Prediction of outcome in Parkinson's disease patients from DAT SPECT images using a convolutional neural network

Matthew P. Adams[1], Bao Yang[1], *Student Member*, *IEEE*,
Arman Rahmim[2], *Senior Member*, *IEEE*, and Jing Tang[1], *Senior Member*, *IEEE*

***Abstract* – Dopamine transporter (DAT) SPECT imaging is widely used for the diagnosis of Parkinson's disease (PD). Investigations on quantitative analysis of DAT images have been performed using radiomic features together with non-imaging features to predict a patient's outcome. The purpose of this study is to take the entire DAT image, without feature extraction, for predicting a patient's motor function indicated by part III of the unified Parkinson's disease rating scale (UPDRS) using a convolutional neural network (CNN). We cast the motor function score prediction as a categorization problem that decides on whether a given patient will have a score below or above 30 in 4 years from baseline. This prediction practice was first conducted using only the baseline UPDRS_III score as the input, which resulted in an accuracy of 64.5%+/-8.1%. The performance was then evaluated using the baseline score with the DAT image to assess whether the latter has added value for prediction. The resulting mean accuracy was significantly improved, reaching to 70.6%+/-7.7%. This CNN-based motor function score predicting scheme demonstrates the contribution of DAT SPECT images, which shows their promise in improving progression tracking for patients with PD.**

I. Introduction

DOPAMINE transporter (DAT) SPECT imaging is extensively used for the diagnosis of Parkinson's disease (PD) [1]. Clinical applications of DAT SPECT images mainly rely on visual inspection, while quantitative analysis may facilitate early stage detection and improve progression tracking. Radiomics based image analysis has shown success in applications such as predicting the Gleason score of Prostate Cancer from MRI images and planning cancer treatment based on FDG-PET images [2][3]. Predicting clinical outcomes in PD patients has opened research opportunities in applying machine-learning based techniques. We investigated the use of radiomic features extracted from DAT SPECT images along with non-imaging features to make predictions about future motor function, and identified which radiomic features added to the performance [4][5]. These efforts highlight the growing interest in moving away from visual inspection of images to automatic data-driven techniques.

While research in using radiomic features to make predictions has demonstrated success, the approach comes with its own challenges. Image features need to be extracted through segmentation and registration. After feature extraction, additional steps are required to determine which features actually contribute to the prediction performance. Demonstrating improved performance in image classification and detection problems, convolutional neural networks (CNNs) could be applied to address these challenges [6]. Before the use of CNNs, image classification algorithms required hand-engineered specialized filters which needs specific image processing expertise. The improvement is due to the ability to automatically identify which features are relevant to the outcome by automatically updating network parameters based on the training data. A well-trained neural network can then generalize to examples that have not been learned on. This capability has led deep learning to wider applications to medical images for disease classification [7]. A CNN based technique was developed to automatically diagnose PD through training by DAT SPECT images without feature extraction [8].

In this study, an approach is taken to predict clinical outcome using the entire DAT SPECT images rather than the extracted radiomic features. We believe this has promise since we have demonstrated that radiomic analysis of DAT SPECT images contributed to predicting motor function scores [4]. We accomplish the goal of learning from the entire image by developing a CNN based prediction technique, where we use the baseline DAT SPECT image together with the motor function score to make a prediction on the patient's motor function at year 4. Taking advantage of the learning and generalization capability of the CNN, our proposed technique is expected to handle subject-specific images with baseline scores and to predict the future motor score categories of patients.

II. Methods and Materials

Before the prediction process, the required data was gathered and pre-processed, which included the baseline DAT SPECT images and the motor part (III) of the unified Parkinson's disease rating scale (UPDRS), at both baseline and year 4. UPDRS_III is a widely accepted measure of a patient's motor function, where a higher number corresponds

Manuscript received December 15, 2018. This work was supported in part by the National Science Foundation ECCS-1454552.

[1]M. P. Adams, B. Yang, and J. Tang (e-mail: jtang@oakland.edu) are with the Department of Electrical and Computer Engineering, Oakland University, Rochester, MI, USA.

[2]A. Rahmim is with the Departments of Radiology and Physics, The University of British Columbia, Vancouver, BC, CA.

978-1-5386-8495-5/18/$31.00 ©2018 IEEE

to more severe symptoms shown in clinical observation [9]. A CNN was trained and tested to perform predictions based on the pre-processed data. Since the CNN inputs could include both the images and the non-imaging data, multiple experiments were conducted to determine whether combining data results in a stronger prediction than from a single data source alone.

A. Patient Data Collection

We gathered data of 293 patients diagnosed with PD from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). There were a number of criteria that a patient needed to meet for inclusion. Any patient that was positively diagnosed with PD was selected, who had baseline and year 4 UPDRS_III clinical evaluation scores available and also DAT SPECT images taken at baseline. Figure 1 shows year 4 versus the baseline motor function scores across the 293 patients, which highlights the challenges this prediction task would face. It is noted that the UPDRS_III scores were only considered usable if the patient was not actively on medication at the time of evaluation.

B. Patient Data Preparation

The data for all PD patients was pre-processed before being used for predictions. The year 4 motor score was converted to a category label based on whether the score is greater than a threshold value of 30, which is to be used as the prediction output. This threshold was chosen because the vast majority of year 4 scores ranged from 0-60. The DAT SPECT images of 91×109×91 were zero padded to the size of 109×109×109, where each image serves as an input with the baseline UPDRS_III score to the CNN.

Before the images were used as inputs to the CNN, each voxel was normalized to fall within the range of [0, 1] using

$$I_{\mathrm{NORM}} = I_{\mathrm{ORIGINAL}} \frac{1}{I_{\mathrm{MAX}} - I_{\mathrm{MIN}}},$$

where for a given image, I_{NORM} represents the normalized voxel intensity, and I_{ORIGINAL}, I_{MAX}, I_{MIN} are the original, maximum, and minimum intensities before normalization. Normalization was adopted to keep the distribution of feature values consistent after they are computed by the CNN, which supposedly leads to faster and more stable learning convergence [10].

Another challenge presented is that the UPDRS_III score is subject to wide variability between evaluations, even among the same patient. This is largely due to the variability in the patient's condition plus the subjective nature of the clinical evaluation. The evaluator must assign a score from 0 to 4 for each motor function task based on their assessment of the patient's performance, but the tasks do not have a hard measurement associated with them [9]. This makes the result a combination of the patient's performance and the evaluator's assessment at the specific time. Such variability was mitigated in other works by averaging each score of interest with any available scores 6 months before or after, if available [4]. This approach was adopted in this study as well.

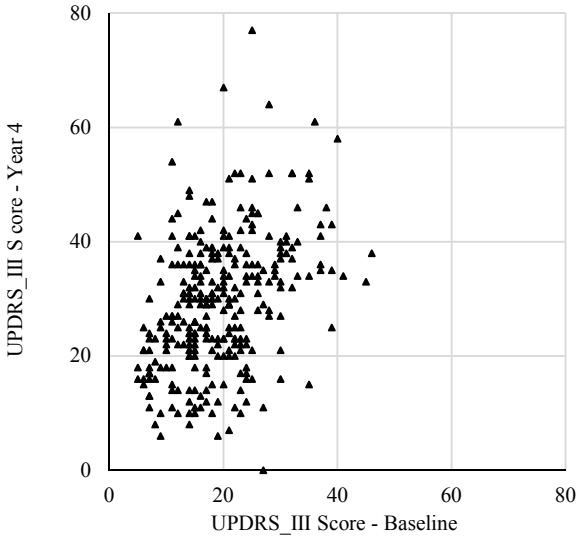

Fig. 1. Baseline and year 4 UPDRS_III score across patient dataset.

C. CNN based Motor Function Prediction

A CNN is a computational model inspired from biological nervous systems and brain structures, which are particularly useful for analyzing images. A typical CNN structure consists of artificial neurons which are organized in a number of layers including an input, an output, and several intermediate layers such as convolutions and pooling. Each neuron after a convolutional layer is fed with the responses from a receptive field of pre-defined size, which is designed based on the size of imaging features that the CNN would identify. During supervised learning, the parameters of these receptive fields are updated so that the CNN can learn the relationship between the input and the output, without explicitly defining which imaging features to focus on [6]. Given the DAT SPECT image with the baseline score as input and year 4 motor function category as output, we designed a CNN to predict whether the year 4 score will be below or above 30.

The CNN structure used in this study is shown in Figure 2. The DAT image forms the input layer and the year 4 score category based on threshold forms the output layer. The hidden layers consist of three 3-D convolutional layers, with filter size of 7×7×7, 5×5×5, and 3×3×3 and stride of three, one, and one voxels, and 16, 32, and 64 filter banks, respectively. Rectified linear unit activation layers are applied after each convolutional layer. Each of these convolutional layers is followed by a max-pooling layer of size 3×3×3 and stride of two voxels. The output of these multiple layers are then flattened into a vector, and the baseline UPDRS_III score is added as a single scalar value. The output of this flattened layer feeds into a fully connected layer with 1024 neurons, which are then connected to the output layer with two outputs, one for each motor function category.

The described CNN was implemented in Tensorflow, which is a widely used practical framework for machine learning [11]. During the training procedure, the loss

backpropagation and Adam optimization algorithms were used to update the weight of every connection, with cross-entropy used as the cost function [12]. The learning rate was initialized at 0.001 empirically, with a total of 50 iterations. The fully connected layer includes a 0.5 dropout, which used only during training.

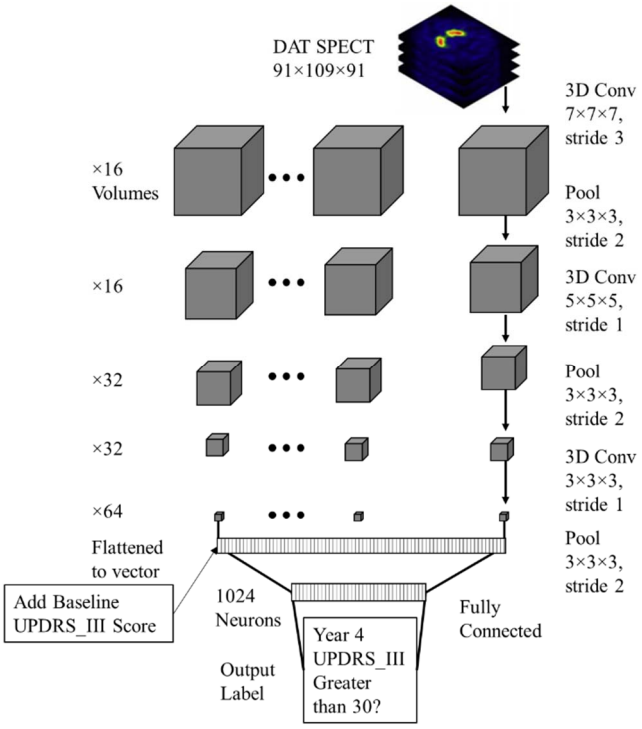

Fig. 2. Convolutional neural network framework for predicting year 4 UPRDS_III score categories.

This procedure was repeated to perform 10-fold stratified cross-validation [13]. The data was randomly split into 10 groups, or folds, of roughly equal size and proportion of patients below and above the year 4 score threshold value, to ensure each split is representative of the overall population. For one run of cross-validation, training and validation was repeated 10 times, with 1 group being used for validation and the other 9 for training. This way ensures that all subjects were tested by the CNN for a comprehensive evaluation. One full round of the 10-fold cross-validation prediction procedure is illustrated in Figure 3. The cross-validation procedure was repeated for a total of 10 runs, which results in 100 total accuracy measurements (10 folds per run).

D. Prediction Evaluation

Following the described cross-validation strategy, each predicted year 4 UPDRS_III score classification was compared with the subject's actual score category to decide whether the CNN made a correct prediction. The metric for evaluating the performance of the trained network is percent accuracy, which is calculated by the number of correctly categorized patients out of the total number of patients under test. The prediction performance was evaluated using the baseline UPDRS_III score as input with the image data or without the image data by setting all voxel values to 0. This allows for a comparison to see if the CNN is able to identify features in the image that contribute to the prediction accuracy. A two sample t-test was used to determine whether there is a statistically significant improvement when the image data is included.

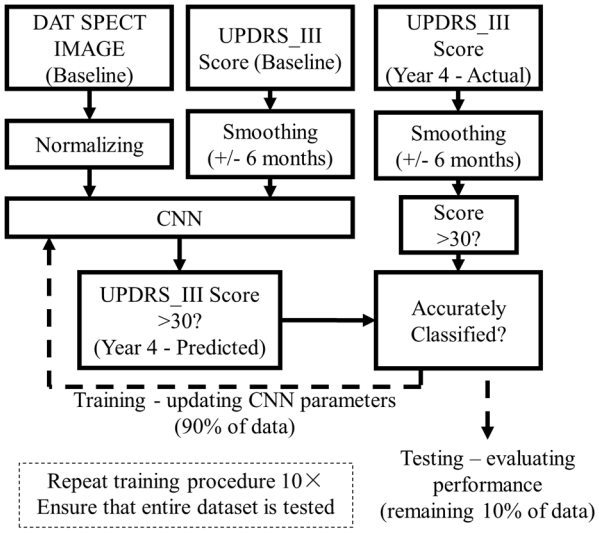

Fig. 3. Overall scheme of CNN based prediction for one round of cross-validation.

III. RESULTS

The resulting CNN performance accuracy across one cross-validation run, is shown in Table I. The prediction of year 4 UPDRS_III score categories demonstrates good accuracy with the provided inputs, which have a wide variety of DAT SPECT images and baseline scores from a diverse patient demographic.

TABLE I
PREDICTION ACCURACY FOR ONE CROSS-VALIDATION RUN

Cross-Validation Fold	Number of Test Patients	Test Accuracy (%)	Test Accuracy w/o Images (%)
1	30	73.3	63.3
2	30	66.7	66.4
3	30	83.3	73.3
4	29	72.4	69.0
5	29	72.4	72.4
6	29	65.5	72.4
7	29	62.1	62.1
8	29	69.0	65.5
9	29	69.0	65.5
10	29	69.0	58.6

The mean accuracy across all cross-validation runs is shown in Table II as 70.7±7.7%, which demonstrates the

effectiveness of the CNN based technique. In addition to the CNN based prediction with DAT SPECT image plus baseline score as the input, the performance of the same prediction conducted with only the baseline score as the input for comparison purpose is also shown in Tables I and II. Omitting all the image data, an accuracy of 64.5±8.1% was resulted. A two sample t-test was conducted on these results which yields a two-tailed p-value of <0.0001, demonstrating the added value of the image data.

TABLE II
MEAN AND SD OF ACCURACY ACROSS THE 10 CROSS-VALIDATION RUNS

Data Input	Mean (%)	SD (%)
UPDRS_III	64.5	8.1
UPDRS_III & DAT SPECT	70.7	7.7

IV. CONCLUSIONS

Using the UPDRS_III score and the DAT SPECT images at baseline, we predicted patients' outcome of year 4. The CNN scheme was developed to accomplish this task, as this approach directly learns from the image, and no segmentation and feature extraction is needed. The network was trained and tested with the data of 293 subjects using stratified cross-validation. The prediction achieved an average accuracy of 70.7% for this challenging task and demonstrated the feasibility of applying CNNs on images and non-imaging features for clinical prediction.

A future work that progresses from this topic is to use a similar method on a more specific prediction task. Based on the results shown in this study towards predicting the year 4 UPDRS_III score category, we would like to investigate if similar improvements would be brought by the images to predict the actual score. We believe that continuing efforts in DAT SPECT image analysis have the potential to enhance diagnostic accuracy and improve outcome prediction in PD.

REFERENCES

[1] L. Wang *et al.*, "SPECT Molecular Imaging in Parkinson's Disease", *Journal of Biomedicine and Biotechnology*, vol. 2012, 412486, 2012.

[2] A. Chaddad *et al.*, "Multimodal Radiomic Features for the Predicting Gleason Score of Prostate Cancer", *Cancers (Basel)*, vol. 10, no. 8, pp. 249, 2018.

[3] C. Lian *et al.*, "Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction", *Medical Image Analysis*, vol. 32, pp. 257-268, 2016.

[4] A. Rahmim *et al.*, "Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images", *NeuroImage: Clinical*, vol. 16, pp, 539-544, 2017.

[5] J. Tang *et al.*, "Artificial neural network based outcome prediction in DAT SPECT imaging of Parkinson's disease", *J. Nucl. Med.*, vol. 58, pp. 292, 2017.

[6] Y. LeCun *et al.*, "Deep learning", *Nature*, 521(7553), pp. 436-444, 2015.

[7] W. Shen *et al.*, "Multi-scale Convolutional Neural Networks for Lung Nodule Classification", *Information Processing in Medical Imaging*, pp. 588-599, 2015.

[8] H. Choi *et al.*, "Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging", *NeuroImage: Clinical*, vol. 16, pp. 586-594, 2017.

[9] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, "The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations", *Movement Disorders,* vol. 18, no. 7, pp. 738-750, 2003.

[10] C. Parmar *et al.*, "Data Analysis Strategies in Medical Imaging", *Clinical Cancer Research*, vol. 24, no. 15, pp. 3492-3499, 2018.

[11] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning", *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association, pp. 265-283, 2016.

[12] D. Kingma & J. Ba, "ADAM: A Method for Stochastic Optimization", *International Conference on Learning Representations*, 2015.

[13] Stone M., "Cross-validatory choice and assessment of statistical predictions", *J. Royal Stat. Soc.*, 36(2), pp. 111–147, 1974.

978-1-5386-8495-5/18/$31.00 ©2018 IEEE