Embedding (T-SNE) algorithm [45]. T-SNE allows visualizing high-dimensional data by converting similarities between data points in the high-dimensional space in distances in a lower-dimensional space. As a result of this mapping, CN and PD samples are differently distributed (radially) and a new sample

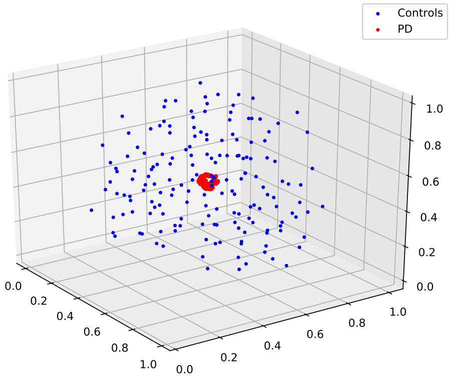

Figure 4: Embeddings of CN and PD subjects performed by deep ranking method. Distances are normalized to [0,1]

can be classified by comparing the mean distance to CN and PD training samples. As shown in this figure, different classes are grouped in different clusters clearly separated in the DR embedding space. Hence, it is possible to distinguish between classes by measuring the distance from a new sample to both clusters. In fact, classification is addressed by measuring the mean distance from a test sample to all the training CN and PD samples in the embedding space.

5. Results and discussion

In this section, classification results using the deep ranking based proposal are shown. Classification experiments between Controls and PD of the PPMI database have been performed. All the architectures in this work have been implemented using Keras[46] and the Python API, and run in a cluster containing 8 Intel(R) Xeon(R) CPU E5-2640v4 at 2.40GHz and 3 Nvidia GTX1080Ti GPUs.

Firstly in this part, we explore the output of the network once trained. As explained in Section 4.1, deep ranking consists of three siamese networks, individually evaluated but jointly updated. This way, once the network is trained, we can use one network to evaluate the output obtained when the network is fed with a specific sample (since the other networks will provide the same output).

A 3D representation of the data manifold shown in Figure 4 demonstrates that the mapping generated by the network is able to separate the two classes efficiently. In fact, two clusters radially distributed can be observed when the dimension of the output space is reduced to 3. These clusters corresponds to CN and PD subjects. In order to continue exploring the properties of the embedding space, we also computed the mean activation of the network when CN and PD test subjects are used as input. This is graphed in Figure 5, clearly showing different activation values for different classes.

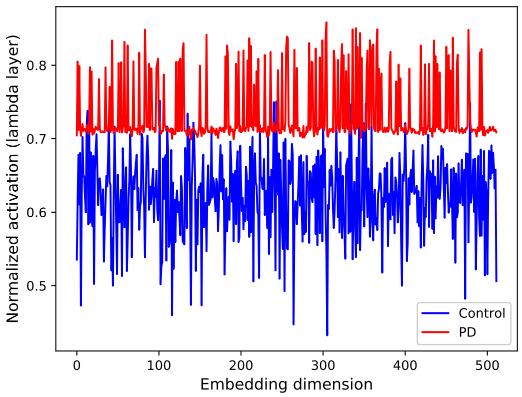

Figure 5: Mean normalized activation of the last dense layer when the network is fed with CN and PD subjects.

Moreover, in order to assess the differences between PD and CN subjects, we present a boxplot (Figure 6) of the activation of each subject using the trained network.

As shown in Figures 4, 5 and 6, the network clearly produces different embedding values for different classes and therefore, these values can be used to classify samples. On the other hand, Figure 7 shows the classification accuracy provided by the model for different number of neurons at the top layer, corresponding to the embedding dimension. Although good classification results are obtained from 128 neurons, the best results are provided when the embedding layer generates 512-dimensional vectors.

5.1. Evaluation

Classification performance has been assessed using the accuracy, sensitivity and specificity metrics considering a binary classification problem (i.e. CN vs. PD). These

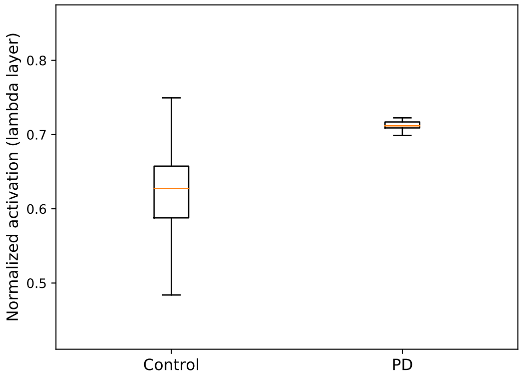

Figure 6: Boxplot computed for the activations of CN and PD subjects

values are derived from the confusion matrix, and then, from the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) of the predictions of the trained model. The definitions of the mentioned metrics can be found in equations 12, 13 and 14.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{13}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{14}$$

The performance has been computed within a cross-validation scheme; in particular, stratified K-fold (k=10) cross-validation is used. Is it worth noting that using stratified cross-validation ensures keeping the class distribution among folds [47]. Regarding the training process, the network was trained during 35 epochs with a batch training algorithm, using a batch size of 8 samples. Figure 8 shows the convergence process across 70 iterations for both training and validation sets.

Unlike methods based on the extraction of specific features from the image such as [48] and [4], or methods that use statistical techniques to select the most discriminative voxels [29], CNN-based methods compute specific features for classification as a re-

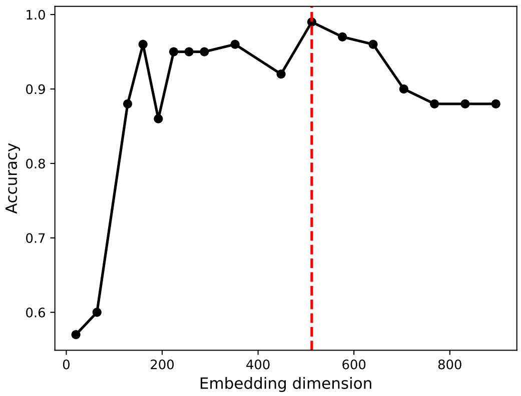

Figure 7: Classification accuracy obtained for different number of neurons at the embedding layer.

Table 2: Classification results using different methods

Method	Accuracy	Sensitivity	Specificity	AUC
LeNet [21, 27]	0.96 ± 0.04	0.84 ± 0.30	0.98 ± 0.02	0.94
AlexNet [22, 27]	0.97 ± 0.01	0.92 ± 0.09	0.97 ± 0.01	0.97
SVC	0.97 ± 0.01	0.92 ± 0.09	0.97 ± 0.01	0.98
Significance M. [29]	0.92	0.95	0.89	0.90
Brahim et. al[48]	0.92	0.94	0.91	-
Textural Patterns[4]	0.95	0.95	0.94	-
EMD [49]	0.95	0.95	0.94	0.94
Label Aided Deep Ranking	**0.99 ± 0.01**	**0.97 ± 0.03**	**0.99 ± 0.01**	**0.99**

sult of the learning process. Moreover, CNN generates features at different abstraction levels, related to discriminative information retrieved from the images. In addition, our proposal process the 3D images, instead of using representative layers as in [49] which exploit all the information contained on the image. Classification using AlexNet network [21, 27] provides an AUC up to 0.97. Our proposal using three siamese AlexNet network outperform the best outcome previously obtained, providing an AUC of 0.99. On the other hand, since the three networks share the weights, the computational bur-

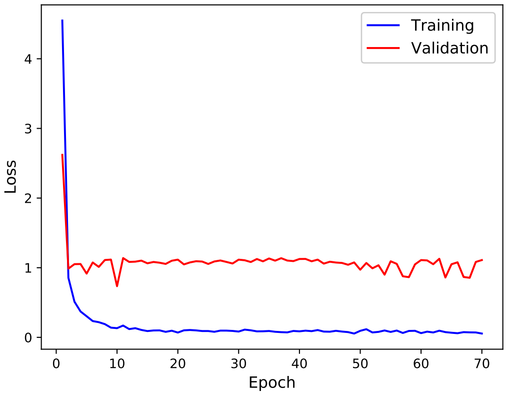

Figure 8: Evolution of the loss function for training and validatgion sets.

den associated to the training process is only slightly increased.

5.1.1. ROC Analysis

ROC analysis depicts the true positive rate (sensitivity) vs. false negative rate (1-specificity) for different thresholds used in the classifier's output score to decide the predicted class. In our case, the neural network does not produce the predicted class, but only the mapping of an input into the embedding space. The prediction is eventually computed by distance measurements from the test sample to the center of the clusters computed during the training. Thus, let $d_C N$ and $d_P D$ be the mean distances from the test sample to all the CN and PD training samples, respectively. We define the score of the classifier as $D = D_C N - D_P D$. Hence, the ROC curve can be computed by ranking the score values for each class, and depicting each point depending on whether it is a TP or FP[50]. ROC curves provides a good visual comparison of the performance among different methods, and it has been used in many works [51, 52, 53, 37, 54]. ROC curves provide another interesting statistic, namely Area Under ROC curve (AUC). This value indicates the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one. In other words, AUC=1.0 indicate a perfect classifier while AUC=0.5 indicates a random classifier. Figure 9 shows the ROC curves for the proposed deep ranking method along with the curves for an AlexNET convolutional network, a LeNET convolutional network and a linear Support Vector Classifier (SVC). It is noteworthy that the AlexNET architecture used in this comparison is identical to the AlexNET network included in the Deep Ranking architecture. Additionally, AUC metrics are also shown at Table 2.

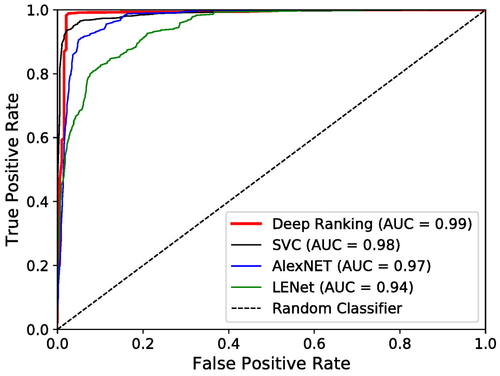

Figure 9: ROC curves obtained using different Deep Learning classification methods. Results using a Support Vector Classifier are also shown for comparison.

Figure 9 and Table 2 show that Deep Ranking outperforms the other compared methods, providing a classification performance close to AUC=1.0. Thus, although the training of the deep ranking architecture requires more computation (in this case, the inputs are propagated through three networks), it achieves a higher accuracy, deserving the higher computational burden.

5.2. Exploring the network insights

Deep learning architectures are frequently critizised, being considered black boxes that merely classify samples with high accuracy, while sometimes it is not clear which features are used to differentiate between classes. Analyzing the network activation on a specific input is a way to take a step beyond the classification results. One usual way to reveal the features computed during the training stage consist in exploring the activation that an input belonging to a specific class produces in different parts of the network. The most frequent procedure in CNNs is to analyze the raw features at the input layer that activate specific neurons at the output layer, which helps to understand the information the network is using to classify samples (i.e. to detect differences between CN and PD subjects). This is addressed here by computing the *saliency* maps [55]. Saliency maps are a representation of changes in the network output with respect to small changes in the input, being able to highlight those regions of the image that play a more important role in the output. In classical CNN for classification, saliency is obtained by computing the gradient of an output category with respect to the input. In our case, the output is a dense layer that computes the representation of an input in the embedding space, rather than being a categorical output containing few neurons (i.e. containing as many neurons as output categories). Thus, saliency is obtained by computing the gradient of the representation in the embedding space with respect to the input.

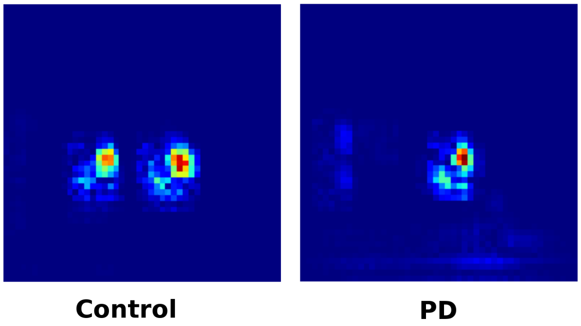

Figure 10: Saliency maps computed for control and PD subjects. The coronal slice with maximum activation values is shown. The gradient is calculated from the most activated filter at the embedding layer

Figure 10 shows the mean saliency map computed for a Controls and PD subjects, indicating the raw features (i.e. the voxels on which the network is actually focusing to compute the embedding that discriminates effectively between classes). As Figure 10 shows, different regions are used to compute the embedding in Controls and in PD subjects, according to different activity in the dopaminergic area revealed by DaTSCAN imaging. In Figures 11 and 12 we overlap the saliency maps of controls and PD respectively to its respective DaTSCAN images, in order to highlight the areas that have a larger influence to compute the embeddings.

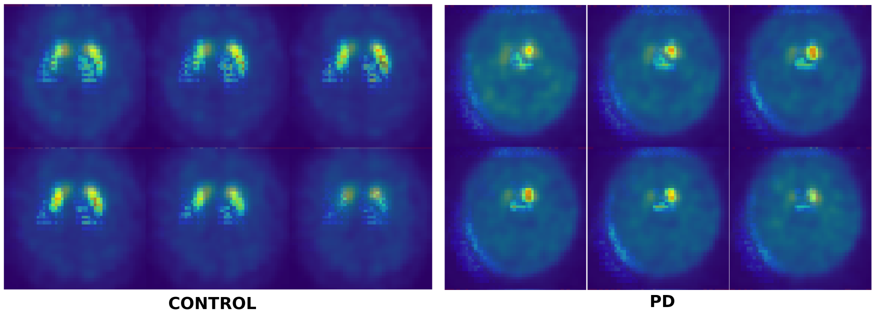

Figure 11: Saliency maps computed for control and PD subjects, overlaying the DaTSCAN imaging. Some relevant slices in the axial plane are shown.

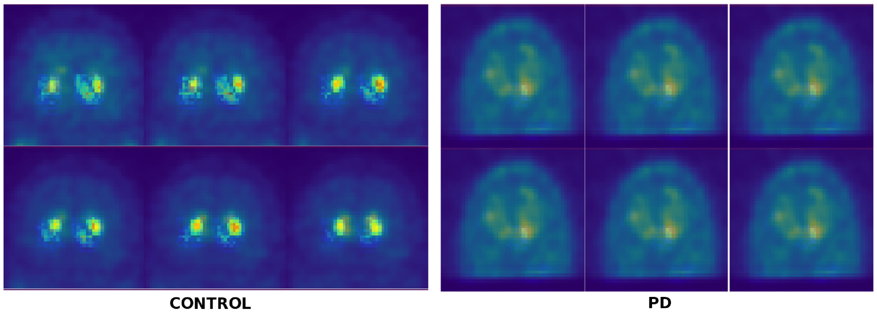

Figure 12: Saliency maps computed for control and PD subjects, overlaying the DaTSCAN imaging. Some relevant slices in the coronal plane are shown.

6. Conclusions and future work

In this paper a deep ranking based method is used to compute the embedding of 3D DaTSCAN images. The method used in this work includes three siamese 3D convolutional networks and a normalization lambda layer at the output, aiming to classify DaTSCAN images. Deep ranking aims to produce an embedding space in which controls are far enough from PD to be able to clearly discriminate betweem them. In our model, we introduced a new loss function that includes label information in the optimization process that defines the embedding space. Experiments performed using data from the PPMI database shows that the proposed method outperforms previous statistical proposals and also outperforms recently proposed deep learning convolutional models. In order to take a further step in the comprehension of the results provided by the proposed model, we analyze the activations of the neural model for a specific input. This has been addressed by computing the saliency maps, which show the group of voxels in the original image that most contribute to the output. Thus, saliency maps depict the regions related to the dopamine transporters that reveal differences between Controls and PD. Due to the good results obtained with the former proposal in DaTSCAN image classification, we plan to apply the same model to the diagnosis of other neurodegenerative diseases such as the Alzheimer's disease using Positron Emision Tomograpgy (PET) or Magnetic Resonance Images (MRI).

Acknowledgements

This work was partly supported by the MINECO/FEDER under TEC2015-64718-R and PSI2015-65848-R projects. We gratefully acknowledge the support of NVIDIA Corporation with the donation of one of the GPUs used for this research. PPMI - a public - private partnership - is funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbott, Biogen Idec, F. Hoffman-La Roche Ltd., GE Healthcare, Genentech and Pfizer Inc.

References

[1] T. Eckert, C. Tang, D. Eidelberg, Assessment of the progression of parkinson's disease: a metabolic network approach, The Lancet Neurology 6 (10) (2007) 926 – 932. `doi:https://doi.org/10.1016/S1474-4422(07)70245-4`.
URL `http://www.sciencedirect.com/science/article/pii/S1474442207702454`

[2] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. G. Río, M. Moreno-Caballero, Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease, Expert Systems with Applications 40 (7) (2013) 2756 – 2766.

[3] C. R. Pereira, D. R. Pereira, F. A. d. Silva, C. Hook, S. A. T. Weber, L. A. M. Pereira, J. P. Papa, A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements, in: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, 2015, pp. 171–176. `doi:10.1109/CBMS.2015.34`.

[4] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, M. Moreno-Caballero, M. Gómez-Río, Parametrization of textural patterns in 123i-ioflupane imaging for the automatic detection of parkinsonism, Medical Physics 41 (1) 1–13.

[5] L. Khedher, J. Ramírez, J. Górriz, A. Brahim, F. Segovia, Early diagnosis of alzheimers disease based on partial least squares, principal component analysis and support vector machine using segmented mri images, Neurocomputing 151 (2015) 139 – 150. `doi:https://doi.org/10.1016/j.neucom.2014.09.072`.
URL `http://www.sciencedirect.com/science/article/pii/S0925231214013137`

[6] F. Martinez-Murcia, J. Górriz, J. Ramírez, A. Ortiz, the Alzheimer's Disease Neuroimaging Initiative, et al., A Spherical Brain Mapping of MR Images for the Detection of Alzheimer's Disease, Current Alzheimer Research 13 (5) (2016) 575–588.

[7] S. Badoud, D. V. D. Ville, N. Nicastro, V. Garibotto, P. R. Burkhard, S. Haller, Discriminating among degenerative parkinsonisms using advanced 123i-ioflupane spect analyses, NeuroImage: Clinical 12 (2016) 234 – 240. `doi:https://doi.org/10.1016/j.nicl.2016.07.004`.
URL `http://www.sciencedirect.com/science/article/pii/S2213158216301231`

[8] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. `doi:10.1038/nature14539`.

[9] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, Deep neural net-

works for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29 (2012) 82–97.

[10] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug Discovery Today 23 (6) (2018) 1241 – 1250. doi:https://doi.org/10.1016/j.drudis.2018.01.039.
URL http://www.sciencedirect.com/science/article/pii/S1359644617303598

[11] B. Alipanahi, A. Delong, M. Weirauch, B. J Frey, Predicting the sequence specificities of dna- and rna-binding proteins by deep learning 33.

[12] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554. doi:10.1162/neco.2006.18.7.1527.
URL http://dx.doi.org/10.1162/neco.2006.18.7.1527

[13] N. Zeng, Z. Wang, H. Zhang, W. Liu, F. E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognitive Computation 8 (2016) 684–692.

[14] Y. Tang, A.-R. Mohamed, Multiresolution deep belief networks, in: N. D. Lawrence, M. Girolami (Eds.), Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, Vol. 22 of Proceedings of Machine Learning Research, La Palma, Canary Islands, 2012, pp. 1203–1211.

[15] A. Ortiz, F. J. Martínez-Murcia, M. J. García-Tarifa, F. Lozano, J. M. Górriz, J. Ramírez, Automated diagnosis of parkinsonian syndromes by deep sparse filtering-based features, in: Innovation in Medicine and Healthcare 2016, Springer, 2016, pp. 249–258.

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
URL http://dl.acm.org/citation.cfm?id=1756006.1953039

[17] G. E Hinton, Learning multiple layers of representation 11 (2007) 428–34.

[18] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A. M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643 – 649. doi:https://doi.org/10.1016/j.neucom.2017.08.043.
URL http://www.sciencedirect.com/science/article/pii/S0925231217314649

[19] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, T. Masquelier, Deep networks can resemble human feed-forward vision in invariant object recognition 6 (2016) 32672.

[20] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Computation 29 (9) (2017) 2352–2449.

[21] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[22] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., USA, 2012, pp. 1097–1105.
URL http://dl.acm.org/citation.cfm?id=2999134.2999257

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.

[25] S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, in: NIPS, 2017.

[26] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 1386–1393. doi:10.1109/CVPR.2014.180.
URL http://dx.doi.org/10.1109/CVPR.2014.180

[27] F. J. Martinez-Murcia, J. M. Gorriz, J. Ramirez, A. Ortiz, Convolutional neural networks for neuroimaging in parkinson's disease: Is preprocessing needed?Exported from https://app.dimensions.ai on 2018/09/20. doi:10.1142/s0129065718500351.
URL https://app.dimensions.ai/details/publication/pub.1105862616

[28] A. Ortiz, J. Munilla, F. J. Martinez-Murcia, J. M. Gorriz, J. Ramirez, Empirical functional pca for 3d image feature extraction through fractal sampling, International Journal of Neural Systems 0 (ja) (0) 1–22. doi:10.1142/S0129065718500405.
URL https://doi.org/10.1142/S0129065718500405

[29] F. Martínez-Murcia, J. Górriz, J. Ramírez, I. Illán, A. Ortiz, Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging, Neurocomputing 126 (2014) 58 – 70, recent trends in Intelligent Data Analysis Online Data Processing. doi:https://doi.org/10.1016/j.neucom.2013.01.054.
URL http://www.sciencedirect.com/science/article/pii/S0925231213007005

[30] F. J. Martinez-Murcia, A. Ortiz, J. M. Górriz, J. Ramírez, F. Segovia, D. Salas-Gonzalez, D. Castillo-Barnes, I. A. Illán, A 3d convolutional neural network approach for the diagnosis of parkinson's disease, in: J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, H. Adeli (Eds.), Natural

and Artificial Computation for Biomedicine and Neuroscience, Springer International Publishing, Cham, 2017, pp. 324–333.

[31] P. The Parkinson Progression Markers Initiative, Imaging technical operations manual, 2 edn (June 2010).

[32] F. Martinez-Murcia, J. Górriz, J. Ramírez, Computer Aided Diagnosis in Neuroimaging, 1st Edition, InTech, 2016, Ch. 7, pp. 137–160. doi:10.5772/64980.

[33] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, et al., A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm), Philosophical Transactions of the Royal Society of London B: Biological Sciences 356 (1412) (2001) 1293–1322.

[34] U. London Institute of Neurology, Statistical parametrix mapping (2012).
URL http://fil.ion.ucl.ac.uk/spm/

[35] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, I. A. Illán, P. Padilla, F. J. Martínez-Murcia, E. W. Lang, Building a FP-CIT SPECT brain template using a posterization approach, Neuroinformatics 13 (4) (2015) 391–402.

[36] P. Padilla, J. Górriz, J. Ramírez, D. Salas-González, I. Illán, Intensity normalization in the analysis of functional datscan spect images: The -stable distribution-based normalization method vs other approaches, Neurocomputing 150 (2015) 4 – 15, bioinspired and knowledge based techniques and applications The Vitality of Pattern Recognition and Image Analysis Data Stream Classification and Big Data Analytics. doi:https://doi.org/10.1016/j.neucom.2014.01.080.
URL http://www.sciencedirect.com/science/article/pii/S0925231214013952

[37] F. Martínez-Murcia, J. Górriz, J. Ramírez, C. Puntonet, D. Salas-González, Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann-Whitney-

Wilcoxon U-Test, Expert Systems with Applications 39 (10) (2012) 9676–9685. doi:10.1016/j.eswa.2012.02.153.

[38] F. Segovia, J. M. Górriz, J. Ramírez, R. Chaves, I. Á. Illán, Automatic differentiation between controls and Parkinson's disease DaTSCAN images using a partial least squares scheme and the fisher discriminant ratio., in: KES, 2012, pp. 2241–2250.

[39] I. A. Illán, J. M. Górriz, J. Ramírez, F. Segovia, J. M. JiménezHoyuela, S. J. O. Lozano, Automatic assistance to parkinsons disease diagnosis in datscan spect imaging, Medical Physics 39 (10) 5971–5980. arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.4742055, doi:10.1118/1.4742055.
URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.4742055

[40] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, 2005, pp. 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

[41] D. G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, 1999, pp. 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410.

[42] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, CoRR abs/1710.05381.

[43] X. Zhu, I. Davidson, Knowledge Discovery and Data Mining: Challenges and Realities, IGI Global, Hershey, PA, USA, 2007.

[44] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 973–978.
URL http://dl.acm.org/citation.cfm?id=1642194.1642224

[45] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-sne.

[46] F. Chollet, et al., Keras, https://github.com/fchollet/keras (2015).

[47] R. Kohavi, G. H. John, Wrappers for Feature Subset Selection (1995).

[48] A. Brahim, J. Ramírez, J. Górriz, L. Khedher, D. Salas-Gonzalez, Comparison between Different Intensity Normalization Methods in 123I-Ioflupane Imaging for the Automatic Detection of Parkinsonism, PLoS One 10 (6: e0130274) (2015) 1–20. doi:10.1371/journal.pone.0130274.

[49] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. G. Río, M. Moreno-Caballero, Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease, Expert Systems with Applications 40 (7) (2013) 2756 – 2766. doi:https://doi.org/10.1016/j.eswa.2012.11.017.
URL http://www.sciencedirect.com/science/article/pii/S0957417412012274

[50] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[51] A. Ortiz, F. Lozano, J. Górriz, J. Ramírez, F. Martínez-Murcia, Discriminative sparse features for Alzheimer's disease diagnosis using multimodal image data, Current Alzheimer Research 15 (1) (2018) 1–24.

[52] A. Ortiz, J. Munilla, J. M. Górriz, J. Ramírez, Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease, International Journal of Neural Systems 26 (07) (2016) 1650025.

[53] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, A. Martínez-Álvarez, Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps, Knowledge-Based Systems 71 (2014) 322 – 338.

[54] M. Liu, D. Zhang, D. Shen, for the Alzheimer's Disease Neuroimaging Initiative, Ensemble sparse classification of alzheimer's disease, Neuroimage 60 (2) (2012) 1106–1116.

[55] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR abs/1312.6034.