A Deep Learning Approach for Grading of Motor Impairment Severity in Parkinson’s Disease

Prithvi Prakash, Rachneet Kaur, Joshua Levy, Richard Sowers, James Brašić, and Manuel E. Hernandez, Member, IEEE

Abstract—Objective and quantitative monitoring of movement impairments is crucial for detecting progression in neurological conditions such as Parkinson’s disease (PD). This study examined the ability of deep learning approaches to grade motor impairment severity in a modified version of the Movement Disorders Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS) using low-cost wearable sensors. A convolutional neural network architecture, XceptionTime, was used to classify lower and higher levels of motor impairment in persons with PD, across five distinct rhythmic tasks: finger tapping, hand movements, pronation-supination movements of the hands, toe tapping, and leg agility. In addition, an aggregate model was trained on data from all tasks together for evaluating bradykinesia symptom severity in PD. The model performance was highest in the hand movement tasks with an accuracy of 82.6% in the held-out test dataset; the accuracy for the aggregate model was 79.7%, however, it demonstrated the lowest variability. Overall, these findings suggest the feasibility of integrating low-cost wearable technology and deep learning approaches to automatically and objectively quantify motor impairment in persons with PD. This approach may provide a viable solution for a widely deployable telemedicine solution.

Parkinson’s disease, deep learning, disease severity, wearable sensors

I. INTRODUCTION

Parkinson’s disease (PD) is a common chronic neurological disorder [1], which is only expected to increase in prevalence as the proportion of adults over 65 years of age increases worldwide [2]. While PD is characterized clinically by primarily motor symptoms such as bradykinesia, rigidity, resting tremor, and postural instability [3], automatic and objective evaluations of motor assessments are still lacking in clinical practice.

Clinical evaluation of patients can often be time-consuming and inefficient. While multiple rating scales have been proposed, the gold standard evaluation of motor impairment in people with PD is the Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS) [4]. However, initial signs and symptoms of PD are subtle, hence further development of tools and methods to detect early changes, track progression, and monitor treatment is needed.

Telemedicine solutions may be well suited to evaluate people with PD, particularly in rural and underserved communities that have usually lacked access to healthcare [5], [6]. While the COVID-19 pandemic, fast tracked the deployment of telemedicine applications to address the needs of people with PD, there are still challenges related to privacy, poor connectivity, and access to technology.

The use of machine learning, together with behavioral data, has demonstrated promise in distinguishing between pathological and physiological motor responses arising from PD [7], [8], [9] and several other neurological conditions [10], [11]. The integration of low-cost wearable technology and deep learning approaches may provide a viable approach towards the development of robust and more widely deployable telemedicine solutions.

In this study, we explored the ability of deep learning approaches to automatically and objectively grade motor impairments in a modified version of the MDS-UPDRS using low cost wearable sensors [12]. We evaluated the use of a convolutional neural network, XceptionTime [13], to classify motor impairment symptoms, based on bradykinesia. Overall, this study demonstrates the feasibility of deep learning in unison with a low-cost movement monitor to automatically and objectively quantify motor impairment in persons with PD.

II. METHODS & MATERIALS

A. Protocol

A custom-built low-cost quantitative measurement system, consisting of a 3-axis accelerometer (ADXL335) and evaluation board (EVAL-ADXL335Z) with USB connectors, was used to collect continuous movement data from the extremities of people with PD [12]. An examiner certified
in the administration of the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) administered the protocol [12] and provided a ground-truth evaluation of movement impairment. Repetitive movements were recorded at 80 Hz using two accelerometers placed on each index finger (Sensor 1) and two inches proximal to each wrist (Sensor 2) for upper extremity movements (e.g., finger tapping [FT], hand movements [HM], and pronation-supination movement of the hands [PS]), or the anterior surface of each tibia two inches proximal to the medial malleolus (Sensor 3) and on the dorsal surface of the proximal phalanx of each big toe (Sensor 4) for lower extremity movements (e.g., toe tapping [TT] and leg agility [LA]). Participants returned after a week or more for repeat testing. Figure 1 depicts our data analysis workflow pipeline.

B. Participants

We used the open-source dataset [14]. Overall, 30 participants (20 persons with PD, 8 healthy controls, and 1 person with multiple system atrophy) were recorded; further, 19 of these had a retest session. The data was collected in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) and the protocol was approved by the Johns Hopkins Institutional Review Board. This study utilized 32 recorded sessions from persons with PD.

C. Data Preprocessing

For each participant and each task (FT, HM, PS, TT, LA), the resultant acceleration for each sensor was calculated as:

\[
\text{Acceleration}_i = \sqrt{a^2_{x_i} + a^2_{y_i} + a^2_{z_i}},
\]

where \( i \) represents the corresponding sensor and \( a_x, a_y, a_z \) represents acceleration in the x-axis, y-axis, and z-axis. The jerk of the resultant acceleration was calculated as the derivative of each time series. The root mean square of the resultant acceleration for each sensor was calculated as:

\[
\text{Jerk}_i = \sqrt{\left(\frac{\text{derivative of each time series}}{\text{root mean square of the acceleration}}\right)^2 + \left(\frac{\text{derivative of each time series}}{\text{root mean square of the acceleration}}\right)^2 + \left(\frac{\text{derivative of each time series}}{\text{root mean square of the acceleration}}\right)^2},
\]

Replicating the movement of the participant were also shared across each of the sensors. Lastly, labels on a scale of 0 to 4 were aggregated to either low (0/1) or high (3/4) levels of motor impairment, while ignoring all signals labelled as severity-level 2.

D. Experiment Design

From the pool of 48 total tests recorded in [14], we only considered tests where the corresponding subject was diagnosed with PD; this filtering resulted in 32 remaining sessions from persons with PD. This diagnosed group of subjects were randomly split into two pools: 1) train-validate pool, with 25 subjects used for training and validating the model, and 2) test pool, with 7 subjects for testing the model. Stratifying the patients this way ensured there was no data leakage while training and validating the model’s performance.

E. Model Training

To train our model, we used data from 25 persons with PD from the train-validate pool. This data was further randomly split into a training and a separate validation pool. For implementation purposes, we used tsai Python library [15], a deep learning package for time series tasks. We experimented with various deep learning architectures. Empirically, we found that the convolutional neural network-based architecture, XceptionTime [13], performed the best on our dataset. For our final models we trained the XceptionTime network for 50 epochs with a learning rate of 0.001 on an NVIDIA Quadro M4000 8 GB GPU machine. The XceptionTime-based binary classifier was optimized using the following binary cross entropy loss function \( L_{BCE} \):

\[
L_{BCE} = -\frac{1}{n} \sum_{i=1}^{n} (Y_i \cdot \log(p_i) + (1 - Y_i) \cdot \log(1 - p_i)),
\]

where \( Y_i \) represent the labels, \( p_i \) the probability of a prediction, and \( n \) the number of samples.

F. Model Inference

The input signals in the test pool were preprocessed and broken into snippets of 400 time steps (i.e., 5 seconds), similar to the Section II-C. These snippets were fed into the trained model for classification for every individual segment. The predictions from each individual segment were then aggregated together in a max-vote fashion to grade the overall severity level of the movement (see Figure 1).

III. RESULTS

The above described experimental setup was used to train a total of 6 distinct models: 5 separate models for each of the rhythmic tasks: FT, HM, PS, TT, LA; and an aggregate model that was trained on data from all tasks together. The training setup for each of the models was repeated 10 times for different train-test splits to ensure there was no sampling bias.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Test (Raw)</th>
<th>Test (Max-Vote)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>0.7006</td>
<td>0.7586</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.7006</td>
<td>0.7586</td>
</tr>
<tr>
<td>InceptionTime</td>
<td>0.7053</td>
<td>0.7586</td>
</tr>
<tr>
<td>XceptionTime</td>
<td>0.7888</td>
<td><strong>0.8620</strong></td>
</tr>
<tr>
<td>OmniScaleCNN</td>
<td>0.7006</td>
<td>0.7413</td>
</tr>
<tr>
<td>RNN</td>
<td>0.3828</td>
<td>0.3103</td>
</tr>
<tr>
<td>RNNPlus</td>
<td>0.6705</td>
<td>0.6896</td>
</tr>
<tr>
<td>RNNFCN</td>
<td>0.7006</td>
<td>0.7586</td>
</tr>
<tr>
<td>RNNFCNPlus</td>
<td>0.8676</td>
<td>0.7586</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.6171</td>
<td>0.6379</td>
</tr>
<tr>
<td>TST</td>
<td>0.5545</td>
<td>0.6551</td>
</tr>
</tbody>
</table>

TABLE I: Performance benchmarks of various Neural Network architectures on a random patient level train-test split. We see the best performance on the test split (Raw and Max-Vote) from the CNN based architecture XceptionTime Architecture.
Fig. 1: A. Schematic of system with data logger, 3-axis accelerometers, and sensor placement on participants (Courtesy of Jenny-Ann Phan, MD-Ph.D.). B. Workflow diagram of analysis and model development for motor impairment classification in Parkinson’s Disease

while training and benchmarking the model’s performance. Table II summarizes the trained model’s performance on the holdout test dataset.

The performance, using max-vote, was highest in the hand movement tasks, with an accuracy of 82.6% ± 6.8% in the holdout test dataset. The aggregate model performance was relatively lower at 79.7% ± 4.0%, however, it demonstrated the lowest variability, as shown in Figure 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Raw Accuracy</th>
<th>Max-vote Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>0.6583 ± 0.1094</td>
<td>0.7807 ± 0.1384</td>
</tr>
<tr>
<td>HM</td>
<td>0.7796 ± 0.079</td>
<td>0.8255 ± 0.0683</td>
</tr>
<tr>
<td>PS</td>
<td>0.6769 ± 0.0937</td>
<td>0.768 ± 0.105</td>
</tr>
<tr>
<td>TT</td>
<td>0.7591 ± 0.0764</td>
<td>0.7331 ± 0.1023</td>
</tr>
<tr>
<td>LA</td>
<td>0.6829 ± 0.1734</td>
<td>0.7537 ± 0.1542</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.751 ± 0.0325</td>
<td>0.7965 ± 0.0405</td>
</tr>
</tbody>
</table>

TABLE II: Test Accuracy Results. Note: FT = finger tapping; HM = hand movements; PS = pronation-supination movements of the hand; TT = toe tapping; LA = leg agility.

IV. DISCUSSION

Overall, we demonstrated the feasibility of integrating low-cost wearable technology with deep learning approaches for automatically and objectively quantifying motor impairment severity, based on bradykinesia, in persons with PD. Consistent with clinical evaluations [4], we found that an aggregate model provided the lowest variability in classification performance against holdout data from unseen participants.

This work found a convolutional neural network architecture (XceptionTime) to successfully classify lower and higher levels of motor impairment in persons with PD, across five distinct rhythmic tasks, consistent with recent work examining the clinical severity of persons with PD in finger tapping, hand movements, and rapid alternating movements [16], [17]. While recent work has achieved high performances in classifying persons with PD versus healthy controls [16], [17], our work focuses on classifying bradykinesia severity in individual motor tasks, so as to better replicate the workflow of clinicians in telemedicine or typical clinical settings. While the integration of wearable devices with machine learning may support PD diagnosis in the future, these systems provide an accurate monitoring solution at home [18], [19].

This study provides further support for the integration of both upper and lower extremity rhythmic tasks for the objective evaluation of bradykinesia in persons with PD. While the initial results are promising, further comparisons in a larger cohort will be useful for evaluating generalizability of the findings. In addition, further refinement of the hardware system will be beneficial to increase the ease of use and
improve deployability in a telemedicine setting.

Future work should focus on examining additional frameworks and symptoms to provide clinicians with a more comprehensive evaluation of motor changes in persons with PD. Furthermore, future work should consider the translational aspects of the technology, particularly in lower resource settings and across different patient populations. The use of wearable sensors for symptom monitoring is in contrast with the use of recent video-based approaches for kinematic evaluation in persons with PD [20], [21], [22]. However, in conditions where challenges still lie due to privacy, poor connectivity, and access to technology, such as in rural and underserved communities [5], [6], low-cost sensors with on-board classification and recording may provide a viable solution for a widely deployable telemedicine solution.

ACKNOWLEDGMENT

The International Parkinson and Movement Disorder Society kindly gave permission to use the Movement Disorder Society-Sponsored Revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS). We would like to thank Timothy Harrigan for instrumentation and data collection, as well as all the volunteers who participated in this study.

REFERENCES