A comparative study for interpreting deep learning prediction of the Parkinson’s disease diagnosis from SPECT imaging

Theerasarn Pianpanit, Sermkiat Lolak, Phattarapong Sawangjai, Apiwat Ditthapron, Sanparith Marukatat, Ekapol Chuangsawanich, and Theerawit Wilaiprasitporn, Member, IEEE

Abstract—The application of deep learning to single-photon emission computed tomography (SPECT) imaging in Parkinson’s disease shows effectively high diagnosis accuracy. However, difficulties in model interpretation were occurred due to the complexity of the deep learning model. Although several interpretation methods were created to show the attention map that contains important features of the input data, it is still uncertain whether these methods can be applied in PD diagnosis. Four different models of the deep learning approach based on 3-dimensional convolution neural network (3D-CNN) of well-established architectures have been trained with an accuracy up to 95-96% in classification performance. These four models have been used as a comparative study for well-known interpretation methods. Generally, radiologists interpret SPECT images by confirming the shape of the I123-Ioflupane uptake in the striatal nuclei. To evaluate the interpretation performance, the segmented striatal nuclei of SPECT images are chosen as the ground truth. Results suggest that guided backpropagation and SHAP which were developed recently, provided the best interpretation performance. Guided backpropagation has the best performance to generate the attention map that focuses on the location of striatal nuclei. On the other hand, SHAP surpasses other methods in suggesting the change of the striatal nucleus uptake shape from healthy to PD subjects. Results from both methods confirm that 3D-CNN focuses on the striatal nuclei in the same way as the radiologist, and both methods should be suggested to increase the credibility of the model.

Index Terms—Parkinson’s disease, computer-aided diagnosis, convolution neural network, interpreting deep learning, SPECT visualization

I. 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 investigation 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. Furthermore, due to the “black box” nature of the algorithm, the adoption rate of using the deep CNN in practice is still low. Also, the EUs General Data Protection Regulation (GDPR), Recital 71, which gives citizens a “right to explanation” will make the “black box” approaches difficult to use 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 still a lack of 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 by applying them to models to find reliable interpretation methods to explain the model. By using the most reliable interpretation method, we also compare the interpretation performance among four 3D-CNN models to suggest the model architecture that has the highest credibility. This comparative 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 II. We present the results and information about the performance in section III. Finally, we draw conclusions of this study in section IV.

II. MATERIALS AND METHOD

A. 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 preprocessing procedure before they are publicly shared via the database [31]. SPECT raw projection data was imported to a HERMES 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 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 × 2 × 2 mm³ and the dimension of 91 × 109 × 91 can be directly downloaded from the publicly shared PPMI database.

B. 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

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

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 [8], is selected as the baseline of the conventional machine learning method to compare with the deep learning approach.

C. 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 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 × 3 × 3) pool size and stride of 2 attached. The first 3D convolution layer has 16 filters with a size of (7 × 7 × 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 × 5 × 5) and a stride of 1. Finally, a 3D convolution layer with 256 filters of size (2 × 3 × 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.

D. 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

1Hermes Medical, Stockholm, Sweden
2PMOD Technologies, Zurich, Switzerland
0.9. The learning rate was initially $1 \times 10^{-4}$ and logarithmically decreased to have $1 \times 10^{-6}$ at the final epoch. Additionally, weight parameters in the model were initiated with a Glorot initialization [35].

E. 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.

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

$$L^c_{\text{Saliency map}} = \frac{\partial S^c}{\partial x}$$  (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 can be used 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. 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

$$G^k = \sum_i \sum_j A^k_{ij}.$$  (3)

The score of predict class $c$ then becomes

$$S^c = \sum_k \sum_i \sum_j w^c_k A^k_{ij},$$  (4)

where $w^c_k$ is the weight of $G^k$ of class $c$. By examining this equation, class activation map (CAM) can be defined as

$$\text{CAM} = \sum_k w^c_k A^k_{ij}.$$  (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 $\partial S^c / \partial A^k_{ij}$. 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:

$$\alpha^c_k = \sum_i \sum_j \frac{\partial S^c}{\partial A^k_{ij}}.$$  (6)

Then Grad-CAM of class $c$ can be generated from

$$L^c_{\text{Grad-CAM}} = \text{ReLU} \left( \sum_k \alpha^c_k A^k \right).$$  (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 backpropagation, 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) 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 \( x = h(x') \). 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

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

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 \( \phi_i \) which is the “attribute” or “contribution score” from each input feature. Two well-known interpretation methods which are based on the concept of Equation 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 \( \Delta x_i \) as \( C_{\Delta x_i, \Delta y} \) and uses the summation of these attributions to give the value of \( \Delta y \), which can be written as;

\[
\sum_{i=1}^{M} C_{\Delta x_i, \Delta y} = \Delta y. \tag{9}
\]

By comparing this with Equation 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 Equation 8. SHAP proposes a way to approximate the Shapley value by minimizing the objective function that satisfies all the properties of Equation 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.

F. Experiment

Clinical characteristics of the subjects are 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 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.

III. RESULTS AND DISCUSSION

A. 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

\[
\text{Sensitivity} = \frac{\text{True positive}}{\text{Total positive}}, \tag{10}
\]

\[
\text{Specificity} = \frac{\text{True negative}}{\text{Total negative}}. \tag{11}
\]

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 ± STD of accuracy, sensitivity, and specificity which were 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 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 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 I, the number of PD subject is 3 times higher than the number of healthy subject. Due to this extreme class imbalance, the specificity of each model was not as high as the sensitivity.

---

### Table I

| Clinical details of all subjects used in this study. |
|------------------------------------------------------|
| 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                                            |
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| 97.11 ± 2.93| 91.46 ± 6.22|
| PD Net             | SPECT         | 95.39 ± 2.88| 97.75 ± 2.36| 89.66 ± 7.13|
| PD Net + Batch Norm| SPECT         | 96.54 ± 2.63| 98.44 ± 2.38| 91.96 ± 6.86|
| Deep PD Net        | SPECT         | 96.71 ± 2.32| 98.43 ± 1.51| 92.16 ± 5.66|
| Deep PD Net + Batch Norm | SPECT    | 96.87 ± 2.13| 99.34 ± 1.07| 90.98 ± 6.71|

TABLE III
THE RESULTS OF MEAN DICE COEFFICIENT USING THE BINARY IMAGE OF THE ATTENTION MAP FOR THE TOP MOST 10% OF CONTRIBUTION SCORES (UPPER) AND TOP MOST 1% OF CONTRIBUTION SCORES (LOWER). THE BOLD NUMBER REFER TO THE HIGHEST DICE COEFFICIENT AMONG ALL THE METHOD.

| 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 | 22.13 ± 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 |

Fig. 2. 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 third row show the binary map generated using only top most 10% of contribution score and using only top most 1% contribution score.
B. 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 ± 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 2. 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 2 (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 2 (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

\[
D = \frac{2 |P \cap G|}{|P| + |G|}. \tag{12}
\]

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 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. Grad-CAM is 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.

The boxplots of the Dice coefficient in Figure 3 also confirm that guided backpropagation performance dominates other methods. Mean absolute error is used as another measure to evaluate the performance between each method as can be seen in Figure 4 and 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.

By examining Figure 5, SHAP is the only method that shows high mean absolute error that locates outside the ground truth segmented region of PD subject. The mean absolute error plot of guided backpropagation and SHAP are compared with the ground truth segmented image as shown in Figure 6. Two red dots in the figure mark the locations where the uptake depletion can be identified and it can be used to distinguish between healthy and PD subject. SHAP mean absolute error peaks around that locations and results in the mean absolute error plot that looks almost like the healthy subject. This confirms that, SHAP outperforms other methods in discriminating the difference between PD and healthy subjects. This study is also consistent with previous study [41], which revealed that SHAP gives the best performance among all other methods of

![Fig. 3. Boxplots of Dice coefficient in different interpretation method using top most 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.](image-url)
The mean segmented image (left) and mean absolute error plot (right) for healthy control group. The mean absolute error was calculated using the binary image from top most 1% contribution pixels to compare with the binary image from the segmented image.

**Fig. 4.** The mean segmented image (left) and mean absolute error plot (right) for healthy control group. The mean absolute error was calculated using the binary image from top most 1% contribution pixels to compare with the binary image from the segmented image.

Fig. 5. Same as Figure 4 but for PD group.
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 the classification performance.

IV. 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 absolute error plot between ground truth segmented images and attention maps, a significant result from SHAP in discriminating the different features between healthy and PD subject was obtained. SHAP correctly shows the uptake depletion regions of PD subjects which is the main characteristic of Parkinson's disease. Furthermore, when using the results of 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 and SHAP can assist in the modification of the model to increase the credibility on PD diagnosis.

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. Boooj, 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. Psarrako s, 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 (123)f-p-cit classification: the beginning of the end for semi-quantification?” EJNMMI Physics, vol. 4, no. 1, p. 29, 2017.
[6] 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-f-p-cit brain SPECT,” European Journal of Nuclear Medicine and Molecular Imaging, vol. 37, no. 11, pp. 2146–2153, Nov 2010.
B. Palumbo, M. L. Fravolini, T. Buresta, F. Pompili, N. Forini, P. Negro, P. Calabresi, and N. Tambascio, “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.

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.

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-toluapate spect imaging,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 1, pp. 276–284, Feb 2018.

F. Martínez-Murcia, J. M. Górriz, J. Ramírez, I. Á. Illán, and A. Ortiz, “Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging,” Neurocomputing, vol. 126, pp. 58 – 70, 2014.

F. Segovia, J. M. Griz, J. Ramírez, I. Iavre, 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.

A. Rojas, J. Griz, J. Ramírez, I. Illn, F. Martínez-Murcia, A. Ortiz, M. G. Ro, and M. Moreno-Caballero, “Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease,” Expert Systems with Applications, no. 4, pp. 2756 – 2763, 2015.

I. A. Blan, J. M. Górriz, J. Ramírez, F. Segovia, J. M. Jiménez-Hoyuela, and S. J. Ortega Lozano, “Automatic assistance to parkinson’s disease’s diagnosis in datscan spect imaging,” Medical Physics, vol. 39, no. 10, pp. 5971–5980, 2012.

F. P. M. Oliveira and 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, vol. 12, no. 2, p. 026008, 2015.

H. D. Tagare, C. Delorenzo, S. Chelikani, L. Saperstein, and R. K. Fulbright, “Voxel-based logistic analysis of ppni control and parkinson’s disease’s datscans,” Neuroimage, vol. 152, pp. 299 – 311, 2017.

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.

Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, p. 436444, 2015.

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

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.

C. F. Cadiou, H. Hong, D. K. L. 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.

G. Litjens, T. Kooi, B. E. Bejnordi, 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.

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 craniofacial imaging,” NeuroImage: Clinical, vol. 16, pp. 586 – 594, 2017.

F. J. Martínez-Murcia, J. M. Griz, 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.

G. Raou, M. van Greven, and P. Haselager, Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. Cham: Springer International Publishing, 2018, pp. 19–36.