

Interpreting deep learning prediction of the Parkinson's disease diagnosis from SPECT imaging

Theerasarn Pianpanit^a, Sermkiat Lolak^b, Phattarapong Sawangjai^a, Apiwat Ditthapron^c, Sanparith Marukatat^d, Ekapol Chuangsuwanich^e, Theerawit Wilaiprasitporn^{a,*}

^a*School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Rayong, Thailand*

^b*Department of Neurosurgery, Taksin Hospital, Bangkok, Thailand*

^c*Computer Department, Worcester Polytechnic Institute, Worcester, MA, USA*

^d*National Electronics and Computer Technology Center, Patumthani, Thailand*

^e*Computer Engineering Department, Chulalongkorn University, Bangkok, Thailand*

Abstract

Parkinson's disease (PD) diagnosis mainly relies on the visual and semi-quantitative analysis of medical imaging using single-photon emission computed tomography (SPECT) with 123I-Ioflupane (DaTSCAN). Deep learning approach has the benefits over other machine learning methods because the model does not rely on feature extraction. However, the complexity of the deep learning model usually results in difficulty of interpretation of the model when uses in clinical. Several interpretation methods were created for this approach to show the attention map which reveals important features of the input data, giving the model interpretability. However, it is still unclear whether these methods can be applied to explain PD diagnosis or not. In this paper, four different models of the deep learning approach based on the 3-dimensional convolution neural network (3D-CNN) of well-established architectures have been trained. All the models give high classification performance of PD diagnosis with accuracy up to 95-96%. These four models have been used to evaluate the interpretation performance of six well-known interpretation methods. In general, radiologists interpret SPECT images for a healthy subject by confirming a homogeneous symmetrical comma type

*Corresponding author

Email address: `theerawit.w@vistec.ac.th` (Theerawit Wilaiprasitporn)

shape of the I123-Ioflupane uptake in the striatal nuclei. Any other shape is interpreted as abnormal. To evaluate the interpretation performance, the segmented striatal nuclei of the SPECT images are chosen to be the ground truth. Guided backpropagation which is one of the interpretation methods shows the best performance among all other methods. Guided backpropagation has the best performance to generate the attention map that focuses on the location of striatal nuclei. By using the result from guided backpropagation, 3D CNN architecture that has the highest classification and interpretation performance can be chosen for SPECT diagnosis.

Keywords: Parkinson’s disease, Diagnosis, Convolutional Neural Network, Explainable model, Deep learning

1. Introduction

Parkinson’s disease (PD) is a chronic neurodegenerative disease caused by the nigrostriatal pathway degeneration and leads to the insufficiency of dopamine in the striatum [1]. The characterization of the disease based on the motor symptoms are tremor, rigidity, and bradykinesia. Moreover, the non-motor symptoms which are depression, apathy, and sleep disorder, are frequently recognized. These symptoms degrade the quality of life of the people who suffer from this disease [2]. Early and accurate diagnosis is crucial for effective treatment. The use of I123-Ioflupane SPECT or sometimes known as DaTSCAN or [123I]FP-CIT images has become reliable as one of the standards for the PD diagnosis [3]. The I123-Ioflupane has a high binding affinity for presynaptic dopamine transporters (DAT) inside the striatum. Healthy subjects are characterized by intense and symmetric uptake of the I123-Ioflupane in the caudate nucleus and putamen in both hemispheres. The striatal transaxial images should appear as the symmetric comma- or crescent-shaped. On the other hand, PD subjects are indicated by the unilateral or bilateral decrease in the uptake of the I123-Ioflupane and usually with more depletion in the putamen rather than the caudate nucleus. The striatal transaxial image often shrinks to a circular or oval shape on one or both sides. In clinical practice, diagnosis using SPECT images is usually evaluated visually and sometimes includes assistance from the semi-quantification method, which relies on computer software to acquire quantification of SPECT images [4].

The study of automated computer-aided diagnosis (CAD) of PD currently focuses on the supervised machine learning algorithm, which receives multi-dimensional input features. The machine learning methods for SPECT images classification between healthy and PD subjects from several studies show very high accuracy normally above 90% [5]. The commonly used features are the striatal binding ratios (SBR) from both left and right caudate and putamen which relate to the ratio of the target region and the reference region. These features were classified with the probabilistic neural network, decision tree [6], and support vector machine (SVM) [7, 8]. Other new methods have been developed to find the features from region of interest (ROI), including shape analysis and surface fitting [9], mean ellipsoid uptake and dysmorphic index [10], Haralick texture features [11], principal component analysis (PCA) [12], independent component analysis (ICA) [13], partial least squares decomposition [14], and empirical mode decomposition with PCA or ICA [15]. These new types of features seem to give the best accuracy with SVM classifier. Furthermore, the image voxels within the region of interest are also used directly as the input features with SVM [16, 17], logistic lasso [18], and single-layer neural network [19] classifier.

Conventional supervised machine learning for the CAD faces the difficulty to process the images in their original form. Hand-engineering is needed to select the region of interest that leads to appropriate features in which the classifier can detect the patterns. Deep convolutional neural network (CNN) which does not rely heavily on hand-engineering has recently become a mainstream method for solving image classification problems [20, 21]. The CNN which composes of the convolutional and pooling layers is inspired by the receptive fields in the visual cortex [22]. The resemblance of the CNN and the primate visual stimuli processing has also been evaluated by using the features of the last convolutional layer from the CNN and the inferior temporal cortex neural responses [23]. Also, the progress in hardware, software and algorithm parallelization, which result in the reduction of the training time to process a huge collection of multi-dimensional data allows CNN to become a high-performance tool in the medical image recognition [24].

Recent studies relevant to the SPECT images confirm the advantages of CNN over the conventional machine learning model. A deep CNN framework called “PD Net” was trained with the whole volume of SPECT images and discriminated the PD subjects from healthy subjects with classification performance exceeding the evaluation from the experts [25]. Further investi-

gation shows that CNN still gives high classification accuracy even without the need for spatial normalization procedure [26]. However, it is still unclear which regions in the images are being detected by the model and whether the CNN understands the pattern in the same way as the visual interpretation from the expert. Unlike the conventional machine learning models in which each input feature is hand-designed and the models are decomposable into interpretable components, the complexity of the CNN seems to diminish its interpretability. Also, the EUs General Data Protection Regulation (GDPR), Recital 71, which gives citizens a “right to explanation” will make the “black box” approaches hardly suitable in clinical diagnosis [27].

Several CNN model interpretation methods have been developed to visualize or interpret the CNN so that the attention map can be generated to understand the important pixels of the input image. This allows the model to become interpretable. These methods were used to increase the credibility of the CNN diagnosis results in several types of medical image [26, 28, 29]. However, due to the variety of model interpretation methods, there is no evidence of which methods can provide the most reliable interpretation result for the application of the medical images.

In this work, we train four 3D-CNN models based on PD Net and compare the classification performance for the PD diagnosis. Then, we explore the interpretation performance of six well-known interpretation methods with these four models. The most reliable interpretation method suggests the best 3D-CNN architecture that has the highest credibility. This study is the first attempt to explore the interpretation methods that assist in the design of the 3D-CNN model that has both high performance in classification and interpretation for the diagnosis of PD using the SPECT image. The rest of the paper is organized as follows. We describe the SPECT data, 3D-CNN models, interpretation methods and experiment procedure in [section 2](#). We present the results and information about the performance in [section 3](#). Finally, we conclude this study in [section 4](#).

2. Materials and Method

2.1. PPMI dataset and image preprocessing

Data that were used in this study were obtained from Parkinsons Progression Markers Initiative (PPMI) database [30]. PPMI is a study from the collaboration of research centers designed to identify PD progression biomarkers and to provide essential tools to improve PD therapeutics.

All SPECT scan data acquired from every center undergo the same pre-processing procedure before they are publicly shared via the database [31]. SPECT raw projection data was imported to a HERMES¹ system for iterative reconstruction using the HOSEM software. Iterative reconstruction was done without applying any filter. The HOSEM reconstructed files were then transferred to PMOD² for further processing. Attenuation correction ellipses were drawn on the images and a Chang 0 attenuation correction was applied. The final 3D-volume SPECT image with the voxel size of $2 \times 2 \times 2$ mm³ and the dimension of $91 \times 109 \times 91$ can be directly downloaded from the publicly shared PPMI database.

2.2. Striatal Binding Ratio

The SBR [31] was calculated by first applying the standard Gaussian 3D 6.0 mm filter to the final preprocessed images. These images were then normalized to standard Montreal Neurologic Institute (MNI) space, so that all scans are in the same anatomical alignment. This was followed by identifying the transaxial slice with the highest striatal uptake. Then the 8 hottest striatal slices around it were averaged to generate a single slice image. Regions of interest (ROI) were then selected for left and right caudate, and left and right putamen. The occipital cortex was selected as the reference region. Count densities for each region were extracted for each region and SBR is calculated as

$$\text{SBR of target region} = \frac{\text{Target region count density}}{\text{Reference region count density}} - 1. \quad (1)$$

The SBR of each subject can be obtained from PPMI database alongside with the SPECT images. The SBR can be used as the input feature for any type of simple classifier. In this work, support vector machine (SVM), which gives very high accuracy when apply to SBR [8], is selected as the baseline of the conventional machine learning method to compare with the deep learning approach.

2.3. Convolution Neural Network architectures

In this work, we focus on model interpretation rather than designing a novel network architecture. Hence, we adopted a deep CNN designed for

¹Hermes Medical, Stockholm, Sweden

²PMOD Technologies, Zurich, Switzerland

PPMI dataset with the same purpose, called PD Net [25]. In the original PD Net, zero padding was applied to make the size of the image to be equal in all dimension. However, this study does not include the zero padding so that the images are all in their original form. Thus, a slight modification of the filter size is made in our model.

PD Net model is composed of three 3D convolution layers connected with a single fully connected layer. Each 3D convolution layer has a different setup of filter size and stride, but all 3D convolution layers have Rectified Linear Unit (ReLU) activation layer and max-pooling layer with $(3 \times 3 \times 3)$ pool size and stride of 2 attached. The first 3D convolution layer has 16 filters with a size of $(7 \times 7 \times 7)$ and a stride of 4. After the first pooling, the images are fed to the second 3D convolution layer which has 64 filters with a size of $(5 \times 5 \times 5)$ and a stride of 1. Finally, a 3D convolution layer with 256 filters of size $(2 \times 3 \times 2)$ and a stride of 1 is attached. This layer produces 256 features which then fully-connect to 2 output node to discriminate the extracted features as illustrated in the left hand side of Figure 1.

In addition to PD Net, we modify PD Net architecture by increasing the network depth as shown in the right hand side of Figure 1. We refer this model as “Deep PD Net”. In this model, the filter size of both 3D convolution layer and max-pooling layer were designed so that the last layer before the fully-connect layer gives 256 features, the same as PD Net.

Batch normalization was proposed to accelerate the training of CNN and was first applied with the image classification task [32]. It can achieve the same accuracy with a much lower learning rate, thus it reduces the number of epochs for training. The batch normalization layer was added to follow each ReLU layer. This study incorporates four different 3D-CNN architectures to be compared in both classification and interpretation performance.

2.4. Training parameters

All the models were implemented with Keras [33], an open source deep learning library written in Python and running on top of Tensorflow [34]. The models were trained for 30 epochs using Stochastic Gradient Descent. The momentum parameter was set to 0.9. The learning rate was initially 1×10^{-4} and logarithmically decreased to have 1×10^{-6} at the final epoch. Additionally, weight parameters in the model were initiated with a Glorot initialization [35].

2.5. Model Interpretation Methods

Due to the black box nature of CNN, using direct investigation to the model cannot explain the importance of the input features that lead to high classification performance. Model interpretation methods have been used in revealing the feature importance and assessing trust of the model prediction results. Hence, the main purpose of the interpretation method is to calculate the “contribution score” [36] of the input features. Vastly used model interpretation methods for CNN can be categorized into two major groups. First one is the gradient-based method which focuses on using backpropagation to calculate the gradient that can be implied back to be the contribution score of input features for the target class. The other group is the additive attribution methods which alternatively construct a simpler model to explain the complex model. Well-known current methods belonging to these two major groups are discussed below.

2.5.1. Gradient based method

The core concept of deep learning is to calculate the gradient of loss function respect to all the weights and biases of the model. These gradients can be used to compute the relation between the input feature and the output prediction class. We categorize the interpretation methods that directly use these gradients from the original model as the gradient based method.

Direct backpropagation (Saliency map): Backpropagation is a method to compute gradients of the loss function for all weights in the network. These gradients can also be backpropagated to the input data layer which contributes the most to the assigned class. This is done by computing the gradient of the output category with respect to a sample input image [37]. If we define input features as x and score for predicting class c as S^c , the map of the contribution score is calculated as

$$L_{\text{Saliency map}}^c = \frac{\partial S^c}{\partial x} \quad (2)$$

Guided backpropagation: For the direct backpropagation, the gradient of the loss function with respect to the parameter of layer $l + 1$ is used to calculate the gradient of loss function with respect to the parameter of layer l . In guided backpropagation, the same calculation with the direct backpropagation is used, but if the gradient of layer $l + 1$ is negative, the gradient of layer l is set to zero [38]. In other word, this method includes the

guidance signal to deeper layer during the backpropagation. This results in the remarkable improvement of the contribution score map.

Grad-CAM: Global average pooling (GAP) is the sum of all the values in a feature map at the last convolution layer. It was proposed to replace the fully-connected layers of the CNN. The use of GAP reduces the total model parameters and results in the prevention of the overfitting from the fully-connected layers in some cases. For a 2D input image, the GAP of the k^{th} feature map A^k can be calculated from the sum of over all the 2D elements i, j or can be written as

$$G^k = \sum_i \sum_j A_{ij}^k. \quad (3)$$

The score of predict class c then becomes

$$S^c = \sum_k \sum_i \sum_j w_k^c A_{ij}^k, \quad (4)$$

where w_k^c is the weight of A_{ij}^k to predict class c . By examining this equation, class activation map (CAM) can be defined as

$$\text{CAM} = \sum_k w_k^c A_{ij}^k, \quad (5)$$

which shows the 2D map of the score that predict class c . CAM represents for the contribution score of the input feature by resizing this 2D map to the original input image. It also has a remarkable ability for object localization of the predict class [39]. However, the structure of GAP tends to reduce the model classification performance. The Gradient-weighted Class Activation Mapping (Grad-CAM) which is a generalized form of CAM was proposed to handle the issue [40]. Grad-CAM directly calculates the gradient using the backpropagation from each neuron of the the last convolution layer feature map, which can be written as $\partial S^c / \partial A_{ij}^k$. Then, these gradients are sum within the k^{th} feature map to generate weight of each map and predict class c , which can be written as;

$$\alpha_k^c = \sum_i \sum_j \frac{\partial S^c}{\partial A_{ij}^k} \quad (6)$$

Then Grad-CAM of class c can be generated from

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right). \quad (7)$$

ReLU function is used to remove the negative contribution scores because Grad-CAM wants to consider only the input features that increase the prediction score of class c . Due to the direct use of the gradient from the back-propagation, Grad-CAM can be applied for the interpretation of any types of CNN (e.g. CNN with recurrent neural networks) without any modifications to the CNN model.

Guided Grad-CAM: The use of the last convolution layer of the Grad-CAM can provide more accurate location of the relevant image regions. However, this last layer does not maintain enough resolution to provide fine-grained importance feature. Although, the guided backpropagation method provides the contribution scores of every individual pixel of the input image, it lacks the localization capability. In order to get the best outcome, it is possible to fuse guided backpropagation with Grad-CAM to create Guided Grad-CAM that has both high-resolution and high capability to locate the related image area.

2.5.2. Additive feature attribution method

When the model becomes more complex, the original model can hardly be used to explain its results. The best way to explain the model is to generate a simpler explanation model from the approximation of the original model. By giving $f(x)$ to be the original model, x to be the original input, $g(x')$ to be the explanation model, and x' to be the simplified input, the equation used to explain the original model can be written as $g(x') = f(x)$. The simplified input must be able to map to the original input through a mapping function $x = h_x(x')$. The simplest way to represent the explanation model is to let the simplified input be the binary vector, which represents the presence or absence of the input features. For the image classification task, these input features can be the pixels or super-pixels. This method of generating the explanation model is defined as the additive feature attribution method [41, 42], in which the explanation model g is written as

$$g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i, \quad (8)$$

where $x' \in \{0, 1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$. This method approximates the output $f(x)$ by using ϕ_i which is the “attribution” or “contribution score” from each input feature. Two well-known interpretation methods which are based on the concept of Eq. (8) are discussed below.

DeepLIFT: Deep Learning Important FeaTures (DeepLIFT) is an interpretation method that avoids discontinuity of the gradient-based approach in the approximation of the feature contribution to the output [36]. By giving reference to the input and output, the contribution scores can be calculated from the difference using this reference. If x_i and $f(x)$ are input feature and model output, x_{i0} and $f(x_0)$ are reference input feature and reference model output, then $\Delta y = f(x) - f(x_0)$ and $\Delta x_i = x_i - x_{i0}$ are defined as the difference between the reference and model output and input feature. DeepLift assigns the attribution of Δx_i as $C_{\Delta x_i \Delta y}$ and uses the summation of these attributions to give the value of Δy , which can be written as;

$$\sum_{i=1}^M C_{\Delta x_i \Delta y} = \Delta y. \quad (9)$$

By comparing this with Eq. (8) with $f(x_0) = \phi_0$ and $C_{\Delta x_i \Delta y} = \phi_i$, DeepLIFT can be categorized as the additive feature attribution method. DeepLIFT uses rules, that are based on the structure of deep learning network, to assign the attribution from each input feature. Thus, DeepLIFT is “model-specific” in the approximation of the contribution score. DeepLIFT also shown to be the modify form with better performance compare to another model-specific method called “layer-wise relevance propagation” [43].

SHAP: SHapley Additive exPlanation (SHAP) was designed to simplify any complex model, not restricted to any model structure [41]. For SHAP, Shapley values are used for the contribution score and they are the only set of values that satisfy the properties of the additive feature attribution or Eq. (8). SHAP proposes a way to approximate the Shapley value by minimizing the objective function that satisfies all the properties of Eq. (8). This objective function does not constrain to any model parameters and only use the result from model output. Thus, SHAP becomes “model-agnostic” in the approximation of the contribution score.

2.6. Experiment

Clinical characteristics of the subjects are summarized in Table 1. Since PPMI is the longitudinal study of the PD subject, only the earliest SPECT image was selected for each subject. After obtaining SPECT images from PPMI, the min-max normalization in the range $[0, 1]$ is applied. The data were divided into training, validation, and testing set with a ratio of 80:10:10. During the training, the model use the validation set to tune the model

to reach to the best classification performance. The experiment is carried out using 10-fold cross-validation. The best model that the validation set provides in each fold is used to calculate both classification and interpretation performance by applying on the testing set.

3. Results and Discussion

3.1. Classification performance

The classification performance of each model is reported using the 10-fold cross-validation. In addition to the accuracy, sensitivity and specificity are used as metrics to compare each model. They are defined as

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{Total positive}}, \quad (10)$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{Total negative}}. \quad (11)$$

Results that were acquired by using SBR as the input feature along with the SVM classifier were used as the benchmark to compare with the deep learning method, which uses whole volume SPECT image as the input feature with 3D-CNN as the classifier. Four types of 3D-CNN architecture were designed based on the PD Net [25] and all of them are described in the previous section. The mean \pm STD of accuracy, sensitivity, and specificity which were calculated from 10-fold of a testing set, are shown in Table 2. The accuracy varies from 95% to 96% with the deep learning approaches, giving a slightly higher accuracy compared to the SVM model. Deep PD Net with batch normalization has the highest accuracy with 96.87%. In this result, attaching the batch normalization shows minor improvement of the model accuracy. This may result from the small value of the learning rate set in this study. Also, the input data may not be complex enough compared to the results of the original paper [32].

From the clinical details shown in Table 1, the number of PD subjects is 3 times higher than the number of healthy subjects. Due to this extreme class imbalance, the sensitivity of each model was not significantly different. However, we can see the improvement of the specificity of Deep PD Net compare to other models.

McNemar’s test [44] was used to compare between SVM and CNN models and the p -value from this test can not reveal any statistical difference in the

classification performance. Thus, we further investigate the ROC curve as can be seen in Figure 2. It reveals a trend of a higher AUC value of 3D-CNN than that of SVM. Also, the Deep PD Net with batch normalization has the highest AUC value which equals to 0.987.

3.2. Interpretation performance

To evaluate the interpretation performance, we generated a ground truth image by segmenting the striatal nuclei. This ground truth image is compared with the attention map from the interpretation methods. The segmented striatal nuclei are created based on a previous study [9]. The slices from 35th to 48th of the SPECT image which cover the striatal nuclei are selected. Then, each slice is normalized to the range from 0 to 1, and a slice averaging image is constructed. This slice averaging image is again normalized to [0,1]. After that, a threshold that determines the segmented area is selected. The mean \pm SD of the thresholds for healthy subjects and PD subjects, which were selected by the experts, were reported in Prashanth et al. [9] as 0.63 ± 0.04 and 0.69 ± 0.05 respectively. In this work, we select the mean threshold values and use them to find the segmented striatal nuclei of the slice averaging image. The results of the slice averaging SPECT images from healthy and PD can be seen in Figure 3. The area that is enclosed by the red irregular ellipse represents for the segmented area. The segmented area is now used as the ground truth to evaluate the interpretation performance.

The slice averaging of the attention map from the interpretation method was also generated similar to the slice averaging of the SPECT images. Examples of grayscale attention map from the Deep PD Net model are shown in the first row of Figure 3 (a) and (b) for a healthy subject and a PD subject respectively. White regions show the most contributed area in the class prediction and they are located near or inside the segmented region.

The pixels that are used to evaluate the interpretation performance need to be selected with another threshold. Shrikumar et al. [36] proposed the threshold of which using only 20% of top values sorted from descending order. In this study, this thresholding technique was used with altering percentages of 10% and 1%. Then two binary images can be generated from an attention map. These binary images for different interpretation methods are shown in the second and third row of Figure 3 (a) and (b) respectively. These figures demonstrate the overlap region between each interpretation method and the segmented area significantly. By considering the figure of top 10% pixels as seen in the second row, we can observe that the majority of the pixels are

located inside the brain area. On the other hand, the results from using the top 1% as seen in the third row show that majority of pixels gather inside the segmented red line area.

Dice coefficient D is widely used as a measure for comparison of a predicted segmented image P with the ground truth segmented image G . It is defined as twice as the size of the intersect area between P and G over the sum of the area of P and G , and can be written as

$$D = \frac{2|P \cap G|}{|P| + |G|}. \quad (12)$$

The coefficient exists in the range of $[0, 1]$ where $D = 1$ indicates identical segmentation. The mean \pm SD of the Dice coefficient is calculated from the test set of all 10-fold. The results are shown in Table 3. The bold value indicates the best result in a given threshold. The upper and lower tables show the results from top 10% and top 1% respectively. The uses of the top 10% and 1% show that guided backpropagation has the highest Dice coefficient which directly relates to the interpretation performance in providing the information of the location of striatal nuclei. The boxplots of the Dice coefficient in Figure 4 also confirm that guided backpropagation performance dominates other methods.

Wilcoxon signed-rank test was used to compare the guided backpropagation with the other methods. It revealed significant differences ($p < 0.01$) for the Dice coefficient between guided backpropagation and all other methods. Then, the test was done to compare the Dice coefficient of guided backpropagation between Deep PD Net and PD Net. We also found the significant difference ($p < 0.01$) of this model. However, the difference between Deep PD Net and Deep PD Net with batch normalization was not significant.

Grad-CAM was the only method that barely focuses on this region. Although Grad-CAM was supposed to perform well in the class-discriminative and localize relevant image regions [40], in this comparison, it seems to lack the ability to show fine-grained importance like guided backpropagation.

Mean absolute error is used as another measure to evaluate the performance between each method as can be seen in Figure 5. The guided backpropagation, which was first designed to improve the quality of the saliency map in feature visualization of deep learning model [38], gives much less error compared to other methods. Inside the striatal nuclei, the error approaches zero which can be interpreted as the Deep PD Net directly focuses on the region and gives more credibility in the prediction results.

We generate a mean segmented image from the binary map of top 1% pixels and overlay on top of the ground truth segmented image as shown in Figure 6. By examining Figure 6(a), saliency map and DeepLIFT can identify the tail of symmetric comma shape in the control group. On the other hand, from Figure 6(b), SHAP can correctly identify the uptake depletion location of the PD group. Since SHAP is the model-agnostic method, it does not rely on 3D-CNN weights, thus the result becomes different from the saliency map and DeepLIFT. These three methods outperform other methods in discriminating the difference between PD and healthy subjects. Even though the performance of these three methods are hard to evaluate, SHAP can generate better quality heatmap at the uptake depletion location, which outperforms other methods in discriminating the difference between PD and healthy subjects. This should consistent with Lundberg et al. [41], which revealed that SHAP gives the best performance among all other methods of showing the class difference between hand-written images of numbers 8 and 3.

By using the results from the comparative study of interpretation methods, the improvement of the interpretation performance can be found by modifying PD Net to Deep PD Net. The guided backpropagation results from Table 3 show that Deep PD Net has the highest interpretation performance. These results suggest that interpretation methods can help in increasing the credibility of the model when the model is modified. We suggest a flow chart for the application of the interpretation method to increase the credibility of the model in Figure 7. In this study, if we follow this flow chart, the Deep PD Net model should be suggested to be used for PPMI data. Furthermore, several studies also have investigated both decreasing [5] and increasing [25] in the classification performance, when applying the well-designed machine learning model from PPMI data to the local data. However, without interpretation methods, it becomes unclear about the feature that affects the change in classification performance.

4. Conclusions

In this work, the Dice coefficient is introduced for the evaluation of the interpretation performance of the interpretation method. The result of the Dice coefficient suggests that guided backpropagation has the highest interpretation performance for the PD diagnosis. By using the mean segmented binary image overlay on top of ground truth segmented images, a significant

Table 1: Clinical details of all subjects used in this study.

	Parkinson’s disease (n=448)	Healthy Control (n=159)
Age	61.6 \pm 9.8	60.5 \pm 11.3
Sex (M/F)	288/160	112/47
MDS-UPDRS part III	21.3 \pm 9.5	
Hoehn and Yahr stage	1.6 \pm 0.5	

Table 2: Classification performance of SVM, PD Net and Deep PD Net.

Method	Input Feature	Accuracy	Sensitivity	Specificity
SVM	SBR Ratio	95.55 \pm 2.48	96.90 \pm 2.61	92.29 \pm 7.73
PD Net	SPECT	95.39 \pm 2.88	95.97 \pm 3.30	93.75 \pm 6.23
PD Net + Batch Norm	SPECT	96.54 \pm 2.63	96.88 \pm 3.20	95.66 \pm 6.09
Deep PD Net	SPECT	96.71 \pm 2.32	97.10 \pm 2.35	95.42 \pm 4.40
Deep PD Net + Batch Norm	SPECT	96.87 \pm 2.13	96.42 \pm 3.01	97.89 \pm 3.61

result from the saliency map, DeepLIFT, and SHAP in discriminating the different features between healthy and PD subjects was obtained. Those three methods correctly show the uptake depletion regions of PD subjects which is the main characteristic of Parkinson’s disease. Furthermore, when using the results of the comparative study of the interpretation method, Deep PD Net is shown to have an improvement in both the classification and interpretation performance compared to the original PD Net model. Considering these results, we can infer that guided backpropagation can assist in the modification of the model to increase the credibility of PD diagnosis.

Table 3: The results of the mean Dice coefficient using the binary image of the attention map for the top 10% of contribution scores (upper) and the top 1% of contribution scores (lower). The bold number refers to the highest Dice coefficient among all the methods.

Model	Saliency Map	Guided Backprop	Grad-CAM	Guided Grad-CAM	DeepLIFT	SHAP
PD Net	17.09 \pm 4.91	23.78 \pm 6.02	3.27 \pm 8.11	22.15 \pm 7.36	16.92 \pm 6.75	16.62 \pm 6.77
PD Net + Batch Norm	12.51 \pm 4.99	23.90 \pm 6.76	4.49 \pm 7.89	20.73 \pm 8.53	14.98 \pm 6.04	15.50 \pm 6.85
Deep PD Net	17.29 \pm 5.00	29.72 \pm 8.95	3.66 \pm 7.77	25.70 \pm 10.15	18.11 \pm 6.59	15.72 \pm 9.60
Deep PD Net + Batch Norm	15.22 \pm 4.36	29.38 \pm 9.00	2.96 \pm 6.49	21.11 \pm 12.16	16.99 \pm 5.23	16.35 \pm 9.39

Model	Saliency Map	Guided Backprop	Grad-CAM	Guided Grad-CAM	DeepLIFT	SHAP
PD Net	38.38 \pm 10.73	53.08 \pm 10.42	1.45 \pm 5.96	49.32 \pm 16.69	32.53 \pm 11.53	26.73 \pm 11.20
PD Net + Batch Norm	22.20 \pm 9.38	54.85 \pm 10.12	1.85 \pm 6.59	47.91 \pm 19.62	26.73 \pm 10.27	22.63 \pm 11.19
Deep PD Net	45.32 \pm 10.02	66.07 \pm 12.62	1.45 \pm 5.99	58.87 \pm 23.86	36.96 \pm 11.00	25.81 \pm 15.54
Deep PD Net + Batch Norm	38.37 \pm 10.22	65.56 \pm 12.32	0.96 \pm 5.11	49.00 \pm 28.71	38.71 \pm 10.28	28.15 \pm 15.82

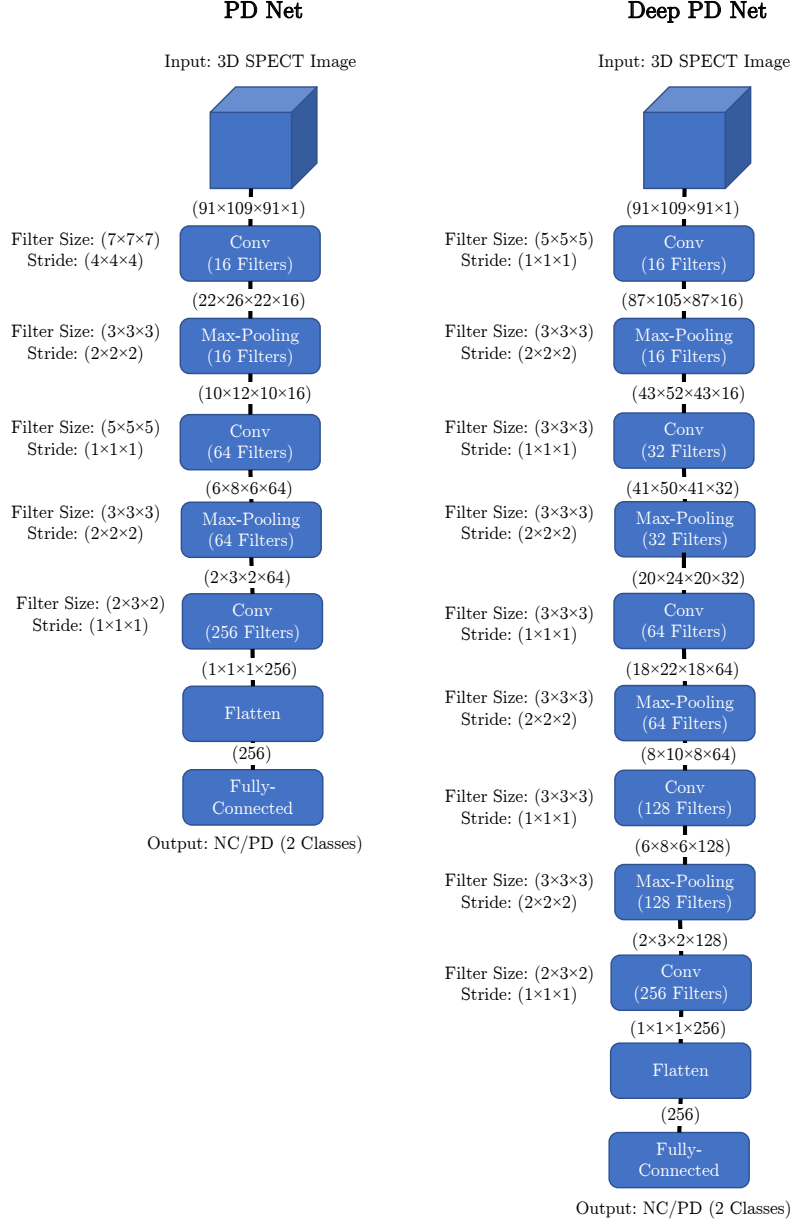


Figure 1: Structure of PD-Net and Deep PD Net used in this work with the details of the size and number of convolution and max-pooling filters. The PD Net has been modified in the last convolution layer so that the image from the database can be used directly without need of the zero-padding.

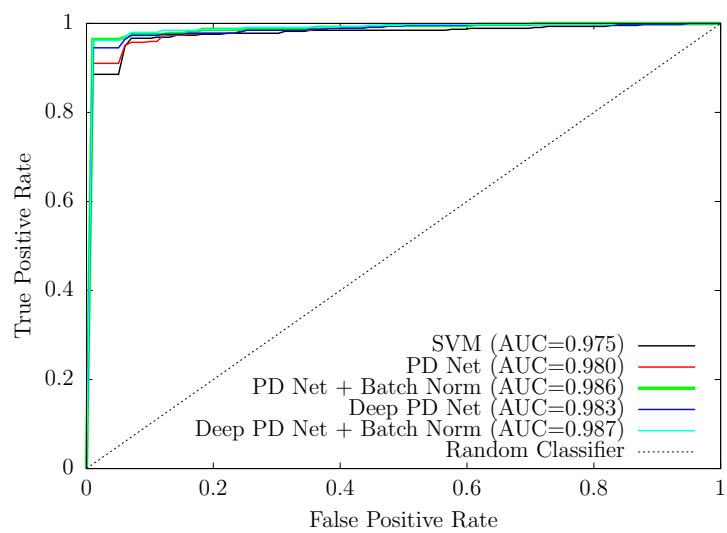


Figure 2: ROC curve for each model.

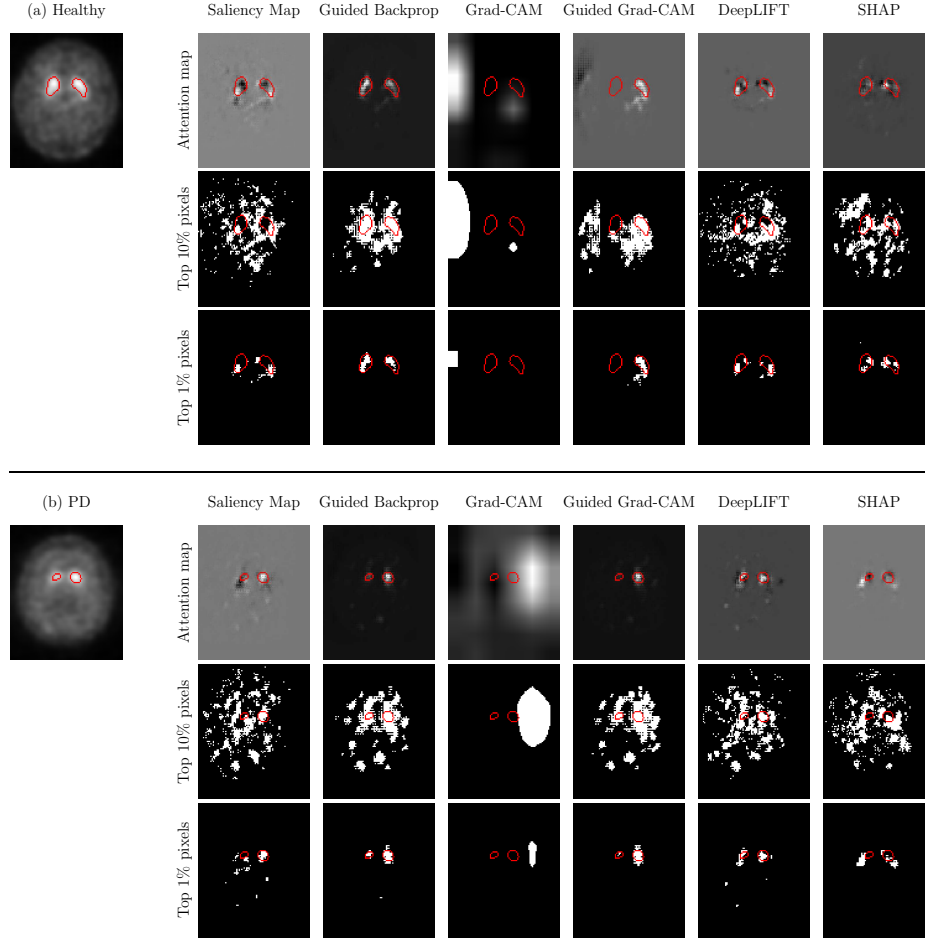


Figure 3: An example of slice averaging SPECT image (left figure) and the attention map (right table) from Deep PD Net model for (a) healthy control and (b) PD. The red line is the segmented line generated from the mean threshold that was reported in Ref. 9. The first row of the right table shows the original map. The second and the third row shows the binary map generated from the top 10% of contribution score and the top 1% contribution score.

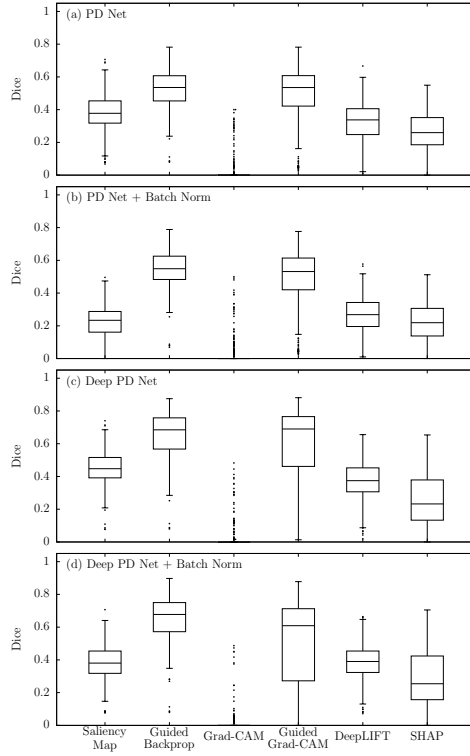


Figure 4: Boxplots of Dice coefficient in different interpretation methods from the top 1% of contribution score for (a) PD Net (b) PD Net + Batch Norm (c) Deep PD Net and (d) Deep PD Net + Batch Norm. Median is the line that locates inside the box and black dots represent outliers outside 1.5 times the interquartile range of the upper and lower quartile.

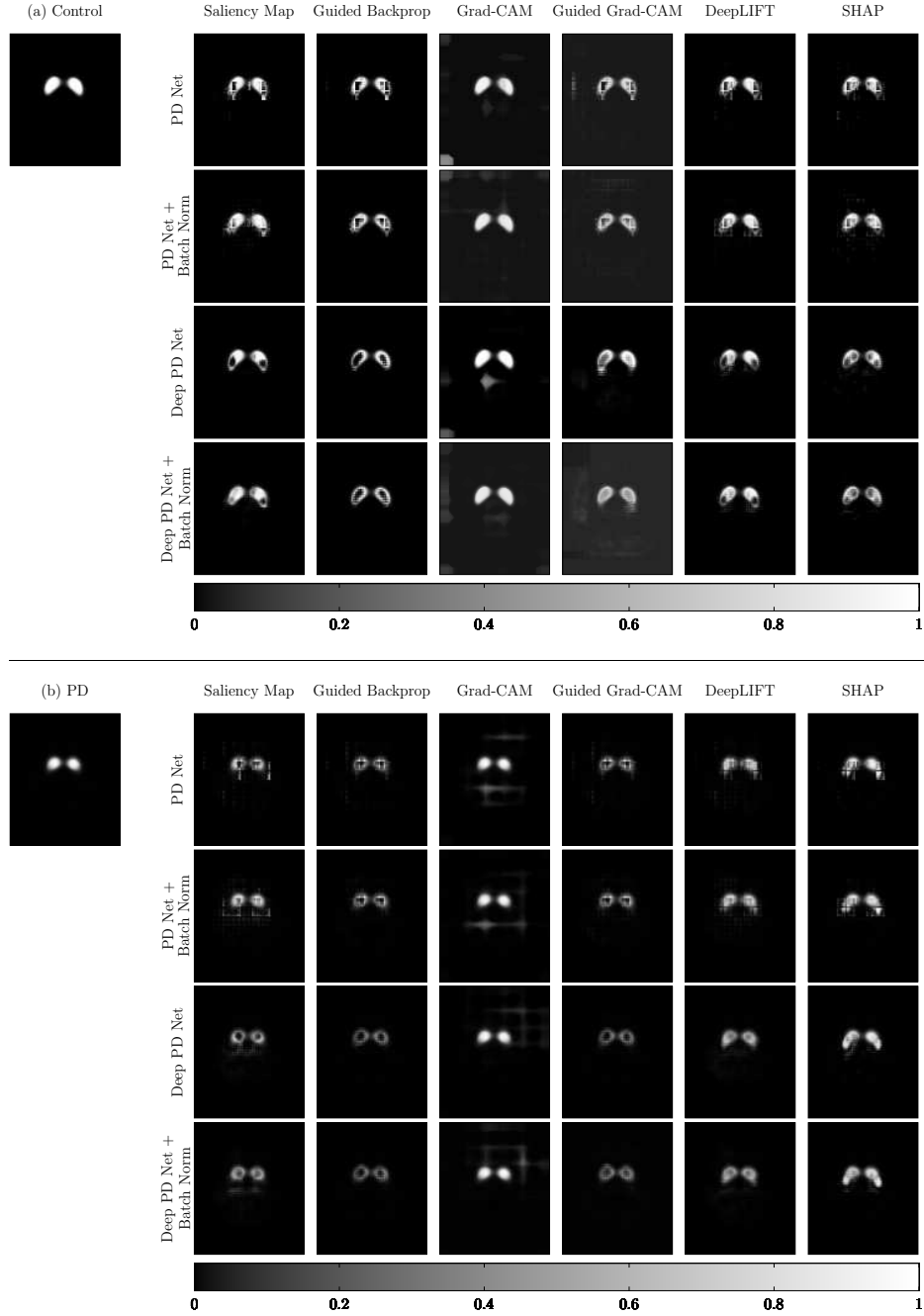


Figure 5: The mean segmented image (left) and mean absolute error plot (right table) for (a) healthy control and (b) PD group. The mean absolute error was calculated using the binary image from the top 1% contribution pixels to compare with the binary image from the segmented image.

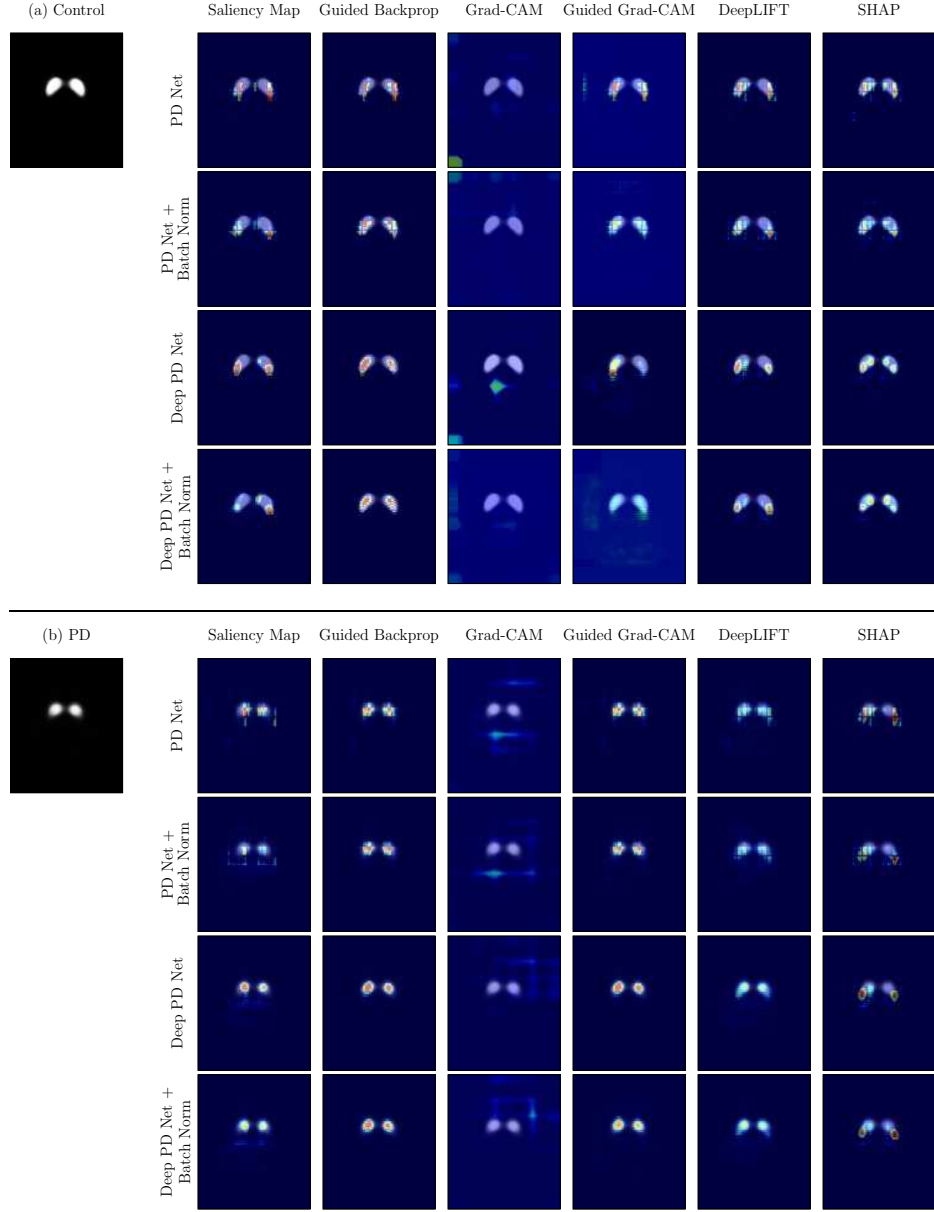


Figure 6: The mean segmented image (left) and mean segmented heatmap (right table) for (a) healthy control and (b) PD group. The mean segmented heatmap was generated using the mean of binary images from the top 1% contribution pixels.

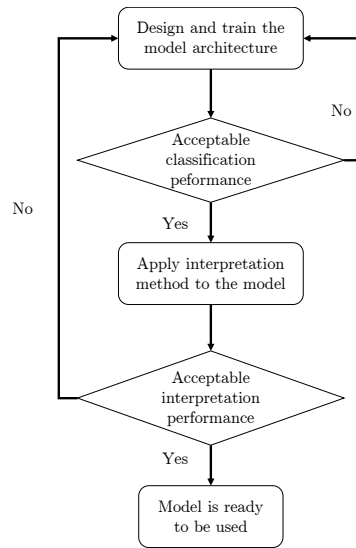


Figure 7: The flow chart for the interpretation method application in increasing the model credibility.

References

- [1] J. A. Obeso, M. C. Rodriguez-Oroz, C. G. Goetz, C. Marin, J. H. Kordower, M. Rodriguez, E. C. Hirsch, M. Farrer, A. H. V. Schapira, G. Halliday, Missing pieces in the parkinson's disease puzzle, *Nature Medicine* 16 (2010) 653–661. doi:[10.1038/nm.2165](https://doi.org/10.1038/nm.2165).
- [2] K. R. Chaudhuri, A. H. Schapira, Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment, *The Lancet Neurology* 8 (2009) 464 – 474. doi:[10.1016/S1474-4422\(09\)70068-7](https://doi.org/10.1016/S1474-4422(09)70068-7).
- [3] D. S. Djang, M. J. Janssen, N. Bohnen, J. Booij, T. A. Henderson, K. Herholz, S. Minoshima, C. C. Rowe, O. Sabri, J. Seibyl, B. N. Van Berckel, M. Wanner, SNM practice guideline for dopamine transporter imaging with 123I-Ioflupane SPECT 1.0, *Journal of Nuclear Medicine* 53 (2012) 154–163. doi:[10.2967/jnumed.111.100784](https://doi.org/10.2967/jnumed.111.100784).
- [4] K. Badiavas, E. Molyvda, I. Iakovou, M. Tsolaki, K. Psarrakos, N. Karatzas, SPECT imaging evaluation in movement disorders: far beyond visual assessment, *European Journal of Nuclear Medicine and Molecular Imaging* 38 (2011) 764–773. doi:[10.1007/s00259-010-1664-1](https://doi.org/10.1007/s00259-010-1664-1).
- [5] J. C. Taylor, J. W. Fenner, Comparison of machine learning and semi-quantification algorithms for (i123)fp-cit classification: the beginning of the end for semi-quantification?, *EJNMMI Physics* 4 (2017) 29. doi:[10.1186/s40658-017-0196-1](https://doi.org/10.1186/s40658-017-0196-1).
- [6] B. Palumbo, M. L. Fravolini, S. Nuvoli, A. Spanu, K. S. Paulus, O. Schillaci, G. Madeddu, Comparison of two neural network classifiers in the differential diagnosis of essential tremor and parkinson's disease by (123)I-FP-CIT brain SPECT, *European Journal of Nuclear Medicine and Molecular Imaging* 37 (2010) 2146–2153. doi:[10.1007/s00259-010-1481-6](https://doi.org/10.1007/s00259-010-1481-6).
- [7] B. Palumbo, M. L. Fravolini, T. Buresta, F. Pompili, N. Forini, P. Nigro, P. Calabresi, N. Tambasco, Diagnostic accuracy of parkinson disease by support vector machine (SVM) analysis of (123)I-FP-CIT brain SPECT data: Implications of putaminal findings and age, *Medicine* 93 (2014) e228. doi:[10.1097/MD.0000000000000228](https://doi.org/10.1097/MD.0000000000000228).

- [8] R. Prashanth, S. D. Roy, P. K. Mandal, S. Ghosh, Automatic classification and prediction models for early Parkinsons disease diagnosis from spect imaging, *Expert Systems with Applications* 41 (2014) 3333 – 3342. doi:[10.1016/j.eswa.2013.11.031](https://doi.org/10.1016/j.eswa.2013.11.031).
- [9] R. Prashanth, S. D. Roy, P. K. Mandal, S. Ghosh, High-accuracy classification of parkinson’s disease through shape analysis and surface fitting in 123i-ioflupane spect imaging, *IEEE Journal of Biomedical and Health Informatics* 21 (2017) 794–802. doi:[10.1109/JBHI.2016.2547901](https://doi.org/10.1109/JBHI.2016.2547901).
- [10] A. Augimeri, A. Cherubini, G. L. Cascini, D. Galea, M. E. Caligiuri, G. Barbagallo, G. Arabia, A. Quattrone, CADA—computer-aided DaTSCAN analysis, *EJNMMI Physics* 3 (2016) 4. doi:[10.1186/s40658-016-0140-9](https://doi.org/10.1186/s40658-016-0140-9).
- [11] F. J. Martínez-Murcia, J. M. Górriz, J. Ramírez, I. A. Illán, C. G. Puntonet, Texture features based detection of parkinson’s disease on datscan images, in: J. M. Ferrández Vicente, J. R. Álvarez Sánchez, F. de la Paz López, F. J. Toledo Moreo (Eds.), *Natural and Artificial Computation in Engineering and Medical Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 266–277. doi:[10.1007/978-3-642-38622-0_28](https://doi.org/10.1007/978-3-642-38622-0_28).
- [12] D. J. Towey, P. G. Bain, K. S. Nijran, Automatic classification of i-123-fp-cit (datscan) spect images, *Nucl Med Commun* 32 (2011). doi:[10.1097/MNM.0b013e328347cd09](https://doi.org/10.1097/MNM.0b013e328347cd09).
- [13] F. Martnez-Murcia, J. Grriz, J. Ramrez, I. Illn, A. Ortiz, Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging, *Neurocomputing* 126 (2014) 58 – 70. doi:[10.1016/j.neucom.2013.01.054](https://doi.org/10.1016/j.neucom.2013.01.054).
- [14] F. Segovia, J. M. Grriz, J. Ramrez, I. lvarez, J. M. Jimnez-Hoyuela, S. J. Ortega, Improved parkinsonism diagnosis using a partial least squares based approach, *Medical Physics* 39 (2012) 4395–4403. doi:[10.1118/1.4730289](https://doi.org/10.1118/1.4730289).
- [15] A. Rojas, J. Grriz, J. Ramrez, I. Illn, F. Martnez-Murcia, A. Ortiz, M. G. Ro, M. Moreno-Caballero, Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson

- disease, *Expert Systems with Applications* 40 (2013) 2756 – 2766. doi:[10.1016/j.eswa.2012.11.017](https://doi.org/10.1016/j.eswa.2012.11.017).
- [16] I. A. Illán, J. M. Górriz, J. Ramírez, F. Segovia, J. M. Jiménez-Hoyuela, S. J. Ortega Lozano, Automatic assistance to parkinson’s disease diagnosis in datscan spect imaging, *Medical Physics* 39 (2012) 5971–5980. doi:[10.1118/1.4742055](https://doi.org/10.1118/1.4742055).
 - [17] F. P. M. Oliveira, M. Castelo-Branco, Computer-aided diagnosis of parkinsons disease based on [123 i]fp-cit spect binding potential images, using the voxels-as-features approach and support vector machines, *Journal of Neural Engineering* 12 (2015) 026008. doi:[10.1088/1741-2560/12/2/026008](https://doi.org/10.1088/1741-2560/12/2/026008).
 - [18] H. D. Tagare, C. DeLorenzo, S. Chelikani, L. Saperstein, R. K. Fulbright, Voxel-based logistic analysis of ppmi control and parkinson’s disease datscans, *NeuroImage* 152 (2017) 299 – 311. doi:[10.1016/j.neuroimage.2017.02.067](https://doi.org/10.1016/j.neuroimage.2017.02.067).
 - [19] Y. C. Zhang, A. C. Kagen, Machine learning interface for medical image analysis, *Journal of Digital Imaging* 30 (2017) 615–621. doi:[10.1007/s10278-016-9910-0](https://doi.org/10.1007/s10278-016-9910-0).
 - [20] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436444. doi:[10.1038/nature14539](https://doi.org/10.1038/nature14539).
 - [21] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, 2016.
 - [22] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex, *The Journal of Physiology* 160 (1962) 106–154. doi:[10.1113/jphysiol.1962.sp006837](https://doi.org/10.1113/jphysiol.1962.sp006837).
 - [23] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, J. J. DiCarlo, Deep neural networks rival the representation of primate IT cortex for core visual object recognition, *PLOS Computational Biology* 10 (2014) 1–18. doi:[10.1371/journal.pcbi.1003963](https://doi.org/10.1371/journal.pcbi.1003963).
 - [24] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Snchez,

- A survey on deep learning in medical image analysis, *Medical Image Analysis* 42 (2017) 60 – 88. doi:[10.1016/j.media.2017.07.005](https://doi.org/10.1016/j.media.2017.07.005).
- [25] H. Choi, S. Ha, H. J. Im, S. H. Paek, D. S. Lee, Refining diagnosis of Parkinson’s disease with deep learning-based interpretation of dopamine transporter imaging, *NeuroImage: Clinical* 16 (2017) 586 – 594. doi:[10.1016/j.nicl.2017.09.010](https://doi.org/10.1016/j.nicl.2017.09.010).
 - [26] F. J. Martinez-Murcia, J. M. Grriz, J. Ramrez, A. Ortiz, Convolutional neural networks for neuroimaging in parkinsons disease: Is preprocessing needed?, *International Journal of Neural Systems* 28 (2018) 1850035. doi:[10.1142/S0129065718500351](https://doi.org/10.1142/S0129065718500351).
 - [27] G. Ras, M. van Gerven, P. Haselager, *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*, Springer International Publishing, Cham, 2018, pp. 19–36. doi:[10.1007/978-3-319-98131-4_2](https://doi.org/10.1007/978-3-319-98131-4_2).
 - [28] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature* 542 (2017) 115–118. doi:[10.1038/nature21056](https://doi.org/10.1038/nature21056).
 - [29] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian, R. G. Gonzalez, M. H. Lev, S. Do, An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets, *Nature Biomedical Engineering* 3 (2019) 173–182. doi:[10.1038/s41551-018-0324-9](https://doi.org/10.1038/s41551-018-0324-9).
 - [30] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, W. Poewe, B. Mollenhauer, P.-E. Klinik, T. Sherer, M. Frasier, C. Meunier, A. Rudolph, C. Casaceli, J. Seibyl, S. Mendick, N. Schuff, Y. Zhang, A. Toga, K. Crawford, A. Ansbach, P. D. Blasio, M. Piovella, J. Trojanowski, L. Shaw, A. Singleton, K. Hawkins, J. Eberling, D. Brooks, D. Russell, L. Leary, S. Factor, B. Sommerfeld, P. Hogarth, E. Pighetti, K. Williams, D. Standaert, S. Guthrie, R. Hauser, H. Delgado, J. Jankovic, C. Hunter, M. Stern, B. Tran, J. Leverenz, M. Baca, S. Frank, C.-A. Thomas, I. Richard, C. Deeley, L. Rees, F. Sprenger, E. Lang, H. Shill, S. Obradov, H. Fernandez, A. Winters, D. Berg, K. Gauss, D. Galasko,

- D. Fontaine, Z. Mari, M. Gerstenhaber, D. Brooks, S. Malloy, P. Barone, K. Longo, T. Comery, B. Ravina, I. Grachev, K. Gallagher, M. Collins, K. L. Widnell, S. Ostrowizki, P. Fontoura, T. Ho, J. Luthman, M. van der Brug, A. D. Reith, P. Taylor, The parkinson progression marker initiative (ppmi), *Progress in Neurobiology* 95 (2011) 629 – 635. doi:[10.1016/j.pneurobio.2011.09.005](https://doi.org/10.1016/j.pneurobio.2011.09.005).
- [31] G. Wisniewski, J. Seibyl, K. Marek, DatScan SPECT Image Processing Methods for Calculation of Striatal Binding Ratio (SBR), Technical Report, Institute for Neurodegenerative Disorders (IND), 2013.
 - [32] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, Preprint in arXiv:1502.03167 (2015). [arXiv:1502.03167](https://arxiv.org/abs/1502.03167).
 - [33] F. Chollet, Keras, 2015. URL: <https://keras.io/>.
 - [34] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous distributed systems, Preprint in arXiv:1603.04467 (2016). [arXiv:1603.04467](https://arxiv.org/abs/1603.04467).
 - [35] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Y. W. Teh, M. Titterton (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 249–256. URL: <http://proceedings.mlr.press/v9/glorot10a.html>.
 - [36] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 3145–3153. URL: <http://proceedings.mlr.press/v70/shrikumar17a.html>.

- [37] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, Preprint in arXiv:1312.6034 (2013). [arXiv:1312.6034](#).
- [38] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, Preprint in arXiv:1412.6806 (2014). [arXiv:1412.6806](#).
- [39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929. doi:[10.1109/CVPR.2016.319](#).
- [40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:[10.1109/ICCV.2017.74](#).
- [41] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-pre>
- [42] S. Lundberg, S. Lee, An unexpected unity among methods for interpreting model predictions, Preprint in arXiv:1611.07478 (2016). [arXiv:1611.07478](#).
- [43] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) 1–46. doi:[10.1371/journal.pone.0130140](#).
- [44] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (1998) 1895–1923. doi:[10.1162/089976698300017197](#).