End-to-End Parkinson Disease Diagnosis using Brain MR-Images by 3D-CNN

Preprint · June 2018

CITATIONS
0

READS
149

3 authors, including:

Soheil Esmaeilzadeh
Stanford University
5 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Ehsan Adeli
Stanford University
73 PUBLICATIONS **490** CITATIONS
SEE PROFILE

End-to-End Parkinson Disease Diagnosis using Brain MR-Images by 3D-CNN

Soheil Esmaeilzadeh
Stanford University
soes@stanford.edu

Yao Yang
Stanford University
yangyao@stanford.edu

Ehsan Adeli
Stanford University
eadeli@stanford.edu

Abstract

In this work, we use a deep learning framework for simultaneous classification and regression of Parkinson disease diagnosis based on MR-Images and personal information (i.e. age, gender). We intend to facilitate and increase the confidence in Parkinson disease diagnosis through our deep learning framework.

1. Introduction

Parkinson's disease (PD) is a long-term degenerative disorder of the central nervous system that affects the motor system [1]. Symptoms of Parkinson's disease include shaking, rigidity, slowness of movement, difficulty with walking, thinking and behavioral problems, dementia, and etc. In 2015, PD affected 6.2 million people and resulted in about 117,400 deaths globally [2]. The disease typically occurs in people who are over 60 years old, of which one percent are affected. The cause of Parkinson disease is not yet known, but mostly is believed to be due to genetic and environmental reasons [3].

2. Background and Related Works

Computer aided diagnosis is getting common in healthcare recently [4] [5]. The diagnosis process for PD includes considering medical history and neurological examination. Computed tomography (CT) scans of people with PD usually appear normal and MRI has become more accurate in diagnosis of PD, especially through iron-sensitive T2 and SWI sequences at a magnetic field strength of at least 3T, both of which can demonstrate absence of the characteristic *swallow tail* imaging pattern in the dorsolateral substantia nigra. There's a 98% sensitivity and 95% specificity for PD on the absence of the pattern. Brain MR-Images (MRI) are theeqrefore widely used for diagnosing Parkinson in order to improve patient treatment strategies. MR-Images are also used to rule out other diseases that can be secondary causes of parkinsonism, most commonly encephalitis, and chronic ischemic insults. Diagnosing diseases based on radiologists' reading on MRI images are oftentimes prone to mistakes. In recent years, machine learning methods become a common tool for early-stage diagnosis to localize the disease in the brain (i.e., localization of disease markers). Salvatorec et al. used a machine learning algorithm that allows individual differential diagnosis of PD by means of MRI which is able to obtain voxel-based morphological biomarkers of PD [4]. Zhang et al. used a machine learning framework based on principal components analysis (PCA) and Support Vector Machine (SVM) for the classification of Parkinson's disease and Essential Tremor (ET). They used statistical analysis and machine learning method to test the differences between PD and ET in some specific brain regions [6]. Another school of research focuses on Region of Interest methods (ROI) where some specific regions of the brain such as the gray matter, hippocampal volume, and cortical thickness are extracted due to a priori knowledge about their effects on brain functionality and memory [7] [8].

2.1. Baseline

For diagnosis of Parkinson disease using MR-Images the state-of-the-art works done by Ahmed et al. [9] and Gil et al. [10] serve as the baselines for our model accuracy. The best performance reported by Ahmed et al. [9] using an ANN model is 70 percent, and by Gil et al. [10] combining the ANN and SVM classifier has an accuracy level of 86.96 percent where both of them use human-engineered feature extraction for training the models.

2.2. Motivation

As mentioned before, most of the current methods focus on using human-engineered features extracted from MRI data, which have less optimal learning performances be-

Sex	Group	Age							
		count	mean	std	min	25%	50 %	75 %	max
F	HC	70	59.2	11.6	31	53	60	68	84
F	PD	160	61.9	9.9	35	56	62	69	84
M	HC	134	61.7	10.9	31	57	63	69	83
M	PD	292	63.3	9.8	36	57	64	71	89

Table 1: Statistical overview of the MR-Images data; HC: Healthy condition, PD: Parkinson disease

cause of the possible high correlation between engineered features and the subsequent classification or regression models. We believe that integrating the feature extraction and the learning of models into one framework can improve the diagnostic performance. Moreover, by extracting the brain's heat-map after the training process, we will be able to find the important parts of the brain that contribute to the Parkinson disease, and this finding can serve as a valuable information for Parkinson diagnosis by medical practitioners. Hence, with an end-to-end approach and without using apriori human-engineered feature extraction techniques we use MR-Images data in order to classify them as Parkinson Disease (PD) or Healthy Condition (HC) and we use the best accuracy found in the previous works, 86.96 percent, as the baseline for Parkinson disease classification. As the input, we use three-dimensional brain MR-Images (section (3)), and use a Convolutional Neural Network as the training model (section (4)).

3. Dataset and Features

3.1. Format of Data

In this work, a set of three-dimensional brain MR-Images is used for Parkinson Disease diagnosis . The dataset for this work is from the PPMI database [11]. Fig. (2(a)) shows an illustration of an MR-Image cut in three sagittal, coronal, and axial planes respectively with cut coordinates of $x = 36, y = 10, z = 36$ where brain tissues together with skull, scalp, and dura are observed. The size of MR-Image is $(120, 120, 270, 1)$ as MR-Image is gray-scale leading to a size of 4 million pixels for each image that will serve as visual features.

3.2. Statistics of Data

We report the demographic information of subjects in Table (1). The dataset consists of 452 Parkinson patients (PD), including 292 males (M) and 160 females (F) and 204 images from people in healthy conditions (HC) with 134 males and 70 females. The average age of the patients is 61 where the minimum age is around 30 and the maximum age is 89 (cref. Table (1)). Fig. (1(a)) shows the number of patients in different classes of Parkinson disease as a function of age and Fig. (1(b)) shows the age distribution of different classes of Parkinson disease. The average age of the patients is around 61 in each class of disease (cref. Fig. (1(b))) among male/female and we include age and gender as extra features besides MR-Images in training of our model. Besides, both male and female groups have approximately similar portions of patients in each of the Parkinson classes; in spite of this, we consider gender also as a feature in training the model in addition to age. Hence, the MR-Images, age, and gender of patients are the features that are used for training of our deep learning model that is presented in section (4).

3.3. Preprocessing of Data

3.3.1 Skull-Stripping

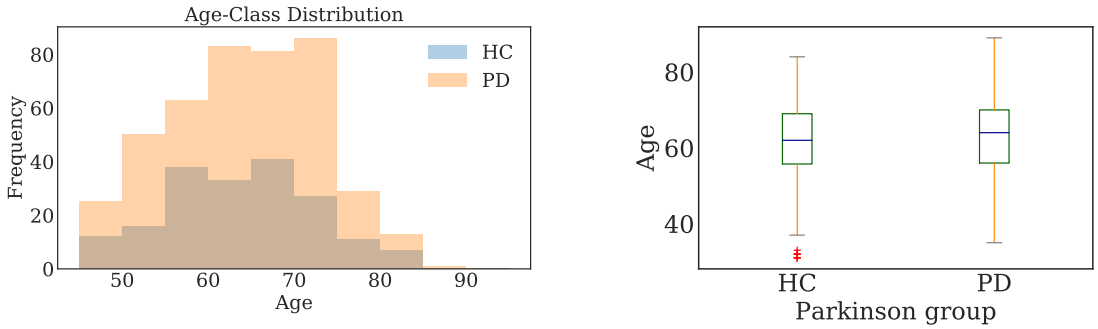
In the preprocessing stage, we carry out skull-stripping to remove non-cerebral tissues like skull, scalp, and dura. In Fig. (2(b)) a skull-stripped version of MR-Image is given. Briefly, skull stripping acts as the preliminary step in numerous medical applications as it increases the speed and accuracy of diagnosis and includes removal of non-cerebral tissues like skull, scalp, and dura from brain images. Skull stripping can be part of the tissue segmentation (e.g. in SPM) but is mostly done by specialized algorithms that delineate the brain boundary. See [12] for a comparison of some brain extraction algorithms (BSE, BET, SPM, and McStrip), which suggests that all algorithms perform well in general but results highly depend on the particular dataset.
In our work we use the Brain Extraction Technique (BET) proposed by Smith in 2002 [13] together with a Statistical Parametric Mapping (SPM) as a voxel-based approach for brain image segmentation and extraction and choose the stripped version with higher brain tissue intensity. Performing skull-stripping on brain MR-Images reduces the size of MR-Images to $(80, 100, 108, 1)$ (i.e. 800,000 pixels), leading to a reduction factor of 4.61.

3.3.2 Data Augmentation

Furthermore, we perform a data-augmentation technique on the MR-Images to expand the training set size. For this purpose, in the skull-stripped MR-Images we flip the right and left hemispheres of a brain for each patient and keep everything else the same. By doing this, we double the size of our dataset and can further train our model on a larger dataset than the original available samples.

4. Machine Learning Model

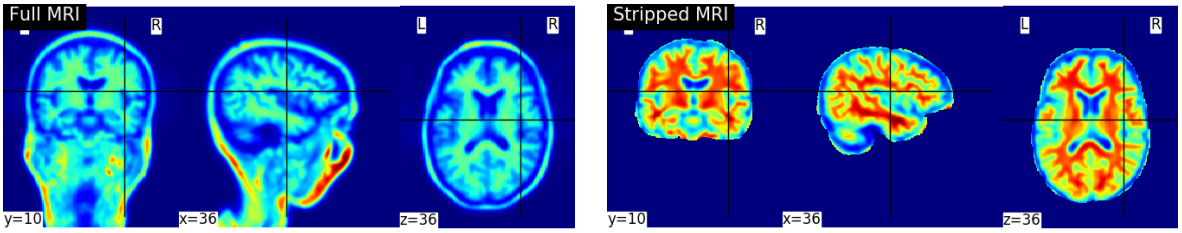
In this work, we divide the dataset into a training (85%), a development (10%) and a test set (5%), and split them into batches of size 8 (due to memory issues we cannot test the effect of different batch sizes on the training process



(a) Age-Frequency distribution

(b) Age-Class distribution

Figure 1: Parkinson disease dataset overview; HC: Healthy condition, PD: Parkinson disease



(a)

(b)

Figure 2: (a) full and (b) skull-stripped brain MR-Image - from *left* to *right*: Coronal, Axial, and Sagittal views

and just use size of 8). During the training, process we keep track of training and development set accuracies and loss function values. F_2-score (Eqs. (1)-(3)) is being used which weighs recall higher than precision (by placing more emphasis on false negatives) to evaluate the performance of our model with true positive, true negative, false positive, and false negative being as TP, TN, FP, and FN respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F_2 = \frac{5 \text{ precision} \times \text{recall}}{4 \text{ precision} + \text{recall}} \tag{3}$$

For training process, we build a 3D Convolutional Neural Network (3D-CNN) shown in Fig. (3) and integrate the feature extraction and the learning of the model into a unified framework. In the 3D-CNN shown in Fig. (3), L_0 is the input layer, which is the 3D MR-Image of brains with input dimension of $(80, 100, 108)$. $L_{1,2,4,5,7,8}$ layers are the conv-layers with *stride* of 1 and *same* padding and $L_{3,6,9}$ are the max pooling layers with *stride* of 2 (dimensions are given in Fig. (3); strides and filter paddings are not tuned as hyperparameters in this work due to lack of enough computational resources). The proposed model repeats the blocks of 2-*conv.* and 1-*max-pooling* layers for three times which are followed with two fully connected layers and the final output layer. For each layer, we use Leaky-Rectified-Linear-Unit (Leaky-ReLU) as the activation function. For the output layer for Parkinson classification we use the Softmax classifier that uses the cross-entropy loss. For the optimization process we use Adam method [14] and all the implementations are done in *TensorFlow* framework.

For the training process of the model shown in Fig. (3), we name the model as the *original model*, where two convolutional layers are followed by a max-pooling layer, repeated three times, and finally followed by two fully connected layers. We have also considered a sub-model of the one shown in Fig. (3) where only one convolutional layer precedes each max-pooling layer, and a three conv-max-pool pairs of them are followed by two fully connected layers, and refer to this sub-model as the *simplified model*, which its number of parameters are considerably lower than the *original model* and lead to shorter training times.

For each of the *original model* and *simplified model*, we perform a random search [15] for hyper-parameter optimization. We look for optimum values of drop-out rate in the fully connected layers, regularization coefficients for kernel and bias in the 3D convolutional layers, α coefficient in the Leaky-ReLU activation function (Eq. (4)), and

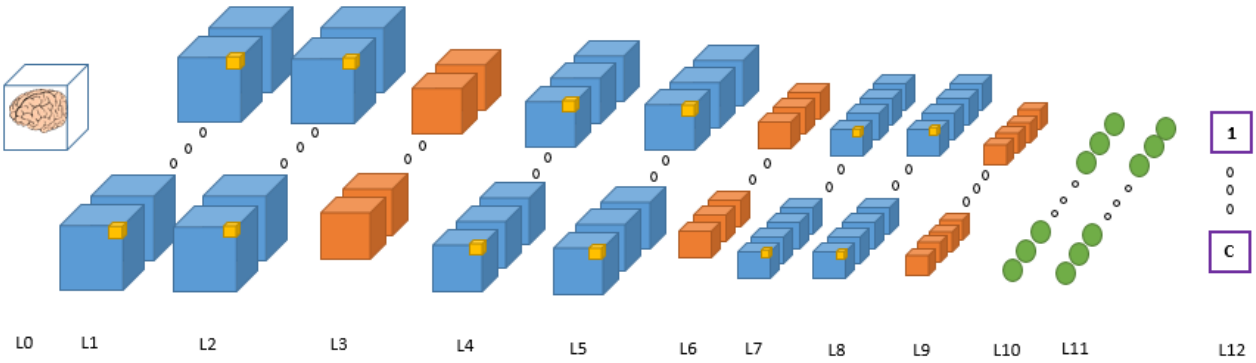


Figure 3: Architecture of the 3D-Convolutional Neural Network model: L_0: MR-Image (80×100×108×1); $L_{1,2}$: Conv. ($3^3 \times 32$); $L_{4,5}$: Conv. ($3^3 \times 64$); $L_{7,8}$: Conv. ($3^3 \times 128$); L_3, L_6, L_9: Max-pool (2^3 ,4^3 ,4^3); L_{10}, L_{11}: F.C.(512 ,128); L_{12}: Output (c) - *same* padding for $L_{1,2,4,5,7,8}$, and strides of two for max-pooling layers

learning-rate (Lr) of the optimizer.

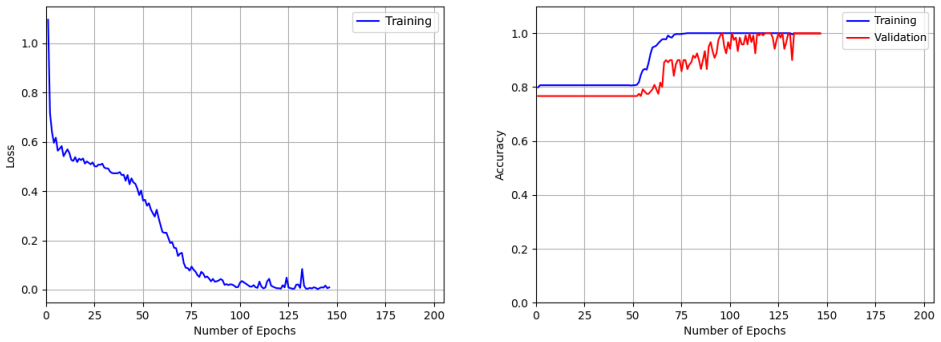
$$\text{Leaky-ReLU: } f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \quad (4)$$

Furthermore, we experiment an exponential learning-rate decay for the optimizer which has the mathematical form of $Lr = Lr_0\, e^{-kt}$ where the initial learning rate (Lr_0) and the decay steps (t) are hyper-parameters that are tuned in case of using decaying learning rate.
Following each convolutional layer, we put a normalization layer as well. For the normalization layer we have experimented batch normalization and group normalization [16] as two different approaches. Moreover, we investigated the effect of adding/removing gender and age of each patient to the model as two extra features in the last fully connected layer.

5. Experiments and Results

5.1. Experiments

Tables (2) and (3) show the experiments and hypter-parameters study results. Starting with the *original model* shown in Fig. (3) we performed a hyper-parameter study to find the optimum value of learning rate (0.00005), and when the training accuracy reached 100 percent we got accuracy of 77.1 percent on the validation set. Repeating the similar procedure with the *simplified model* described at the end of section (4) we reached to a higher validation accuracy i.e. 82 percent. This implies that probably the original model due to having too many parameters built in it might overfit the training data, leading to the lack of generalization and theeqrefor not performing well on the validation set.
In the next step of experiments, we added age and gender of the patients as two additional features to the last fully connected layer. By adding age and gender the validation accuracies of the *original model* and *simplified model* increased by 5% and 2.5 % respectively, leading to 81.2 and 84.1 percent accuracy. It was also observed the age and gender addition led to faster training process. Worth mentioning that, neglecting MR-Images and doing a simple logistic regression analysis on status of Parkinson with respect to patinets age led to 72% accuracy.
Afterwards, we added normalization layers after the convolutional layers in both *original* and *simplified* models. As the normalization layers, we experimented the training with Batch Normalization and Group Normalization. In general, both type of normalizations led to slight increase in the validation accuracies and decrease in training time (100 % train accuracy was achieved in a fewer number of epochs).
In the next set of experiments, we added bias and kernel regularizations to the layers of both *simplified* and *original* models and by hyper-parameter study we found the optimum regularization coefficients for each, shown in Table (3). Adding regularization, in both of models delayed the overfit on train dataset and led to further learning and higher validation accuracy in each.
Later on, we added drop-out, which as a regularization term affects the last two fully connected layers in both of the *simplified* and *original* models. Letting the keep probability rate ($1-$dropout rate) to vary for both of the fully connected layers, we carried out a parameter study to find the proper keep probability rates, which are not too big to lead to early overfit and not too small to inhibit learning. With the optimum values of regularization coefficient and keep probability rates shown in Table (3) the simplified model which includes age and gender of the patients led to 100 % accuracy on both the train and validation sets. We eqrefer to this as the best trained model. Just for the sake of illustration of the difference between using group and batch normalization, for the best trained model, the training loss and train and validation set accuracies as the function of number of epochs are shown in Fig. (4(a)). It can be seen that after



(a)

Figure 4: Loss value and accuracy of the training-set and validation-set during the training process

training set accuracy reaches to 100 % the model reaches to full accuracy on the validation set as well (100 % validation accuracy), and ultimately the loss function reaches to zero values. This model, on the test-set of 56 samples also led to 100 % test accuracy (see Fig. (5(c)))

For the best trained model found above, the classification accuracies on both groups of Healthy patients (HC) and Parkinson diagnosed patients (PD) is presented in section (5.2) by confusion matrices and ROC curves. Moreover, a heat-map for sensitivity analysis of the best trained model's output using image occlusion technique is presented in section (5.3).

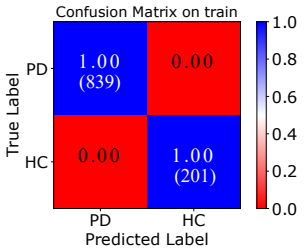
5.2. Model Evaluation

In this part, to evaluate the performance of the best trained model found in section (5.1), we find the confusion matrix where each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class or vice versa. Figs. (5(a)), (5(b)), and (5(c)) show the normalized confusion matrices of the Parkinson classification results of the best trained model for training-set, validation-set, and test-set respectively. As it can be seen the trained model performs pretty well on the validation and test sets (120 & 56 cases respectively) with classifying the MR-Images as for Healthy Conditions (HC) and Parkinson Disease (PD) with accuracy of 100 percent, and leading to zero false negatives and zero false positives.

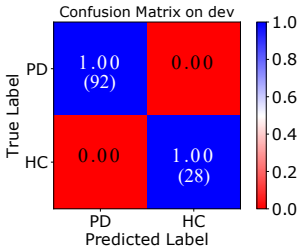
Fig. (6(a)) shows ROC curves for training, validation, and test sets where the AUC is 1, shows perfect TPR v.s FPR.

5.3. Parkinson Heat-Map of Brain

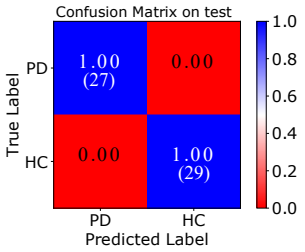
Similar to the approach in the work by Matthew et al. in 2013 [17] on visualizing and understanding Convolutional Networks, we perform a sensitivity analysis of the classifier output by occluding portions of the input image, revealing which parts of the brain are important for Parkinson diagnosis. For this reason, we performed an image occlusion



(a)



(b)



(c)

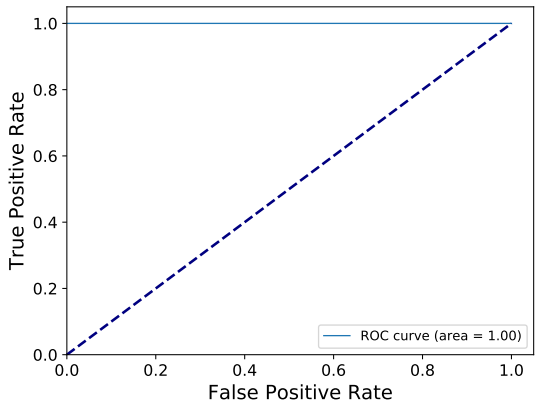
Figure 5: Normalized confusion matrix - *top* to *bottom*: train-set (1040 cases), dev-set (120 cases), test-set (56 cases) - HC: Healthy condition, PD: Parkinson disease

No.	Experiment	Accuracy	
		Train.	Val.
1	Original Model (OM)	1.0	0.771
2	Simplified Model (SM)	1.0	0.820
3	OM + Gender & Age (OM-GA)	1.0	0.812
4	SM + Gender & Age (SM-GA)	1.0	0.841
5	OM-GA + Batch Normalization (OM-GA-B)	1.0	0.821
6	SM-GA + Batch Normalization (SM-GA-B)	1.0	0.847
7	OM-GA + Group Normalization (OM-GA-G)	1.0	0.820
8	SM-GA + Group Normalization (SM-GA-G)	1.0	0.849
9	OM-GA-G + Regularization (OM-GA-GR)	1.0	0.895
10	SM-GA-G + Regularization (SM-GA-GR)	1.0	0.935
11	OM-GA-GR + Drop-out (OM-GA-GRD)	1.0	0.947
12	SM-GA-GR + Drop-out (SM-GA-GRD)	1.0	1.000

Table 2: F_2-score values of different experiments on training set and validation set; OM: Original Model, SM: Simplified Model, GA: Gender & Age included, B: Batch Normalization, G: Group Normalization, R: Bias & Kernel Regularization, D: with Drop-out

No.	Experiment	Lr	α	Rc	kp_1	kp_2
1	OM	0.00005	0	0	1	1
2	SM	0.00005	0	0	1	1
3	OM-GA	0.00020	0	0	1	1
4	SM-GA	0.00005	0	0	1	1
5	OM-GA-B	0.00005	0.01	0	1	1
6	SM-GA-B	0.00001	0.01	0	1	1
7	OM-GA-G	0.00001	0.01	0	1	1
8	SM-GA-G	0.00001	0.01	0	1	1
9	OM-GA-GR	0.00001	0.01	0.05	1	1
10	SM-GA-GR	0.00001	0.01	0.001	1	1
11	OM-GA-GRD	0.00001	0.01	0.05	0.2	0.35
12	SM-GA-GRD	0.00001	0.01	0.001	0.45	0.5

Table 3: Parameters of different experiments of Table (2); Lr: Learning-rate, α Leaky-ReLU function's coefficient, Rc: Regularization coefficient, kp_i: keep probability of the $i-th$ fully connected layer (keep probability = 1−dropout probability)



(a)

Figure 6: ROC curve for training-set (1040 cases), dev-set (120 cases), and test-set (56 cases)

analysis on our best-trained model found in section (5.1) by translating a box with the size of $2\times2\times2$ zero-valued voxels along the whole MR-Image of a Parkinson patient that was correctly labeled as Parkinson Disease (PD) by the trained model. In the heat-maps in Fig. (7) White areas are irrelevant as they don't change the confidence of the prediction, red areas increase the confidence, and blue areas decrease the confidence of the model suggesting that they are areas that are important for diagnosing of PD.

By looking at Fig. (7) in both Coronal and Axial views we see that *Basal Ganglia* and *Substantia Nigra* (bottom blue regions) together with *Superior Parietal* part on right hemisphere of the brain are found to be of critical importance in diagnosis of Parkinson, where the former one is completely corroborated by medical studies that when dopamine receptors in the striatum are not adequately stimulated those parts get either under- or over-stimulated and lead to Parkinson. However, our latter finding (i.e. *Superior Parietal* part) is a novel finding which asserts that not only the *Basal Ganglia* but also *Superior Parietal* part of the brain play role in Parkinson disease.

6. Conclusion

In this work, we successfully could build a machine learning model to diagnose Parkinson in patients using MR-Images. We achieved 100 % accuracy on the validation and test sets, built a brain heat-map for Parkinson diagnosis and verified that *Basal Ganglia* and *Substantia Nigra* part of brain as already were known by medical experts are important in diagnosis of Parkinson, and for the first time we found out that *Superior Parietal* part on right hemisphere of the brain is also very critical in diagnosis of Parkinson.

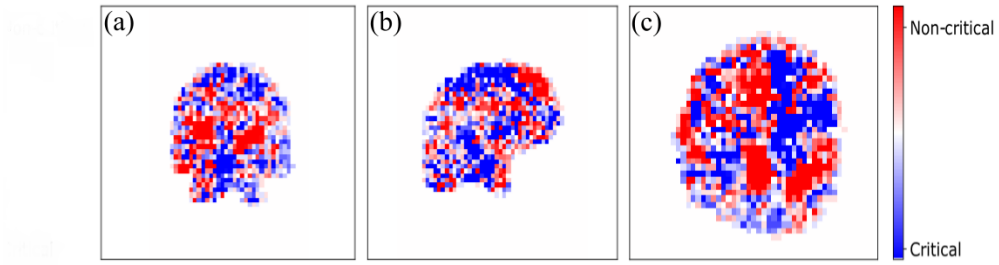


Figure 7: Brain's Heat-map for Parkinson diagnosis - from *left* to *right*: (a) Coronal, (b) Axial, and (c) Sagittal views

References

[1] "Parkinson's disease information page." https://www.ninds.nih.gov/Disorders/All-Disorders/Parkinsons-Disease-Information-Page.

[2] G. . Disease, I. Incidence, and P. Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 19902015: a systematic analysis for the global burden of disease study 2015." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5055577/, 2015.

[3] "Parkinson's disease information page." https://www.ninds.nih.gov/Disorders/All-Disorders/Parkinsons-Disease-Information-Page.

[4] C. Salvatorec, "Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy,"

[5] S. Esmaeilzadeh, O. Khebzegga, and M. Moradshahi, "Clinical parameters prediction for gait disorder recognition," 2018.

[6] L. zhang, "Classification of parkinson's disease and essential tremor based on structural mri,"

[7] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Deep multi-task multi-channel learning for joint classification and regression of brain status. lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)),"

[8] B. Peng, "A multilevel-roi-features-based machine learning method for detection of morphometric biomarkers in parkinsons disease,"

[9] M. N. Ahmed and A. A. Farag, "Two-stage neural network for volume segmentation of medical images," *Neural Networks, International Conference on pp 1373- 1378 vol.3.*

[10] D. Gil and M. Johnsson, "Diagnosing parkinson by using artificial neural networks and support vector machines," *Global Journal of Computer Science and technology*, vol. 9, pp. 63–71, 2009.

[11] K. Marek, "The parkinson progression marker initiative (ppmi).," *Progress in Neurobiology*, 2011.

[12] K. Boesen, K. Rehm, and K. Schaper, "Quantitative comparison of four brain extraction algorithms," *NeuroImage*, vol. 22, no. 3, pp. 1255–1261, 2004.

[13] S. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, 2002.

[14] P. iederik and J. Kingma, "Adam: A method for stochastic optimization," 2014.

[15] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research 13 (2012) 281-305.*

[16] Y. Wu and K. He, "Group normalization." https://github.com/taki0112/Group_Normalization-Tensorflow, journal: arXiv:1803.08494v2 [cs.CV] 24 Apr 2018.

[17] R. F. Matthew D. Zeiler, "Visualizing and understanding convolutional networks," *arXiv:1311.2901v3 [cs.CV] 28 Nov 2013*, 2013.