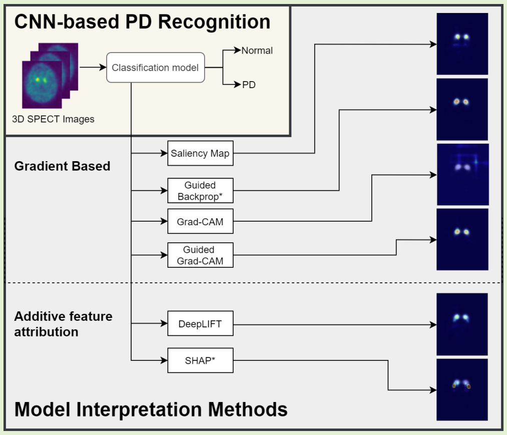
Parkinson's Disease Recognition Using SPECT Image and Interpretable-AI: A Tutorial

Theerasarn Pianpanit, Sermkiat Lolak, Phattarapong Sawangjai, Thapanun Sudhawiyangkul* and Theerawit Wilaiprasitporn*, *Member, IEEE*

***Abstract*—** **Parkinson's disease (PD) diagnosis mainly relies on the visual and semi-quantitative medical imaging analysis using single-photon emission computed tomography (SPECT). The deep learning approach benefits other machine learning methods because it does not rely on feature engineering. However, the deep learning model's complexity usually results in difficult model interpretation when used in clinical. The model interpretability depends on the interpretation method to reveal each pixel's contribution in the input image from an attention map. This tutorial aims to demonstrate the procedure to choose a suitable interpretation method for the PD recognition model. We exhibit four DCNN architectures as an example and introduce six well-known interpretation methods. We categorized the introduced methods into two significant groups. The first one is the gradient-based method, which focuses on using backpropagation to calculate the gradient that implies the input score of the target class's input features. The other group is the additive attribution methods, which alternatively construct a simpler model to explain the predictive model. Finally, we propose an evaluation method to measure the interpretation performance and a method to use the interpreted feedback for assisting in model selection. Shortly, the introduced interpretation methods can contribute to sensor data processing in an AI Era (interpretable-AI) as feedback in constructing well-suited deep learning architectures for specific applications.**



***Index Terms*—Parkinson's disease, SPECT image, computer-aided diagnosis (CAD), explainable AI (XAI), deep learning Tutorial**

I. Introduction

Parkinson's disease (PD) is a chronic neurodegenerative disease caused by the nigrostriatal pathway degeneration and leads to dopamine's insufficiency 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 PD diagnosis standards [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. 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, generally above 90% [5]. Conventional supervised machine learning for the CAD faces the difficulty of processing the images in their original form. Hand-engineering is needed in selecting the region of interest that

This work was supported by PTT Public Company Limited, The SCB Public Company Limited, Thailand Science Research and Innovation (SRI62W1501) and Office of National Higher Education Science Research and Innovation Policy Council (C10F630057).

T. Pianpanit is with Department of Applied Radiation and Isotopes, Faculty of Science, Kasetsart University, Bangkok, Thailand.

S. Lolak is with Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.

P. Sawangjai, T. Sudhawiyangkul, and T. Wilaiprasitporn are with Bio-inspired Robotics and Neural Engineering (BRAIN) Lab, School of Information Science and Technology (IST), Vidyasirimedhi Institute of Science & Technology (VISTEC), Rayong, Thailand.

*corresponding authors: thapanun.s at vistec.ac.th, theerawit.w at vistec.ac.th

leads to appropriate features in which the classifier can detect the patterns. Deep convolutional neural network (DCNN), which does not rely heavily on hand-engineering, has recently become a mainstream method for solving image classification problems [6], [7].

The DCNN composing the convolutional and pooling layers is inspired by the receptive fields in the visual cortex [8]. The resemblance of the DCNN and the primate visual stimuli processing has also been evaluated using the last convolutional layer's features from the DCNN, and the inferior temporal cortex neural responses [9]. In addition, the progress in hardware, software, and algorithm parallelization, which reduces the training time to process a massive collection of multi-dimensional data, allows DCNN to become a high-performance tool in medical image recognition [10], [11]. Further investigation shows that DCNN still gives high classification accuracy even without the need for spatial normalization procedure [12]. However, it is still unclear which regions in the images are being detected by the model and whether the DCNN understands the pattern in the same way as the expert's visual interpretation. 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 DCNN seems to diminish its interpretability. Also, the EU's 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 [13].

Several DCNN model interpretation methods have been developed to visualize or interpret the DCNN so that the attention map can be generated to understand the essential pixels of the input image. These methods were used to interpret the model's decision and increase the credibility of the DCNN diagnosis results in several types of medical image [12], [14], [15]. However, due to the variety of model interpretation methods, there is no evidence of which methods can provide the most reliable interpretation results for medical image applications.

This tutorial aims to demonstrate the procedure for selecting the most suitable interpretation method for SPECT image PD recognition model. The tutorial can be divided into three main parts as follows:

1) Classification models overview and example classification scenario (section II). In this part, we provide a brief introduction to the traditional method and DCNN models for PD recognition. Next, we give an example scenario by training four DCNN models and comparing the classification performance on the PD diagnosis. Note that this tutorial will focus on the interpretation method and will not go deeply into PD classification models. However, this part could give a general idea to those who are new or not actively involved in PD recognition.
2) Model interpretation methods overview and example interpretation scenario (section III). In this part, we provide an overview of the concept of six well-known conventional interpretation methods. To illustrate each interpretation method's difference, we incorporate the interpretation methods to four DCNN models and display each method's visual interpretation result. Lastly, we demonstrate the methods for evaluating the interpretation performance.
3) Discussion on example scenario (section IV). In this part, we provide some insights into the example scenario's interpretation results and how to decide the most suitable interpretation method. We also give a method to utilize the interpreted feedback to aid in model selection.

The code for all four DCNN models with six interpretation methods was uploaded and can be download publicly[1]. Furthermore, the introduced deep neural network interpretation methods can contribute to the future of data processing in an AI Era (interpretable-AI) as one of the core modules in sensors-related studies. For example, Grezmak *et al.* had reported the interpretable CNN for a machine fault diagnosis [16], Alharthi *et al.* had reported an interpretable time series model for gait-induced ground reaction force (GRF) in Parkinson's disease (PD) recognition [17], and Lee *et al.* had utilized interpretable AI in glucose management for diabetes patient [18]. All of the examples demonstrate the usefulness of the model interpretation methods as feedback in constructing well-suited deep learning architectures.

II. PD Recognition Methods and an Example Scenario

A. PPMI Dataset and Image Preprocessing

The public SPECT image dataset commonly used in PD recognition studies [19]–[23] are from Parkinson's Progression Markers Initiative (PPMI) database [24]. 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 preprocessing procedure before they are publicly shared via the database [25]. SPECT raw projection data was imported to a HERMES[2] system for iterative reconstruction using the HOSEM software. Iterative reconstruction was done without applying any filter. The HOSEM reconstructed files were then transferred to PMOD[3] 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 $2 \times 2 \times 2$ mm^3 and the dimension of $91 \times 109 \times 91$ can be directly downloaded from the publicly shared PPMI database.

B. Traditional Classification Method

The most commonly used features for the traditional classification method 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 [26], and support vector machine (SVM) [27],

[1] https://github.com/IoBT-VISTEC/PPMI_DL, We will publish all source codes and data sources immediately after getting an acceptance letter from SJ

[2] Hermes Medical, Stockholm, Sweden

[3] PMOD Technologies, Zurich, Switzerland

[28]. Other new methods have been developed to find the features from region of interest (ROI), including shape analysis and surface fitting [29], mean ellipsoid uptake and dysmorphic index [30], Haralick texture features [31], principal component analysis (PCA) [32], independent component analysis (ICA) [33], partial least squares decomposition [34], empirical mode decomposition with PCA or ICA [35] or circularity features obtained from DAT [36]. These new types of features seem to give the best accuracy with the SVM classifier. Furthermore, the image voxels within the ROI are also used directly as the input features with SVM [37], [38], logistic lasso [39], and single-layer neural network [40] classifiers.

In this tutorial, we utilize the most commonly used SBR feature with an SVM classifier as an example of the traditional classification method. The SBR [25] can be 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, 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, and SBR is calculated as

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

The SBR of each subject can be obtained from the PPMI database alongside the SPECT images. It was proved that applying SBR to SVM gives very high accuracy [28]; therefore, we will use this classification method as a baseline for comparing and evaluating with the deep learning approach.

C. CNN Architectures for PD recognition

There are several DCNN based models for PD recognition using SPECT images. From 2017, Martinez-Murcia *et al.* proposed utilization of DCNN on SPECT image to diagnose PD [41]. They trained their model with 301 SPECT images (158 PD, 111 normal control (NC), and 32 scans without evidence for dopaminergic deficit (SWEDD)) from the PPMI database, and their network achieved up to 95.5% accuracy (96.2% sensitivity). Choi *et al.* proposed a deep DCNN model "PD Net" which was trained with the whole volume of SPECT images to discriminated the PD subjects from healthy subjects [23]. The model was trained with 624 subjects (431 PD and 193 normal control (NC)) from the PPMI database, resulting in 96.0% accuracy (94.2% sensitivity) comparable to the evaluation from the experts.

Later in 2018, Wenzel *et al.* proposed a large DCNN model with 2,872,642 parameters trained by 645 subjects from PPMI (438 PD and 207 NC). Despite the fact that this model yield 97.7% accuracy (96.6% sensitivity), slightly better than PD Net, the model is large and resource-consuming. Recently in 2021, Mohammed *et al.* proposed the present state-of-the-art model with minimal architecture [22]. Their model consisted of three convolutional layers with a filter size of (3×3) and two dense layers. Their model's input image was normalized to enhance the ROI and provide a distinguishing feature to the model. A 10-fold cross-validation was used to evaluate the performance of the model. This state-of-the-art model was trained by 2723 SPECT images from the PPMI database (1359 PD and 1364 NC) and can provide 99.3% accuracy (99.0% sensitivity).

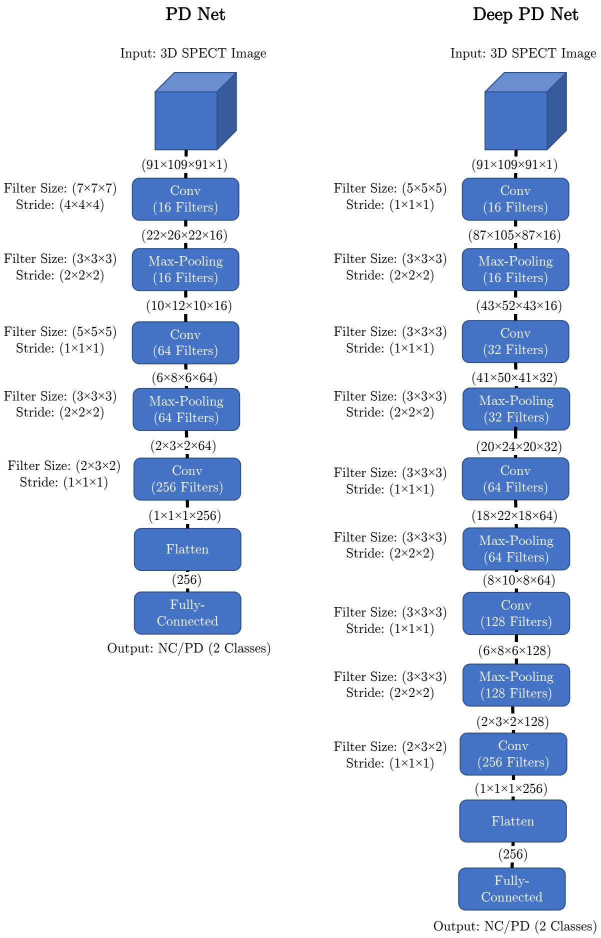


Fig. 1. Structure of PD-Net and Deep PD Net used as examples in this tutorial with the details of the size and number of convolution and max-pooling filters. The PD Net has been modified in the last convolution layer so that the image from the database can be used directly without the need for zero-padding.

D. Models Comparison: Example Scenario

For a demonstration, we incorporate four different DCNN architectures based on PD Net [23] for comparing in both classification and interpretation performance. As a tutorial, we choose these four DCNN architectures so that the classification performance is not significantly different and difficult to evaluate. Later in the tutorial, we will show another benefit of model interpretation: to interpret feedback as an evaluation metric. For further study and development, we suggest utilizing the state-of-the-art model [22].

The first model in this tutorial is the PD Net illustrated

TABLE I
CLINICAL DETAILS OF ALL SUBJECTS USED IN THIS TUTORIAL.

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

on the left-hand side of Figure 1. In the original PD Net, zero-padding was applied to make the image's size equal in all dimensions. However, this tutorial does not include zero-padding so that the images are all in their original form. Thus, a slight modification of the filter size is made in our PD Net 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 a 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, 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 nodes to discriminate the extracted features.

The second model is a modified PD Net architecture by increasing the network depth as shown on the right-hand side of Figure 1. We refer this model as "Deep PD Net". In this model, the filter size of both the 3D convolution and max-pooling layers was designed so that the last layer before the fully-connect layer gives 256 features, the same as PD Net.

The third and fourth models are PD Net and Deep PD Net with batch normalization. Batch normalization was proposed to accelerate DCNN's training and was first applied with the image classification task [42]. 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. The training parameters for all DCNN models are elaborated in Appendix.

All of the models were trained by the subjects with clinical characteristics summarized in Table I. 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 uses the validation set to tune the model to reach 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 it to the testing set.

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

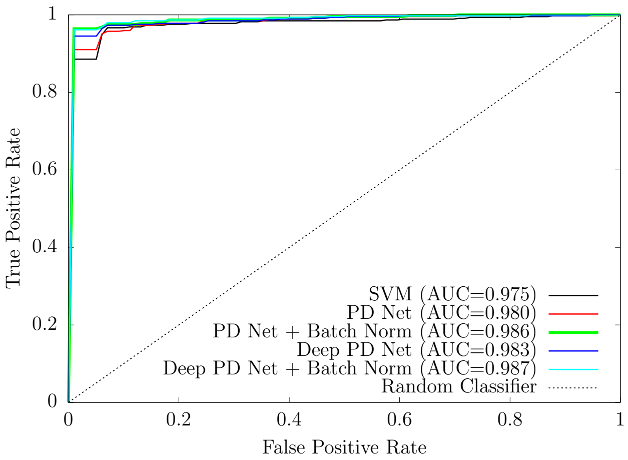


Fig. 2. ROC curve for each model.

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{Total positive}}, \tag{2}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{Total negative}}. \tag{3}$$

Results that were acquired 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 DCNN as the classifier. Four types of DCNN architecture were designed based on the PD Net [23] and all of them are described in the previous section. The mean ± STD of accuracy, sensitivity, and specificity calculated from 10-fold of a testing set, are shown in Table II. The accuracy varies from 95% to 96% with the deep learning approaches, giving a slightly higher accuracy than the SVM model. Deep PD Net with batch normalization has the highest accuracy with 96.87%. The sensitivity of each model was not significantly different. However, we can see the improvement of the specificity from 93% to 97% of the Deep PD Net model.

McNemar's test [43] was used to compare between SVM and DCNN 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 shown in Figure 2. It reveals a trend of a higher AUC value of DCNN than that of SVM. The Deep PD Net with batch normalization has the highest AUC value, which is 0.987.

III. MODEL INTERPRETATION METHODS

A. Interpretation Methods Overview

Despite the fact that DCNN models can provide highly accurate classification results, due to DCNN's black-box nature, it is difficult to directly explain the importance of the input features that lead to high classification performance. Model interpretation methods have been used to reveal the feature importance and assess the trust of the model prediction results. Hence, the primary purpose of the interpretation method is to calculate the "contribution score" [44] of the input features. Vastly used model interpretation methods for

TABLE II
CLASSIFICATION PERFORMANCE OF SVM, PD NET AND DEEP PD NET.

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

DCNN can be categorized into two major groups. The first one is the gradient-based method, which focuses on using backpropagation to calculate the gradient that can be implied back to be the input score of the target class's input features. 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.

1) Gradient based method: The core concept of deep learning is to calculate the gradient of the loss function with respect to all the model's weights and biases. 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 [45]. 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

$$L^c_{\text{Saliency map}} = \frac{\partial S^c}{\partial x} \tag{4}$$

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 the 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 [46]. In other words, this method includes the guidance signal to the deeper layer during the backpropagation resulted 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 DCNN. GAP reduces the total model parameters and results in preventing 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 the 2D elements i, j or can be written as

$$G^k = \sum_i \sum_j A^k_{ij}. \tag{5}$$

The score of predict class c then becomes

$$S^c = \sum_k \sum_i \sum_j w^c_k A^k_{ij}, \tag{6}$$

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

$$\text{CAM} = \sum_k w^c_k A^k_{ij}, \tag{7}$$

which shows the 2D map of the score that predict class c. CAM represents the input feature's contribution score by resizing this 2D map to the original input image. It also has a remarkable ability for object localization of the predict class [47]. 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 [48]. Grad-CAM directly calculates the gradient using the backpropagation from each neuron of the last convolution layer feature map, which can be written as $\partial S^c / \partial A^k_{ij}$. Then, these gradients are summed within the k^{th} feature map to generate the weight of each map and predict class c, which can be written as;

$$\alpha^c_k = \sum_i \sum_j \frac{\partial S^c}{\partial A^k_{ij}} \tag{8}$$

Then Grad-CAM of class c can be generated from

$$L^c_{\text{Grad}-\text{CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right). \tag{9}$$

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 backpropagation, Grad-CAM can be applied to interpret any type of DCNN (e.g., DCNN with recurrent neural networks) without any modifications to the DCNN model.

Guided Grad-CAM: The use of the last convolution layer of the Grad-CAM can provide a more accurate location of the relevant image regions. However, this last layer does not maintain enough resolution to provide a 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) 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 original model's approximation. 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, representing the presence or absence of the input features. For the image classification task, these input features can be pixels or super-pixels. This method of generating the explanation model is defined as the additive feature attribution method [49], [50], in which the explanation model g is written as

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i {x'}_i, \tag{10}$$

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 Equation 10 are discussed below.

DeepLIFT: Deep Learning Important FeaTures (DeepLIFT) is an interpretation method that avoids discontinuity of the gradient-based approach in approximating the feature contribution to the output [44]. 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;

$$\sum_{i=1}^{M} C_{\Delta x_i \Delta y} = \Delta y. \tag{11}$$

By comparing this with Equation 10 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" [51].

SHAP: SHapley Additive exPlanation (SHAP) was designed to simplify any complex model, not restricted to any model structure [49]. 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 Equation 10. SHAP proposes a way to approximate the Shapley value by minimizing the objective function that satisfies all the properties of Equation 10. This objective function does not constrain any model parameters and only use the result from the model output. Thus, SHAP becomes "model-agnostic" in the approximation of the contribution score.

B. Interpretation performance Comparison: Example Scenario

Since there are various model interpretation methods available, it is not possible to decisively express which method is the best for SPECT image classification without actual comparison. Therefore, we demonstrate the interpretation performance of the six well-known interpretation methods mentioned above in this tutorial.

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 [29]. 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 ± SD of the thresholds for healthy subjects and PD subjects, which the experts selected, were reported in [29] as 0.63 ± 0.04 and 0.69 ± 0.05, respectively. In this tutorial, 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 the segmented area. The segmented area is now used as the ground truth to evaluate the interpretation performance.

The slice averaging the attention map from the interpretation method was also generated similar to the slice averaging the SPECT images. Examples of grayscale attention maps 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 located near or inside the segmented region show the most contributed area in the class prediction.

C. Evaluation Methods for Interpretable Models

The pixels that are used to evaluate the interpretation performance need to be selected with another threshold. [44] 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 the 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

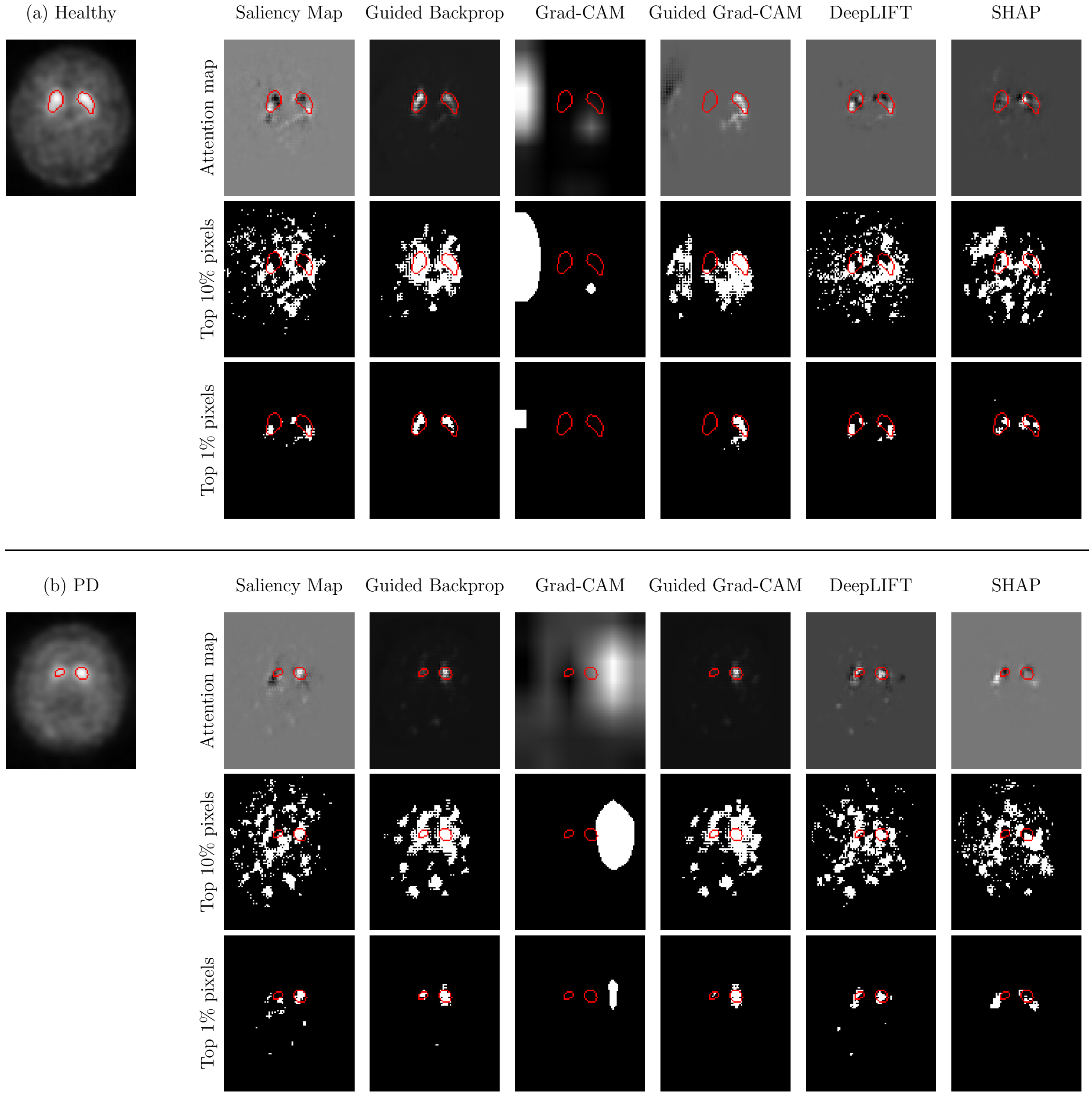


Fig. 3. An example of slice averaging SPECT image (left figure) and the attention map (right table) from Deep PD Net model for (a) healthy control and (b) PD. The red line is the segmented line generated from the mean threshold reported in Ref. 29. The first row of the right table shows the original 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.

TABLE III

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 ± 4.91	**23.78 ± 6.02**	3.27 ± 8.11	22.15 ± 7.36	16.92 ± 6.75	16.62 ± 6.77
PD Net + Batch Norm	12.51 ± 4.99	**23.90 ± 6.76**	4.49 ± 7.89	20.73 ± 8.53	14.98 ± 6.04	15.50 ± 6.85
Deep PD Net	17.29 ± 5.00	**29.72 ± 8.95**	3.66 ± 7.77	25.70 ± 10.15	18.11 ± 6.59	15.72 ± 9.60
Deep PD Net + Batch Norm	15.22 ± 4.36	**29.38 ± 9.00**	2.96 ± 6.49	21.11 ± 12.16	16.99 ± 5.23	16.35 ± 9.39

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

show that the majority of pixels gather inside the segmented red line area.

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

$$D = \frac{2\,|P \cap G|}{|P| + |G|}. \tag{12}$$

The coefficient exists in the range of $[0, 1]$ where $D = 1$ indicates identical segmentation. The mean ± SD of the Dice coefficient is calculated from the test set of all 10-fold. The results are shown in Table III. The bold value indicates the best result in a given threshold. The upper and lower tables show the results from the 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 Dice coefficient's boxplots 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. This test was then used 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.

Mean absolute error is used as another measure to evaluate each method's performance as demonstrated in Figure 5. The guided backpropagation shows that the error approaches zero inside the striatal nuclei, which can be interpreted as the Deep PD Net directly focuses on the region and gives more credibility to 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.

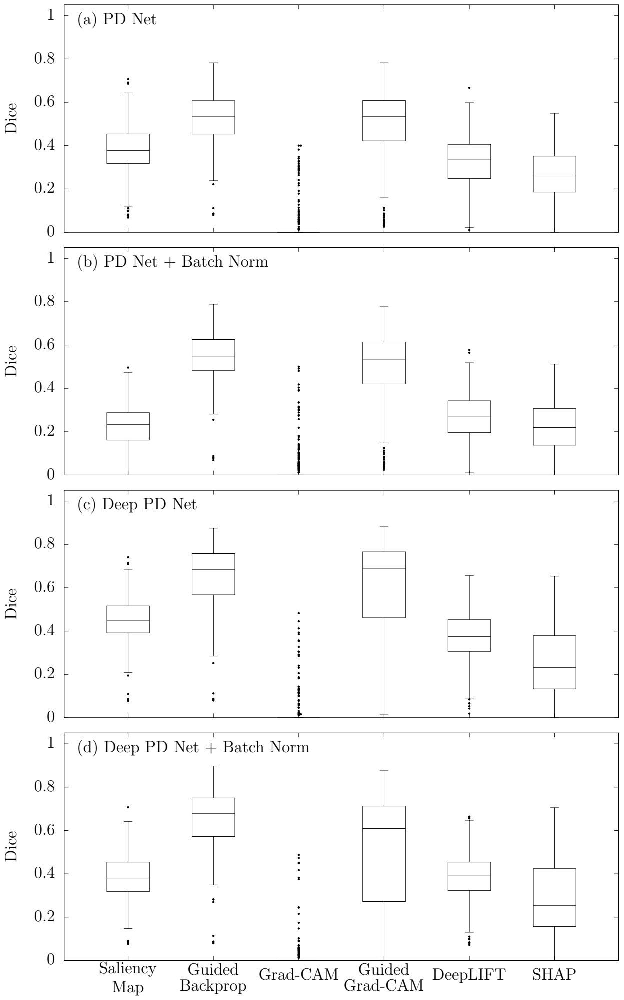


Fig. 4. Boxplots of Dice coefficient in different interpretation methods from the top 1% of contribution score for (a) PD Net (b) PD Net + Batch Norm (c) Deep PD Net and (d) Deep PD Net + Batch Norm. Median is the line that locates inside the box, and black dots represent outliers outside 1.5 times the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

IV. Discussion on Example Scenario

In section II, we demonstrate the comparison of classification performance between 4 types of DCNN architecture based on the PD Net. DCNN may cause the overfitting of the data [52]. However, Deep PD Net, which attaches more convolution and max-pooling layers to increase the network depth, yields better performance than PD Net without overfitting. The addition of batch normalization to the model shows a minor improvement of the model accuracy, which might be due to 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 batch normalization study [42]. From the clinical details shown in Table I, the number of PD subjects is 3 times higher than the number of healthy subjects. Due to this extreme class imbalance, we can observe only the increase in the specificity but not the sensitivity.

For the comparison of interpretation performance, the Dice coefficient in Table III and mean absolute error in Figure 5 show that guided backpropagation outperforms other methods. Guided backpropagation was first designed to improve the saliency map's quality in feature visualization of the deep learning model [46]. In this tutorial, it has the best ability to show fine-grained importance. It also gives much less error in the mean absolute error plot compared to other methods. On the other hand, Grad-CAM was the only method that barely focuses on the crucial region. Although Grad-CAM was supposed to perform well in the class-discriminative and localize relevant image regions [48], it barely focuses on fine-grained importance. Another issue of Grad-CAM is that it heavily relies on the resolution of the last feature map. Since PD Net was designed with the last feature map of size $(1 \times 1 \times 1)$, we need to select the feature map from the upper convolution layer with size $(6 \times 8 \times 6)$ which may reduce the

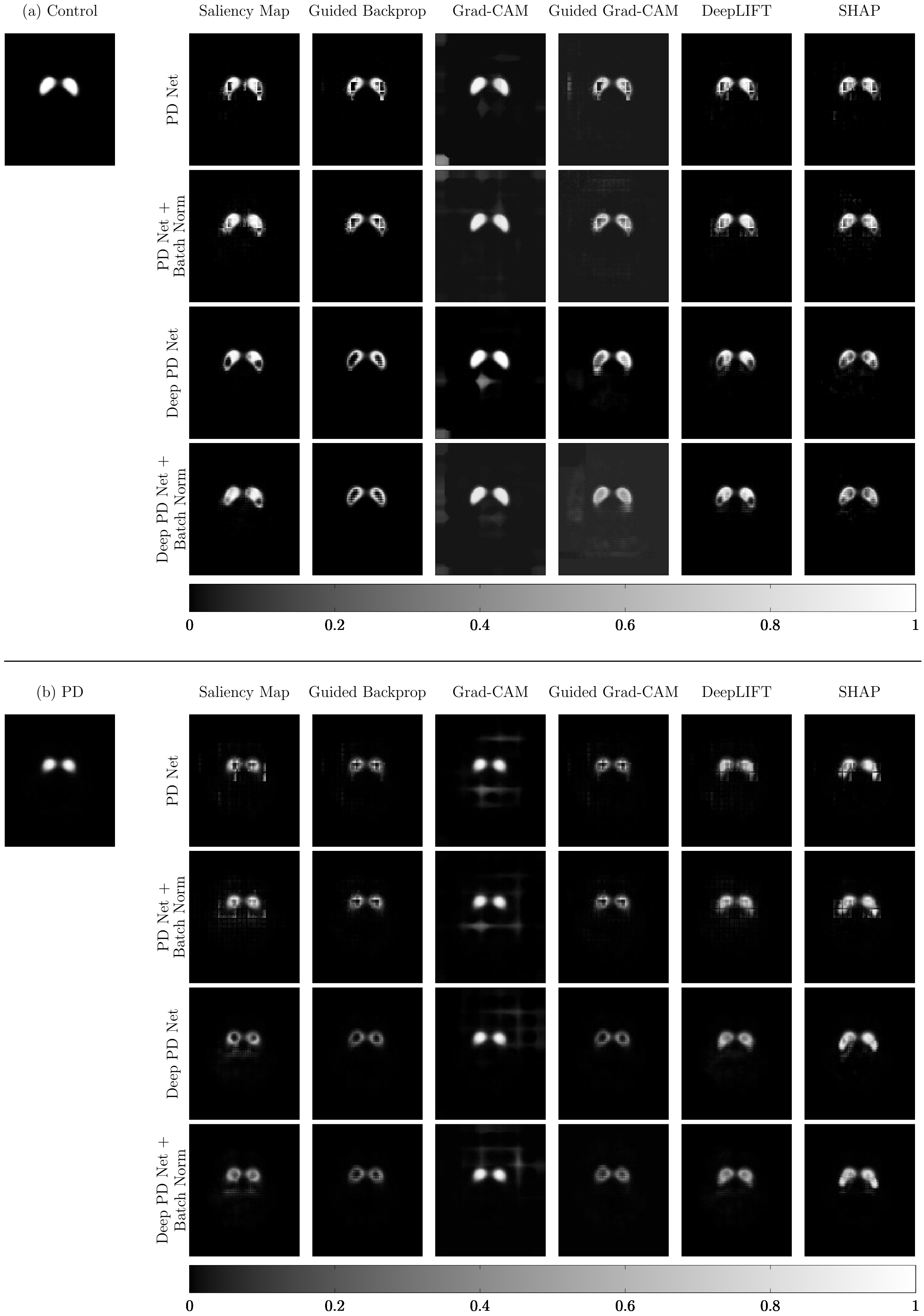


Fig. 5. The mean segmented image (left) and mean absolute error plot (right table) for (a) healthy control and (b) PD group. The mean absolute error was calculated using the binary image from the top 1% contribution pixels to compare with the binary image from the segmented image.

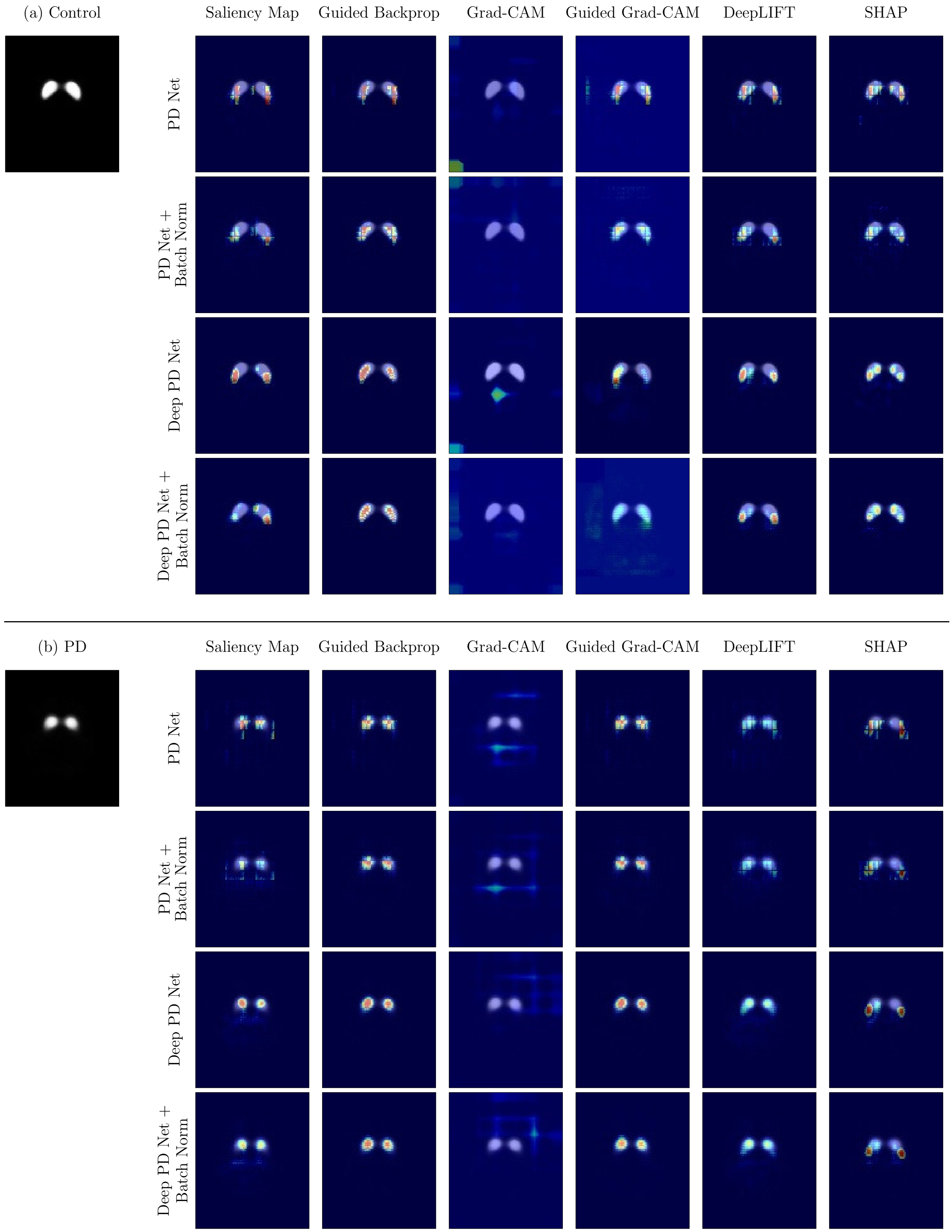


Fig. 6. The mean segmented image (left) and mean segmented heatmap (right table) for (a) healthy control and (b) PD group. The mean segmented heatmap was generated using the mean of binary images from the top 1% contribution pixels.

interpretation performance.

When we plot the mean segmented image from the binary map of the top 1% pixels, saliency map, DeepLIFT, and SHAP also able to discriminate the difference between classes. Since SHAP is the model-agnostic method, it does not rely on DCNN weights; thus, the result becomes different from the saliency map and DeepLIFT. Even though these three methods' performance is hard to evaluate, SHAP can generate a better quality heatmap at the uptake depletion location, which outperforms other methods in discriminating the difference between PD and healthy subjects. This should be consistent with [49], 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. In conclusion, both guided backpropagation and SHAP are suitable interpretation methods for the architectures in this tutorial. Nevertheless, in other medical image applications, the other interpretation methods might overcome both methods.

Another interesting aspect of the model interpretation is to use the interpreted feedback as an evaluation metric to choose the best model. For example, from Table II and Figure 2, PD Net with batch normalization shows better specificity and AUC value comparing to Deep PD Net, while Deep PD Net shows better accuracy and sensitivity. The performance of the two models is not significantly different and is difficult to evaluate. To this end, the feedback from the interpretation method can provide a decisive answer to the evaluation. From Table III, when utilizing the guided backpropagation method, Deep PD Net has the highest interpretation performance, which is significantly higher than that of PD Net with batch normalization. This result suggests that Deep PD Net is the best suit for PD recognition among the tutorial's DCNN architecture.

We suggest a flow chart for applying the interpreted feedback to assist in model evaluation in Figure 7. In this tutorial, if we follow the flow chart, the Deep PD Net model should be suggested to be used for PPMI data. This flow chart and the concept of model interpretation methods can be utilized not only in medical image application but also in other DCNN application where the credibility of DCNN need to be verified as well, for example, DCNN applications in biomedical [14], [15] or bioinformatics [53].

V. Conclusions

The purpose of this tutorial is to demonstrate the procedure for selecting the most reliable interpretation method for SPECT image PD recognition model. To this end, we introduce the traditional and DCNN approach for PD diagnosis and give an example scenario with four DCNN models. Then, we introduce six well-known interpretation methods and exhibit these six methods' interpretation performance on those four DCNN models mentioned above. We propose evaluation methods for measuring the interpretation performance using the Dice coefficient and mean segmented binary image overlay on top of ground truth segmented image. Finally, we discuss about utilizing interpreted feedback for deciding the most suitable model for the intended task. The interpretation and evaluation methods displayed in this tutorial can be applied for other tasks aside from PD recognition, and contribute to sensor data processing in an AI Era.

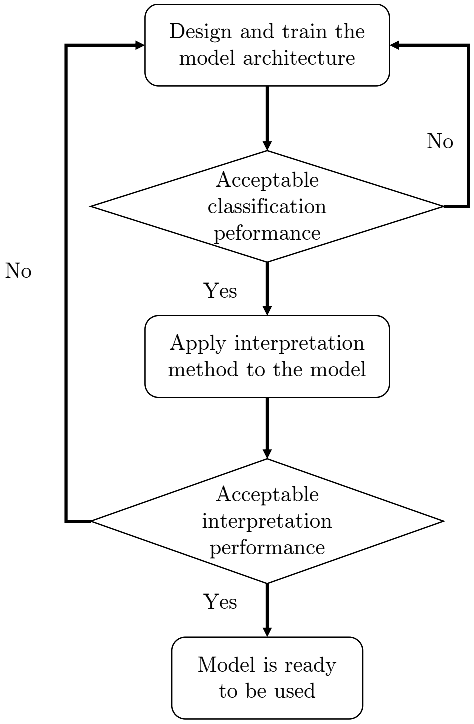

Design and train the model architecture
Acceptable classification peformance
No
Yes
Apply interpretation method to the model
Acceptable interpretation performance
No
Yes
Model is ready to be used

Fig. 7. The flow chart for the interpretation method application for assisting in model evaluation.

Appendix

All the DCNN models were implemented with Keras [54], an open-source deep learning library written in Python and running on top of Tensorflow [55]. 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 [56]. The loss function also is weighted for class imbalance during the training. These training parameters are the same with [23] and every model uses the same parameters for a fair comparison.

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, and G. Halliday, "Missing pieces in the Parkinson's disease puzzle," *Nature Medicine*, vol. 16, pp. 653–661, 2010.

[2] K. R. Chaudhuri and A. H. Schapira, "Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment," *The Lancet Neurology*, vol. 8, no. 5, pp. 464 – 474, 2009.

[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, and M. Wanner, "SNM practice guideline for dopamine transporter imaging with 123I-Ioflupane SPECT 1.0," *Journal of Nuclear Medicine*, vol. 53, no. 1, pp. 154–163, 2012.

[4] K. Badiavas, E. Molyvda, I. Iakovou, M. Tsolaki, K. Psarrakos, and N. Karatzas, "SPECT imaging evaluation in movement disorders: far beyond visual assessment," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 38, no. 4, pp. 764–773, 2011.

[5] J. C. Taylor and 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*, vol. 4, no. 1, p. 29, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016.

[8] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[9] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLOS Computational Biology*, vol. 10, no. 12, pp. 1–18, 12 2014.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[11] J. S. Duncan, M. F. Insana, and N. Ayache, "Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue]," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 3–10, Jan 2020.

[12] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, and A. Ortiz, "Convolutional neural networks for neuroimaging in Parkinson's disease: Is preprocessing needed?" *International Journal of Neural Systems*, vol. 28, no. 10, p. 1850035, 2018.

[13] G. Ras, M. van Gerven, and P. Haselager, *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*. Cham: Springer International Publishing, 2018, pp. 19–36.

[14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Jan 2017.

[15] 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, and S. Do, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomedical Engineering*, vol. 3, no. 3, pp. 173–182, 2019.

[16] J. Grezmak, J. Zhang, P. Wang, K. A. Loparo, and R. X. Gao, "Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3172–3181, 2020.

[17] A. S. Alharthi, A. J. Casson, and K. B. Ozanyan, "Gait spatiotemporal signal analysis for parkinson's disease detection and severity rating," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1838–1848, 2021.

[18] S. Lee, J. Kim, S. W. Park, S. M. Jin, and S. M. Park, "Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 536–546, 2021.

[19] I. Klyuzhin, N. Shenkov, A. Rahmim, and V. Sossi, "Use of deep convolutional neural networks to predict parkinson's disease progression from datscan spect images," *Journal of Nuclear Medicine*, vol. 59, no. supplement 1, pp. 29–29, 2018. [Online]. Available: https://jnm.snmjournals.org/content/59/supplement_1/29

[20] A. Ortiz, J. Munilla, M. Martínez-Ibañez, J. M. Górriz, J. Ramírez, and D. Salas-Gonzalez, "Parkinson's disease detection using isosurfaces-based features and convolutional neural networks," *Frontiers in Neuroinformatics*, vol. 13, p. 48, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2019.00048

[21] M. Wenzel, F. Milletari, J. Krüger, C. Lange, M. Schenk, I. Apostolova, S. Klutmann, M. Ehrenburg, and R. Buchert, "Automatic classification of dopamine transporter spect: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics," *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 13, pp. 2800–2811, 2019.

[22] F. Mohammed, X. He, and Y. Lin, "An easy-to-use deep-learning model for highly accurate diagnosis of parkinson's disease using spect images," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101810, 2021.

[23] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, "Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging," *NeuroImage: Clinical*, vol. 16, pp. 586 – 594, 2017.

[24] K. Marek *et al.*, "The parkinson progression marker initiative (PPMI)," *Progress in Neurobiology*, vol. 95, no. 4, pp. 629 – 635, 2011.

[25] G. Wisniewski, J. Seibyl, and K. Marek, "DatScan SPECT image processing methods for calculation of striatal binding ratio (SBR)," Institute for Neurodegenerative Disorders (IND), Tech. Rep., 2013.

[26] B. Palumbo, M. L. Fravolini, S. Nuvoli, A. Spanu, K. S. Paulus, O. Schillaci, and 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*, vol. 37, no. 11, pp. 2146–2153, Nov 2010.

[27] B. Palumbo, M. L. Fravolini, T. Buresta, F. Pompili, N. Forini, P. Nigro, P. Calabresi, and 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*, vol. 93, no. 27, p. e228, Dec 2014.

[28] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3333 – 3342, 2014.

[29] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "High-accuracy classification of Parkinson's disease through shape analysis and surface fitting in 123i-ioflupane SPECT imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 794–802, May 2017.

[30] A. Augimeri, A. Cherubini, G. L. Cascini, D. Galea, M. E. Caligiuri, G. Barbagallo, G. Arabia, and A. Quattrone, "CADA—computer-aided DaTSCAN analysis," *EJNMMI Physics*, vol. 3, no. 1, p. 4, Feb 2016.

[31] F. J. Martínez-Murcia, J. M. Górriz, J. Ramírez, I. A. Illán, and C. G. Puntonet, "Texture features based detection of Parkinson's disease on datscan images," in *Natural and Artificial Computation in Engineering and Medical Applications*, J. M. Ferrández Vicente, J. R. Álvarez Sánchez, F. de la Paz López, and F. J. Toledo Moreo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 266–277.

[32] D. J. Towey, P. G. Bain, and K. S. Nijran, "Automatic classification of I-123-FP-CIT (DaTSCAN) SPECT images," *Nucl Med Commun*, vol. 32, 2011.

[33] F. Martínez-Murcia, J. Górriz, J. Ramírez, I. Illán, and A. Ortiz, "Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging," *Neurocomputing*, vol. 126, pp. 58 – 70, 2014.

[34] F. Segovia, J. M. Górriz, J. Ramírez, I. Álvarez, J. M. Jiménez-Hoyuela, and S. J. Ortega, "Improved parkinsonism diagnosis using a partial least squares based approach," *Medical Physics*, vol. 39, no. 7, pp. 4395–4403, 2012.

[35] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. G. Río, and M. Moreno-Caballero, "Application of empirical mode decomposition (emd) on datscan SPECT images to explore parkinson disease," *Expert Systems with Applications*, vol. 40, no. 7, pp. 2756 – 2766, 2013.

[36] T. Shiiba, Y. Arimura, M. Nagano, T. Takahashi, and A. Takaki, "Improvement of classification performance of Parkinson's disease using shape features for machine learning on dopamine transporter single photon emission computed tomography," *PLOS ONE*, vol. 15, no. 1, pp. 1–12, 01 2020.

[37] I. A. Illán, J. M. Górriz, J. Ramírez, F. Segovia, J. M. Jiménez-Hoyuela, and S. J. Ortega Lozano, "Automatic assistance to Parkinson's disease diagnosis in datscan SPECT imaging," *Medical Physics*, vol. 39, no. 10, pp. 5971–5980, 2012.

[38] F. P. M. Oliveira and M. Castelo-Branco, "Computer-aided diagnosis of Parkinson's 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*, vol. 12, no. 2, p. 026008, 2015.

[39] H. D. Tagare, C. DeLorenzo, S. Chelikani, L. Saperstein, and R. K. Fulbright, "Voxel-based logistic analysis of PPMI control and Parkinson's disease datscans," *NeuroImage*, vol. 152, pp. 299 – 311, 2017.

[40] Y. C. Zhang and A. C. Kagen, "Machine learning interface for medical image analysis," *Journal of Digital Imaging*, vol. 30, no. 5, pp. 615–621, Oct 2017.

[41] F. J. Martinez-Murcia, A. Ortiz, J. M. Górriz, J. Ramírez, F. Segovia, D. Salas-Gonzalez, D. Castillo-Barnes, and I. A. Illán, "A 3d convolutional neural network approach for the diagnosis of parkinson's disease," in *Natural and Artificial Computation for Biomedicine and Neuroscience*, J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, and H. Adeli, Eds. Cham: Springer International Publishing, 2017, pp. 324–333.

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Preprint in arXiv:1502.03167*, 2015.

[43] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[44] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 3145–3153.

[45] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Preprint in arXiv:1312.6034*, 2013.

[46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *Preprint in arXiv:1412.6806*, 2014.

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2921–2929.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.

[49] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[50] S. Lundberg and S. Lee, "An unexpected unity among methods for interpreting model predictions," *Preprint in arXiv:1611.07478*, 2016.

[51] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 07 2015.

[52] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[53] N.-Q.-K. Le, Q.-T. Ho, and Y.-Y. Ou, "Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins," *Journal of Computational Chemistry*, vol. 38, no. 23, pp. 2000–2006, 2017.

[54] F. Chollet. (2015) Keras. [Online]. Available: https://keras.io/

[55] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *Preprint in arXiv:1603.04467*, 2016.

[56] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256.