Enhancing Model Generalizability In Parkinson's Disease Automatic Assessment: A Semi-Supervised Approach Across Independent Experiments

1st Gianluca Amprimo
Politecnico di Torino
Turin, Italy
gianluca.amprimo@polito.it

2nd Giulia Masi
Politecnico di Torino
Turin, Italy
giulia.masi@polito.it

3rd Gabriella Olmo
Politecnico di Torino
Turin, Italy
gabriella.olmo@polito.it

4th Claudia Ferraris
CNR-IEIIT Italy
Turin, Italy
claudia.ferraris@cnr.it

***Abstract*—Machine learning in Parkinson's disease assessment uses data from clinically-coded movements, such as finger tapping, to objectively measure motor impairment. Video-based models showed promise in several experiments, but the lack of a unified test benchmark hinders proving generalizability. Additionally, new telemedicine systems may easily collect large amounts of unsupervised data, while obtaining ground truth labels for supervised learning remains time-consuming and requires specialized clinicians. This study explores semi-supervised learning to enhance the generalizability of a Light Gradient Boosting model for video-based finger tapping staging, while reducing its need for supervised data labelling. Specifically, this work employs the Self-training schema in two trials using openly-available finger tapping datasets from three independent experiments. This method significantly improves model performance across various metrics, achieving notable accuracy gains (e.g., from 87.62% to 92.05%) when tested on unseen data from a different experiment. Semi-supervision proves valuable when limited labelled data (less than 10%) from the test distribution are available during training.**

***Index Terms*—Semi-supervised Learning, Finger Tapping, Parkinson's Disease, Machine Learning, Health Informatics**

I. Introduction

Parkinson's disease (PD) is the second most prominent neurodegenerative disease worldwide [1]. The clinical assessment of PD-related motor impairment, crucial for refining pharmacological therapy, relies on the evaluation of the motor tasks outlined in Section III of the Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [2]. This assessment takes place during a neurological examination conducted by well-trained clinical personnel, who assign a MDS-UPDRS score from 0 (no impairment) to 4 (severe impairment) to each task performed by the patient. Even if standardised, this assessment is susceptible to intra- and inter-rater variability [3].

The advent of Machine Learning (ML) has sparked the interest in developing automatic and objective assessment models, able to estimate MDS-UPDRS scores from simple video recordings [4]. The popularity of this approach stems from the ease of video recording clinical tasks either during inpatient visits or directly *at-home* through telemedicine systems [5] [6]. Following collection, videos undergo automatic analysis using deep-learning-based body tracking frameworks (e.g., MediaPipe [7]) to extract raw kinematic data and derive handcrafted features for training the scoring model.

Finger Tapping (FT) (i.e., the repetitive tapping of thumb and index finger) is among the most studied motor tasks for this purpose [8], since PD progression affects fine motor control. Alterations in FT may manifest as bradykinesia (i.e., slowness or halt during executions) as well as irregular tapping frequency. While numerous studies have proposed solutions for video-based MDS-UPDRS score estimation [9]–[13], the lack of open benchmark datasets hinders the complete validation of model generalizability across independent experiments with similar endeavours. Moreover, privacy regulations pose significant limitations on video data sharing. This constrains the size of trainable learning models, with most solutions relying on shallow models to classify small self-collected datasets [14] [15]. Furthermore, proposed models are supervised and require ground-truth scoring by clinicians. Such labelling process is time-consuming, and requires two or more well-trained raters to reduce inter-rater variability. This aspect represents a substantial limitation, especially when considering the speed that last-generation telemedicine systems could offer to collect large amounts of data directly at the patient's home.

To the best of the authors' knowledge, this work is the first to propose semi-supervised learning in the specific domain of video-based MDS-UPDRS score estimation for the FT task. Semi-supervision exploits unlabelled and labelled data during training to improve ML model performance and generalization capabilities, reducing the need for extensive data labelling [16]. In particular, this work applied the *Self-training* paradigm to the state-of-the-art Light Gradient Boosting (LGB) model proposed in [10], combining their data with an unsupervised original dataset. The aim was to verify if this methodology could enhance the model performance when testing on an open-access benchmark dataset from an independent but similar experiment [13].

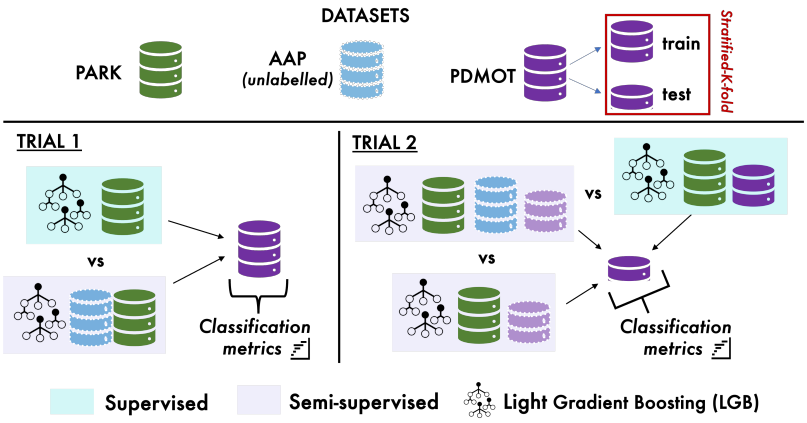

Fig. 1. Experiment summary: Trial 1 compares supervision and semi-supervision by training LGB models on *PARK* and *AAP* then testing over *PDMOT*; Trial 2 investigates the performance of semi-supervision vs supervision when training data are sampled from the test distribution as well using *Stratified-K-fold.*

II. Materials & Methods

A. FT Video Datasets

Most of the FT video datasets in the literature include few data samples and subjects [14] [15]. Some larger studies exist [9] [15], but their data are not open-access. This work employed two openly-available labelled datasets from the literature, (*PARK* and *PDMOT*), as well as an unlabelled dataset without associated clinical scores (*AAP*).

PARK dataset [10] consists of video recordings from 250 subjects (PD subjects: 172, healthy controls: 78), collected worldwide through the *Parktest* web application. Raw video recordings are private, but the authors shared the features extracted from 489 samples and their MDS-UPDRS scores, as assigned by three expert raters. Moreover, the code to reproduce their feature extraction pipeline and their trained optimal LGB model are open-access.

PDMOT dataset [13] consists of 611 FT videos collected from 368 PD patients at Tiantan Hospital (Beijing, China) and scored by three expert raters. The dataset is the first open-access benchmark for FT and provides anonymized video recordings of the task. However, the low quality and the corruption of some videos due to automatic face anonymization reduce the number of valid samples to 543. The authors did not share the code of the optimal classifier they trained on these data, and such model is not perfectly reproducible due to the corrupted original samples.

AAP dataset results from a collaboration between the authors of this work and the Associazione Amici Parkinsoniani Piemonte Onlus (Turin, Italy). It includes 338 FT videos provided voluntarily by 40 PD subjects. MDS-UPDRS scores from experts are currently missing, therefore only the diagnosis of Parkinson's disease is known for these subjects. Raw videos will be released shortly as part of a larger dataset, but extracted numerical features are available as a CSV file alongside the code to reproduce this work at https://rb.gy/f9s4oi.

Given the absence of source videos in the *PARK* dataset, the feature extraction pipeline proposed alongside this dataset was applied to achieve the same numerical features for *PDMOT* and *AAP* as well. Such a pipeline consists of running MediaPipe [7] for extracting hand joints, followed by computing and segmenting the trajectory over time of the angle between the thumb and index finger tips with respect to the wrist. Established features from the literature (e.g., tapping speed and acceleration, movement freezing events, aperiodicity) are then derived, and only the 22 most relevant are retained. Additional details can be found in [10].

B. Semi-supervised learning

Semi-supervised learning involves training models on datasets encompassing a combination of labelled and unlabelled instances, since these latter are usually more abundant and overlooked during supervised learning [16]. Indeed, incorporating such data may augment the model's ability to discern underlying patterns and relationships while reducing the need for human supervision in the training process. The basic idea consists in assigning confident *pseudo-labels* to the unlabelled samples, which can then contribute to the traditional supervised training of the model. The literature encompasses several semi-supervised techniques suitable both for shallow and deep learning models [17].

This work exploited the *Self-training* paradigm [18]: a model is initially trained on a small set of labelled data and used to predict *pseudo-labels* for the unlabelled data points. The samples with a confident predictions then contribute to the retraining of the supervised model, together with the other labelled instances. This process iterates, gradually expanding the labelled dataset and hopefully improving the model's performance. Confident predictions are either the *K-best* for each iteration or those crossing a predefined confidence threshold. Furthermore, a maximum number of iterations guarantees the end of the process even when the algorithm does not converge. The approach performs well if the initial supervised model is robust enough to perform some correct and confident predictions on the unlabelled data. Moreover, the choice of the hyperparameters (i.e., the confidence criterion, the maximum

number of iterations) is crucial to avoid the potential accumulation of labelling errors during the iterative process.

C. Experiment overview

The experiment, depicted in Figure 1, comprised two trials. In Trial 1, the optimal LGB model trained via supervised learning on *PARK* in [10] served as the baseline. A second LGB model used Self-training to learn from both *PARK* and *AAP* (*PARK+AAP*), while *PDMOT* served as the test benchmark. The aim was to evaluate how unlabelled samples from a comparable yet independent experiment could improve the predictive performance of an optimal model when confronting with a previously unseen benchmark.

In Trial 2, *PDMOT* was split into train ($PDMOT_{trn}$) and test ($PDMOT_{tst}$) subsets using *Stratified-K-fold* validation, with $K \in [2, 10]$ to assess different train and test sizes. Splitting was repeated five times for each fold to guarantee independence from random sampling. The main objective was to compare the efficacy of unlabelled data sampled from the same distribution as the testing benchmark versus unlabelled samples from a similar but different experiment. Additionally, the trial sought to establish a threshold where abundant unlabelled data proved equivalent or more useful than labelled data from the test distribution. Therefore, a supervised LGB model was trained on $PDMOT_{trn}$+*PARK*. Expectations were that this baseline, leveraging labelled training data from the same experiment as the test data, would outperform the lower bound baseline and potentially represent the best-performing model of the whole experiment (upper bound). Alongside this, two additional semi-supervised LGB models were learned on *PARK+AAP*+$\overline{PDMOT}_{trn}$ and *PARK*+$\overline{PDMOT}_{trn}$. The overline notation for $\overline{PDMOT}_{trn}$ means that samples were used as if they were unlabelled during semi-supervision. All models tested their performance on $PDMOT_{tst}$.
Optimal hyperparameters for Self-training in both trials were determined through an extensive grid-search procedure, selecting combinations minimizing Mean Absolute Error (MAE).

D. Evaluation strategy

Model performance assessment uses the same core metrics of [10], namely MAE and Pearson's Correlation Coefficient (PCC). Moreover, acceptable accuracy (i.e., accuracy within a ± 1 prediction error) is included, since commonly used in FT studies to support performance claims [9] [15]. Indeed, its error tolerance accounts for the potential disagreement among expert raters. Furthermore, results for Trial 1 include the confusion matrices of the predictions on *PDMOT* from the two trained models to discern per-score differences. For Trial 2, testing metrics were evaluated across all iterations of the *Stratified-K-fold* cross-validation for a given K value, ensuring unbiased results across different train-test splits.

III. Results & Discussion

For the sake of brevity, trained models are discussed in this section by referring to them through the name of their training data. In both trials, the optimal Self-training consisted

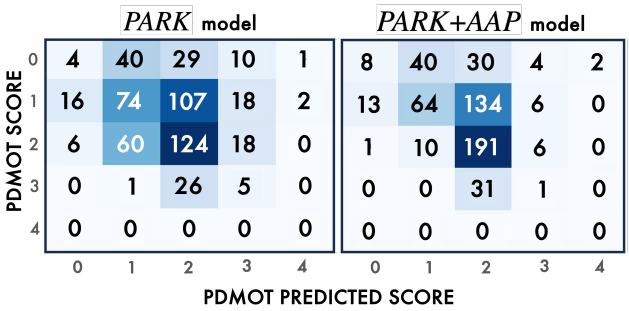

Fig. 2. Test predictions of MDS-UPDRS scores on *PDMOT*: left, predictions by supervised *PARK* model; right, predictions by semi-supervised *PARK+AAP* model. MDS-UPDRS measures severity in a range from 0 (low) to 4 (high).

in adding the $K = 10$ most confident *pseudo-labels* for each iteration. The maximum number of iterations was set equal to the cardinality of the unlabelled data divided by K. Such an outcome indicates that the model effectively learnt by slowly using small batches of confident *pseudo-labels* to refine its predictions over the iterations of the algorithm.

A. Trial 1

When testing on *PDMOT*, *PARK* model achieved MAE=0.77, PCC=0.15, and 87.62% acceptable accuracy. This performance, as expected, denotes a lack of model generalizability. Indeed, these results are much worse than those achieved in [10], where the model scored MAE=0.58 and PCC=0.63 in a *Leave-One-Subject-Out* validation on *PARK*-only data. In contrast, the semi-supervised *PARK+AAP* model attained MAE=0.61, PCC=0.35, and 92.05% acceptable accuracy on *PDMOT*, which means approximately a 20%, 133%, and 5% improvement respectively in each metric. This result demonstrates the efficacy of Self-training to enhance the generalizability of the model, by exploiting the inherent information contained in *AAP* unlabelled data. Moreover, acceptable accuracy for *PARK+AAP* model is comparable to that found in [9] (81%) and [15] (98.3%) for their test data. In Figure 2, the confusion matrices illustrate that Self-training reduces the number of scoring errors larger than one point and improves recognition of scores 0 and 2, albeit with a slight decline in scores 1 and 3. This outcome explains the overall improvement observed, especially on acceptable accuracy. Results cannot be derived for score 4, which is not present in *PDMOT* and also underrepresented in *PARK*.

B. Trial 2

Figure 3 presents line plots showing the trend of PCC (mean value and confidence range in the repeated *Stratified-K-Fold* splits) as the cardinality of $PDMOT_{trn}$ grows. This value is expressed as percentage of *PDMOT* size. Acceptable accuracy and MAE show similar plots, not reported for brevity. Notably, when only 10% of labelled training data were sampled from the test distribution, semi-supervision performed as well as supervision (red rectangle in the figure). In this scenario, 10% implies 50 labelled samples versus the 338 unlabelled

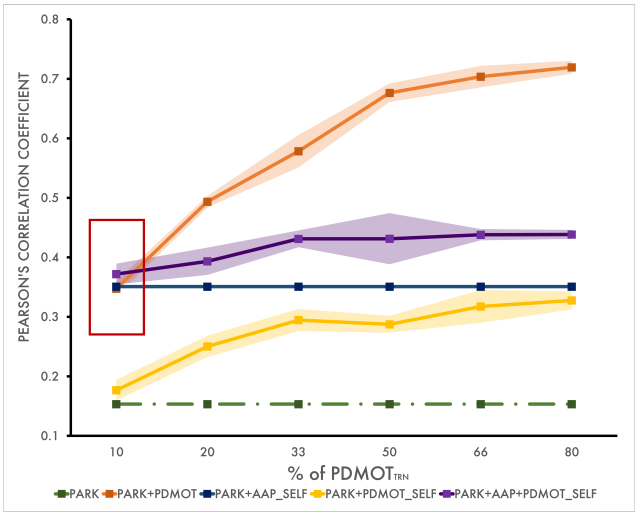

Fig. 3. Line plots for PCC when testing supervised and semi-supervised models of Trial 1 and Trial 2, as the dimension of $PDMOT_{trn}$ grows.

samples from *AAP* (1:7 proportion). Comparing the semi-supervised model, *PARK+AAP+*$\overline{PDMOT}_{trn}$ (violet line) performed better than the *PARK+AAP* model from Trial 1 (blue line). Indeed, the former considers a larger pool of unlabelled data, including data sampled from the same experiment of the test data. Its performance reached an asymptotic behaviour when $PDMOT_{trn}$ hit a dimension above 50% (around 270 in $PDMOT_{trn}$, total: 608 unlabelled samples). The same behaviour was observed in the supervised $PDMOT_{trn}$*+PARK* model (dotted orange line). This outcome may likely depend on the progressive reduction of $PDMOT_{tst}$ size. The semi-supervised *PARK+*$\overline{PDMOT}_{trn}$ (yellow line) model exhibited improved performance as $PDMOT_{trn}$ increased size. When its cardinality was similar to *AAP* dataset (i.e., 66% of *PDMOT*, 362 samples), it achieved results comparable to the *PARK+AAP* model. Therefore, unlabelled data from both experiments appear to share similar informative power. Overall, all supervised models significantly enhanced classification compared to the supervised *PARK* baseline of Trial 1 (dotted green line).

C. Limitations

This work is not without limitations. First, semi-supervision usually exploits a pool of unlabelled data more numerous than the labelled one. This condition, however, was met only in Trial 2, for *PARK+AAP+*$\overline{PDMOT}_{trn}$ model, when more than 33% of data from *PDMOT* were used for training. This aspect may limit the observable performance improvement in comparison to the supervised $PDMOT_{trn}$*+PARK* model. However, even in such an unfavourable scenario, a significant improvement was observed for all metrics. In addition, the use of *Stratified-K-fold* assumes that training data will have the same distribution as test data, which may not happen in a real scenario. This limitation may translate in partially over-optimistic results for semi-supervised models trained on *PARK+AAP+*$\overline{PDMOT}_{trn}$ and *PARK+*$\overline{PDMOT}_{trn}$.

IV. Conclusions

This work investigated whether semi-supervised learning could improve the automatic staging of Parkinson's disease from finger tapping videos. This approach may reduce the need for extensive data labelling by expert clinical raters and the bias introduced by using uncertain clinical scores as ground truth, exploiting intrinsic information in the unlabelled samples. Results obtained using data from three independent experiments suggest that this approach is feasible and beneficial to model generalizability. Future works should further explore semi-supervised learning, possibly exploiting deep-learning-based methods. Such an endeavour, however, will be possible only if new and larger open-access finger-tapping datasets will gradually become available to the research community.

References

[1] D. K. Simon *et al.*, "Parkinson disease epidemiology, pathology, genetics, and pathophysiology," *Clinics in geriatric medicine*, vol. 36, no. 1, pp. 1–12.

[2] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric assessment," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, 2008.

[3] B. Post *et al.*, "Unified parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?," *Movement disorders: official journal of the Movement Disorder Society*, vol. 20, no. 12, pp. 1577–1584, 2005.

[4] K. Sibley *et al.*, "Video-based analyses of parkinson's disease severity: A brief review," *Journal of Parkinson's disease*, vol. 11, no. s1, pp. S83–S93, 2021.

[5] K. Sibley *et al.*, "An evaluation of kelvin, an artificial intelligence platform, as an objective assessment of the mds updrs part iii," *Journal of Parkinson's Disease*, no. Preprint, pp. 1–12, 2022.

[6] G. Amprimo *et al.*, "Assessment tasks and virtual exergames for remote monitoring of parkinson's disease: An integrated approach based on azure kinect," *Sensors*, vol. 22, no. 21, p. 8173, 2022.

[7] F. Zhang *et al.*, "MediaPipe hands: On-device real-time hand tracking," 2020.

[8] J. Mei *et al.*, "Machine learning for the diagnosis of parkinson's disease: a review of literature," *Frontiers in aging neuroscience*, vol. 13, p. 633752, 2021.

[9] G. Morinan *et al.*, "Computer vision quantification of whole-body parkinsonian bradykinesia using a large multi-site population," *NPJ Parkinsons Dis.*, vol. 9, no. 1, p. 10, 2023.

[10] M. S. Islam *et al.*, "Using ai to measure parkinson's disease severity at home," *NPJ digital medicine*, vol. 6, p. 156, 08 2023.

[11] Z. Li *et al.*, "An automatic evaluation method for parkinson's dyskinesia using finger tapping video for small samples," *J. Med. Biol. Eng.*, vol. 42, no. 3, pp. 351–363, 2022.

[12] M. H. Monje *et al.*, "Remote evaluation of parkinson's disease using a conventional webcam and artificial intelligence," *Frontiers in Neurology*, vol. 12, p. 742654, 2021.

[13] N. Yang *et al.*, "Automatic detection pipeline for accessing the motor severity of parkinson's disease in finger tapping and postural stability," *IEEE Access*, vol. 10, pp. 66961–66973, 2022.

[14] G. Vignoud *et al.*, "Video-based automated assessment of movement parameters consistent with mds-updrs iii in parkinson's disease," *Journal of Parkinson's Disease*, no. Preprint, pp. 1–12, 2022.

[15] H. Li *et al.*, "Automated assessment of parkinsonian finger-tapping tests through a vision-based fine-grained classification model," *Neurocomputing*, vol. 441, pp. 260–271, 2021.

[16] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[17] X. Yang *et al.*, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[18] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *33rd annual meeting of the association for computational linguistics*, pp. 189–196, 1995.