Predicting the Progression of Parkinson's Disease MDS-UPDRS-III Motor Severity Score from Gait Data using Deep Learning

Rana Zia Ur Rehman, Lynn Rochester, Alison J. Yarnall, Silvia Del Din, *Member, IEEE*

Abstract— **Parkinson's disease (PD) is a common neurodegenerative disease presenting with both motor and non-motor symptoms. Among PD motor symptoms, gait impairments are common and evolve over time. PD motor symptoms severity can be evaluated using clinical scales such as the Movement Disorder Society Unified Parkinson's Rating Scale part III (MDS-UPDRS-III), which depend on the patient's status at the time of assessment and are limited by subjectivity. Objective quantification of motor symptoms (i.e. gait) with wearable technology paired with Deep Learning (DL) techniques could help assess motor severity. The aims of this study were to: (i) apply DL techniques to wearable-based gait data to estimate MDS-UPDRS-III scores; (ii) test the DL approach on longitudinal dataset to predict the progression of MDS-UPDRS-III scores. PD gait was measured in the laboratory, during a 2 minute continuous walk, with a sensor positioned on the lower back. A DL Convolutional Neural Network (CNN) was trained on 70 PD subjects (mean disease duration: 3.5 years), validated on 58 subjects (mean disease duration: 5 years) and tested on 46 subjects (mean disease duration: 6.5 years). Model performance was evaluated on longitudinal data by quantifying the association (Pearson correlation (r)), absolute agreement (Intraclass correlation (ICC)) and mean absolute error between the predicted and true MDS-UPDRS-III. Results showed that MDS-UPDRS-III scores predicted with the proposed model, strongly correlated (r=0.82) and had a good agreement (ICC(2,1)=0.76) with true values; the mean absolute error for the predicted MDS-UPDRS-III scores was 6.29 points. The results from this study are encouraging and show that a DL-CNN model trained on baseline wearable-based gait data could be used to assess PD motor severity after 3 years.**

Clinical Relevance— **Gait assessed with wearable technology paired with DL-CNN can estimate PD motor symptom severity and progression to support clinical decision making.**

I. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease and presents both motor and non-motor symptoms [1]. Symptom severity assessment is performed by expert clinicians using clinical scales such as the Movement Disorder Society Unified Parkinson's disease rating scale (MDS-UPDRS). The MDS-UPDRS consists of four parts: mental dysfunction and mood, (Part I), motor experiences of daily living (Part II), motor symptoms severity (Part III: UPDRS-III), and treatment related motor and non-motor complications (Part IV). Although highly relevant and the clinical "gold standard", there are some notable limitations. Assessment is performed at one point in time and requires a face to face visit. This increases patient burden; does not capture fluctuations and is subjective in nature. Collectively this reduces reliability. Therefore, novel, objective, less time-consuming and inexpensive methods are required to accurately and reliably quantify PD motor severity and predict motor symptom decline in order to support clinical management and decision making.

Among PD motor symptoms, gait impairments are common and evolve over time [2, 3]. Gait performance is considered a marker of global health and brain function [4]. Wearable technology allows gait to be measured in the clinic and continuously in real-world [5]. Gait analysis has been shown to be sensitive to PD progression [2, 3]. Previous work showed how machine learning (ML) techniques applied to wearable-based gait data can accurately classify PD from healthy older adults [6-9]. ML and Deep Learning (DL) approaches have also been utilized to estimate UPDRS-III [10-12]. A study by Parisi *et al.* [10] used gait, leg agility, and sit-to-stand movements from 34 PD subjects measured with wearable technology to automatically quantify the UPDRS-III. Models trained on kinematic features showed only poor to moderate correlation between the estimated MDS-UPDRS-III and the real scores. A study by Zhan *et al.* [11], used voice, balance, gait, reaction time, and finger tapping data collected from 23 PD subjects, during scprited tests, by a smart phone over a period of 6 months to estimate a new disease severity score. Interestingly, Zhan *et al.* showed that gait features contributed most in the new mobile score and, the estimated score had high correlation (r=0.88) with the MDS-UPDRS-III. A study by Hssayeni *et al.* [12] used simulated activities of daily living measured in 24 PD subjects with sensors attached to the most affected wrist and ankle for estimating UPDRS-III. Models trained on signal-based extracted features showed a correlation between the estimated and real scroes of 0.62, with mean abosulte error of 7.5 points.

Based on previous research, it is clear that most of the studies had a low sample size (n=23-34) and mainly relied on feature engineering (i.e. signal-based features extracted from the wearables). In addition, trained models were not validated or tested on longitudinal data, but testing was based on k-fold cross-validation.

The overall objective of this work was to propose a DL CNN based model architecture that receives as an input raw

R.Z.U. Rehman, L. Rochester, A.J. Yarnall and S. Del Din are with Translation and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK (e-mail: rana.zia-ur-rehman, silvia.del-din@newcastle.ac.uk)

L. Rochester, and A.J. Yarnall are also with Newcastle upon Tyne Hospitals NHS Foundation Trust (e-mail: lynn.rochester, alison.yarnall@newcastle.ac.uk)

sensor data, rather than extracted features, in order to predict PD motor severity (i.e. MDS-UPDRS-III score). Specifically, we aim to: (i) apply DL techniques by training the model using time-point 1 (T1) gait data and MDS-UPDRS-III scores; (ii) validate the proposed approach on a longitudinal dataset: trained model will be tested on gait data collected after 3 years (T3) from T1, to predict MDS-UPDRS-III progression.

II. METHODS

A. Data Collection

119 PD participants were recruited between June 2009 and December 2011 from the Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-Gait study (ICICLE-GAIT), which was a collaborative study with ICICLE-PD study [13]. Ethical approval was obtained from the Newcastle and North Tyneside research ethics committee (REC reference: 09/H0906/82). All the experiments were conducted according to declaration of Helsinki. All PD subjects gave informed written consent before participating in the study.

Participants were assessed longitudinally every 18 months after baseline assessment up to 72 months. For this study, lab-based gait data collected with wearable sensors were included from the 36 month assessment onwards (at 36 (T1), 54 (T2) and 72 months (T3)). In this way, we included more severe PD subjects for the training, validation and testing of CNN, in order to provide a better generalization of our findings. Assessments were performed within one hour of dopaminergic medication intake. Participants clinical (e.g. MDS-UPDRS assessment) and demographic data were collected. For the gait assessment, subjects were asked to wear a tri-axial accelerometer (Axivity, AX3, dimensions: 23.0 × 32.5 × 7.6 mm, range: ±8g, sampling frequency: 100 Hz) attached on the

Fig. 1. Schematic diagram of the proposed CNN based model

lower back (L5) [5]. Participants were instructed to walk at their comfortable pace continuously for two minutes on a 25 m oval circuit. Only those who completed at least one minute of continuous walk were included in the analysis.

B. Data Pre-Processing

From the AX3 tri-axial (x (vertical), y (medio-lateral), and z (antero-posterior)) raw data, a combined signal magnitude vector was calculated from the three axis. In previous work, for motor fluctuation [14] and motor severity assessment [12], a 5-seconds (approximately 5-10 walking steps) non-overlapping sliding window had been used for feature extraction. Therefore, continuous signal magnitude vector was segmented into non-overlapping windows of 5-seconds (500 samples) that were inputted into the DL model. Training data comprised of 1748 signal magnitude windows (T1 data), while validation (T2 data) and test data (T3 data) of 1427 and 1179 windows, respectively.

C. Proposed Model Architecture

Previous work showed promising results by using CNN based deep neural networks based on raw movement signals for PD gait analysis [15]. A similar strategy was adopted in this study where we propose a new architecture of CNN for prediction of UPDRS-III, where the first bloc of convolutional layers is followed by a fully connected layers bloc (Fig. 1). Convolutional layer included four 1D-convolutional layers where each two convolutional layers were followed by the 1D max-pooling layer. This 1D convolutional bloc extracted deep features from each 5-second window (500, 1) of signal magnitude. Within each layer, a piecewise linear activation function (rectified linear activation function (ReLU)) was utilized [16]. In the second bloc, the model deep-learned the relationship between the spatial features extracted from the first bloc and the output regression (MDS-UPDRS-III predicted score) using two fully connected layers. The output layer included one neuron with ReLU activation to predict the MDS-UPDRS-III.

Separate experiments were designed to tune the model design and hyperparameters of the network. Hyperparameters considered for optimization were the filters/units in the convolution layer and fully connected layers; size of kernel in the convolutional layer; dropout rate; learning rate; batch size; and number of epochs. An adaptive learning rate optimization algorithm (Adam) was used to minimize the mean absolute error loss function [17]. Hyperparameters optimization was achieved by using the validation dataset which included 1427 segmented sequences of signal magnitude.

III. RESULTS AND DISCUSSION

Demographic and clinical data are reported in Table 1. At T1, the average age of PD participants was 70 years, height, mass, and BMI across three time points was comparable. Medication intake increased over time; however, their average motor severity (MDS-UPDRS-III score) ranged between 37-38 across the three time-points. Final testing of the optimized model was performed on the separate testing datasets, which included 1179 sequences of the segmented signal from 46 PD subjects. These 46 subjects' signals (average disease duration 6.5 years) were used to evaluate the performance of the model which was trained on 70 PD subjects (average disease duration 3.5 years). This section is divided into two parts: the first part
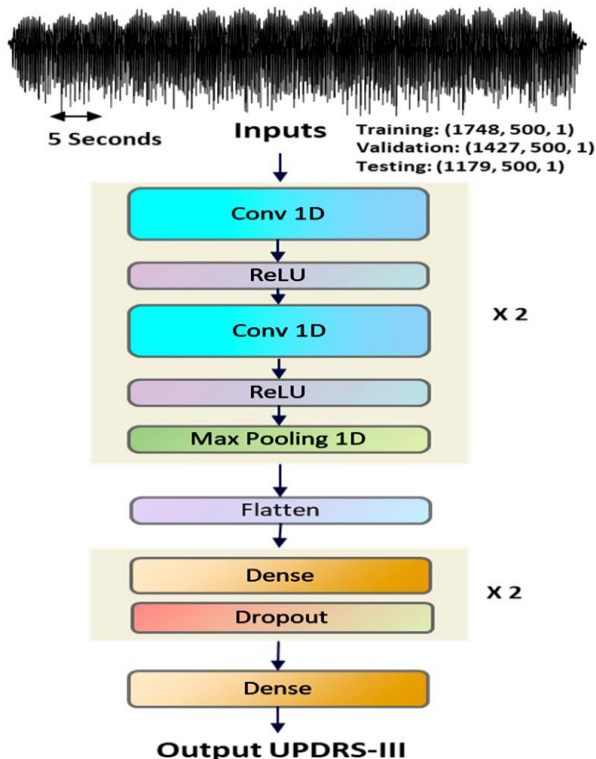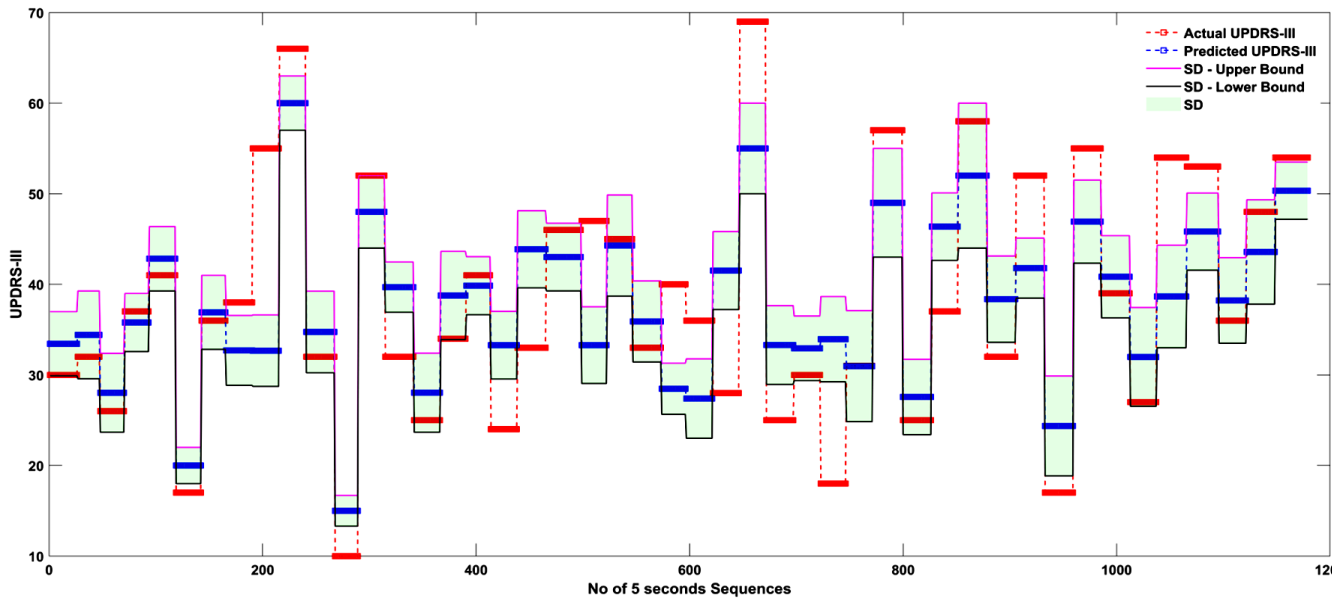
Fig 2. Predicted and actual UPDRS-III score (Y axis) for each sequence of segmented window testing data (after 3 years) for each subject

Table 1. DEMOGRAPHIC CHARACTERISTICS OF THE PD SUBJECTS, (LEDD: LEVODPA EQUIVALENT DAILY DOSE)

Demographics	Time Point 1 N = 70 PD	Time Point 2 N = 58 PD	Time Point 3 N = 46
Disease Duration (y)	3.54 ± 0.45	4.98 ± 0.43	6.48 ± 0.48
Sex (M/F)	46/24	38/20	32/14
Age (years)	70.04 ± 9.78	68.58 ± 9.36	69.98 ± 9.37
Height (m)	1.69 ± 0.09	1.69 ± 0.08	1.67 ± 0.09
Mass (kg)	77.62 ± 15.95	77.10 ± 16.04	77.47 ± 14.48
BMI (kg/m2)	27.21 ± 5.19	26.28 ± 5.87	27.51 ± 4.18
MDS UPDRS-III	37.56 ± 12.13	36.52 ± 12.69	38.11 ± 13.38
Hoehn &Yahr (HY) Stages	HY-I: 1 HY-II: 62 HY-III: 7	HY-I: 1 HY-II: 50 HY-III: 7	HY-II: 34 HY-III: 11 HY-IV: 1
LEDD (mg/day)	487.8 ± 243.3	633.2 ± 270.4	730.8 ± 337.8

highlights the CNN model tuning results (A), and the second part (B) presents the UPDRS-III prediction results.

A. Model Parameters Optimization and Training

For each first and second layer of the convolutional bloc, the filters/units selected were 8, and 16 respectively with grid search from [8,16,32,64,100,128,256]. The kernel size within these layers was 11, selected by grid search from [3,5,7,9,11]. The number of units selected by tuning for dense layers were 100. Learning rate hyperparameter of 0.001 was grid searched from [0.00001,0.0001,0.001,0.01, 0.05,0.08,0.1]. Dropout rate of 0.5 was selected by searching from [0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.0.8,0.9] for fully connected layers. Batch size of 8 was suitable based on the grid search from [8,16,32,64,128,256,512]. Epoch size selected was 50 by searching from 50 to 500 with increment of 50. However, early stopping of model training with patience of 10 was also applied in Keras framework to avoid overlearning of the model, when the model performance was not improving on the validation dataset.

B. Testing of the Proposed Model

The tuned trained model was used to predict the MDS-UPDRS-III score (T3) for each sequence within subjects for time-point 3 (46 PD subjects) as shown in Fig. 2. The

performance of the CNN model within subject fluctuated based on the 5-second signal magnitude, even though CNN was trained against one true value of MDS-UPDRS-III for each subject. Where true labelled MDS-UPDRS-III score (red color) was close to or within the range of upper and lower bound of one standard deviation (green color) from the average predicted MDS-UPDRS-III score (blue color). However, the average values of the predicted MDS-UPDRS-III scores within each subject were compared with the actual scores for further analysis.

Linear trend (correlation) between actual and predicted MDS-UPDRS-III is presented in Fig. 3. In terms of relative agreement, a strong correlation ($r = 0.82$, $p < 0.001$) between predicted and actual scores was observed. Absolute agreement was good (ICC (2,1) = 0.76, $p < 0.001$). When the model was tested on T3 data, the mean absolute error for the predicted scores was 6.29, which is clinically significant. The proposed
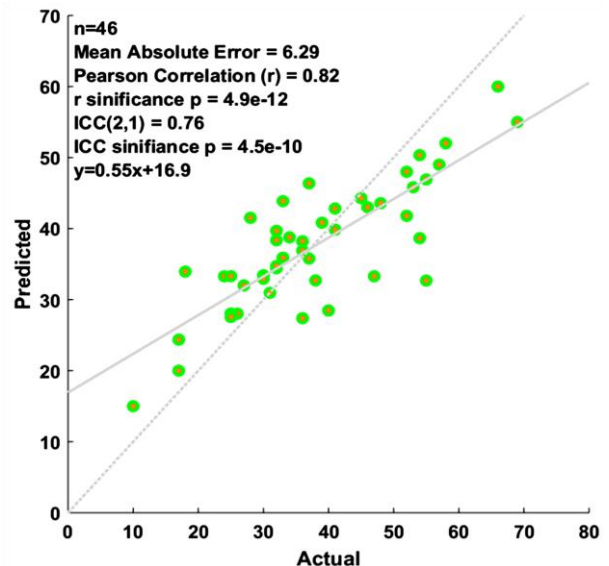

Fig. 3. Scatter plot showing the association between the predicted (Y axis) and actual UPDRS-III (X axis). 46 PDs are represented in green.
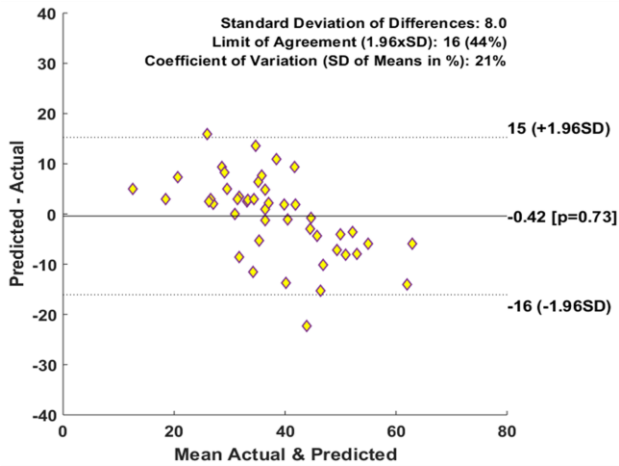
Fig. 4. Bland-Altman plot showing the absolute agreement and limits of agreement (dotted lines) between the predicted and actual UPDRS-III.

CNN-based model architecture, based on the use of raw signals, gave better performance compared to previous work based on feature engineering [10, 12].

Mean differences and limits of agreement between the predicted and actual UPDRS-III are presented in Fig. 4. A non-statistically (p = 0.73) significant bias of -0.42 was observed. A negative bias means that CNN slightly underestimated the UPDRS-III scores compared to actual scores. This could be due to the fact that the model was trained on data from PD subjects with less severe motor impairments (MDS-UPDRS-III: 37.56±12.13) compared to the patients in the testing dataset (UPDRS-III: 38.11±13.38). However, their difference is not clinically significant. Low constant bias (Fig. 4), strong correlation (Fig. 3), and good ICC (2,1) (Fig. 3) indicated that the model could reliably predict T3 UPDRS-III scores.

IV. Conclusion

The aim of this work was to propose a novel and objective DL-CNN model for quantifying PD motor severity (MDS-UPDRS-III score), in order to support clinicians for better management of PD patients. PD gait assessed longitudinally, with wearable technology, over a period of three years, was used to train and test the CNN based deep neural network architecture in order to estimate PD motor severity. The trained model tested on prospective data (after 3 years) gave reliable robust predicted MDS-UPDRS-III scores. Wearable sensor based gait data paired with deep learning techniques showed promising results for its adoption to support PD clinical management. Further work is required to extend this approach to real-world gait data for continuous monitoring of PD motor severity and compare it with traditional feature engineering based machine learning approaches.

References

[1] E. R. Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Neurology,* vol. 17, no. 11, pp. 939-953, 2018/11/01/ 2018.
[2] J. Wilson *et al.*, "Cholinergic Basal Forebrain Volumes Predict Gait Decline in Parkinson's Disease," *Movement Disorders,* 2020.
[3] J. Wilson *et al.*, "Gait Progression Over 6 Years in Parkinson's Disease: Effects of Age, Medication, and Pathology," *Frontiers in aging neuroscience,* 2020.
[4] S. Studenski *et al.*, "Gait speed and survival in older adults," *Jama,* vol. 305, no. 1, pp. 50-58, 2011.
[5] S. Del Din, A. Godfrey, and L. Rochester, "Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and Parkinson's disease: toward clinical and at home use," *IEEE journal of biomedical and health informatics,* vol. 20, no. 3, pp. 838-847, 2016.
[6] R. Z. U. Rehman *et al.*, "Turning Detection During Gait: Algorithm Validation and Influence of Sensor Location and Turning Characteristics in the Classification of Parkinson's Disease," *Sensors,* vol. 20, no. 18, p. 5377, 2020.
[7] R. Z. U. Rehman *et al.*, "Accelerometry-Based Digital Gait Characteristics for Classification of Parkinson's Disease: What Counts?," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 1, pp. 65-73, 2020.
[8] R. Z. U. Rehman *et al.*, "Comparison of Walking Protocols and Gait Assessment Systems for Machine Learning-Based Classification of Parkinson's Disease," *Sensors,* vol. 19, no. 24, p. 5363, 2019.
[9] R. Z. U. Rehman, S. Del Din, Y. Guan, A. J. Yarnall, J. Q. Shi, and L. Rochester, "Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson's Disease: A Comprehensive Machine Learning Approach," *Scientific Reports,* vol. 9, no. 1, p. 17269, 2019/11/21 2019.
[10] F. Parisi *et al.*, "Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and Gait tasks in Parkinson's disease," *IEEE journal of biomedical and health informatics,* vol. 19, no. 6, pp. 1777-1793, 2015.
[11] A. Zhan *et al.*, "Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score," *JAMA neurology,* vol. 75, no. 7, pp. 876-880, 2018.
[12] M. D. Hssayeni, J. Jimenez-Shahed, M. A. Burack, and B. Ghoraani, "Symptom-based, Dual-channel LSTM Network for The Estimation of Unified Parkinson's Disease Rating Scale III," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1-4: IEEE.
[13] A. J. Yarnall *et al.*, "Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study," *Neurology,* vol. 82, no. 4, pp. 308-316, 2014.
[14] S. Patel *et al.*, "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors," *IEEE transactions on information technology in biomedicine,* vol. 13, no. 6, pp. 864-873, 2009.
[15] M. Gadaleta, G. Cisotto, M. Rossi, R. Z. U. Rehman, L. Rochester, and S. Del Din, "Deep learning techniques for improving digital gait segmentation," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 1834-1837: IEEE.
[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.