Comparing laboratory and in-the-wild data for continuous Parkinson's Disease tremor detection

Ada Zhang, Fernando De la Torre, and Jessica Hodgins
Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

***Abstract*— Passive, continuous monitoring of Parkinson's Disease (PD) symptoms in the wild (*i.e.*, in home environments) could improve disease management, thereby improving a patient's quality of life. We envision a system that uses machine learning to automatically detect PD symptoms from accelerometer data collected in the wild. Building such systems, however, is challenging because it is difficult to obtain labels of symptom occurrences in the wild. Many researchers therefore train machine learning algorithms on laboratory data with the assumption that findings will translate to the wild. This paper assesses how well laboratory data represents wild data by comparing PD symptom (tremor) detection performance of three models on both lab and wild data. Findings indicate that, for this application, laboratory data is not a good representation of wild data. Results also show that training on wild data, even though labels are less precise, leads to better performance on wild data than training on accurate labels from laboratory data.**

***Clinical relevance*— Results in this paper suggest that, when building a system for in-the-wild PD symptom detection, it is better to train machine learning algorithms on data from the wild as opposed to from the lab, even though wild labels are less precise. This paper also presents a newly released dataset for PD tremor detection in lab and wild environments.**

I. Introduction

An estimated 10 million people worldwide live with Parkinson's Disease [1], a chronic, neurodegenerative disorder that leads to both non-motor and motor symptoms. These symptoms include, but are not limited to, depression, anxiety, sleep disorders, slowness, muscle rigidity, postural instability, and tremors. While there is no cure, medication can provide symptomatic relief. However, dosages need to be adjusted as a patient's disease progresses and symptoms worsen. We believe that continuous PD motor symptom monitoring would enable clinicians to better adjust medication and thereby improve their patients' quality of life. Our eventual goal is to build a system for continuous monitoring of PD motor symptoms "*in the wild*" – *i.e.*, in natural environments without requiring any specific interaction from the patient. Such a system would use machine learning algorithms to detect symptoms in data collected from wearable accelerometers (see Fig. 1).

Machine learning algorithms for symptom detection typically require accurate labels: *i.e.*, the start and end of each symptom. Labels are time consuming to annotate and the exact onset of a symptom can be subjective. These issues are compounded in wild settings. Therefore, researchers often use data collected in laboratory settings for training. Fewer researchers have explored the use of wild data for training because labels for these data are typically supplied by PD patients via paper diaries [2], [3]. Therefore, only *approximate* (typically ± 1 hour) timestamps of symptom occurrences are available. That is, these data are *weakly* labeled.

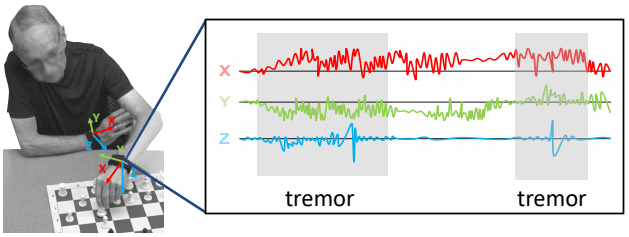

Fig. 1. Depiction of automated tremor detection with wearable sensors during everyday activities

In this paper, we assess how well laboratory data represents wild data by comparing PD symptom detection performance of three models on laboratory versus wild data. Similar performance across datasets would imply that findings on laboratory data should transfer to the wild. Results from each of the models, however, show that laboratory data is not representative of wild data. Of the three models, we find that the one trained on weakly labeled wild data has better performance on wild data than the ones trained on accurately labeled laboratory data. While this paper focuses on upper-limb PD tremor, we expect results to translate to other PD symptoms. The findings in this paper may also generalize to other problems in human activity understanding, such as monitoring of other motor impairments, activity tracking, or sports performance analysis.

II. Related Work

With the advent of wearable sensors and recent advances in machine learning, there has been a preponderance of interest in automated and objective analysis of PD. Many researchers have attempted to diagnose PD, evaluate symptom severity, or detect symptoms. Indeed, the use of wearable devices for PD monitoring and assessment has been well-studied, and we refer the reader to the following review papers for more information: [4]–[7]. Many studies, however, are conducted within laboratory settings, which may not accurately reflect a patient's at-home disease state. Some researchers have explored systems for automated at-home disease analysis [8]–[15]. These studies, all within the past five years, reflect the great potential of smartphones and other portable devices to facilitate in-home monitoring by enabling motor tests that can be performed in-home and analyzed automatically. However, each of these studies requires patients

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

to perform specific actions for analysis throughout the day, which can be burdensome and lead to noncompliance.

We believe the future in PD monitoring is through passive, continuous monitoring "*in the wild*," where the system automatically detects symptoms through a smartphone or smartwatch without requiring any specific interaction from the user. Several studies have collected this form of naturalistic data [13], [15]–[21]. However, none of these data collections asked participants to label when they experienced symptoms. Therefore, none attempt to explicitly detect symptoms.

Some researchers have asked study participants to label the approximate time their symptoms occur in wild settings. These approximate labels are called *weak labels*. Fisher et al. [2] used a neural network to detect dyskinesia in weakly labeled data collected in the homes of subjects with PD, but treated the data as accurately labeled before training the network. Das et al. [3] compared several weakly supervised learning techniques on in-home data collected from two subjects. In a follow-up paper, Zhang et al. [22] compared several weakly supervised learning algorithms, including a novel "stratified" algorithm, on a larger dataset collected in a laboratory setting. In this paper, we extend previous work and assess how well laboratory data represents data collected in the wild, and we compare the performance of algorithms trained on these two types of data.

III. Methods

Here, we describe our data collection procedures, our data processing and feature extraction methods, our algorithms and data partitioning protocols, and our performance metrics.

A. Data Collection

We collected data both in laboratory and wild settings[1]. More details about the data collection protocol are given in [23]. This research was approved by the Carnegie Mellon University Institutional Review Board.

1) Laboratory recordings (LAB): Data were collected from 12 subjects (eight male, four female, ages 66 to 85) who had been diagnosed with PD two to five years prior. Each subject self-reported tremor in one or both hands. The subjects wore one Axivity AX3 accelerometer on each wrist while they completed several actions, some of which were taken from the UPDRS and others from daily living (playing cards, making/eating a sandwich, *e.g.*). Data were collected at 100 Hz. Three cameras were used to record the subjects so as to minimize occlusion (see Fig. 2 for views from the three cameras). These video data were used to annotate tremor events, thereby providing ground truth data. Table I shows a summary of the collected, labeled data. Note that subjects 1 and 6 did not exhibit any tremor during the data collection and were thus excluded from this study. These data are subsequently referred to as *LAB*.

2) In-the-wild recordings (WILD): Subjects 2, 4, 5, 10, 11 and 12 agreed to participate in an in-home study, which involved wearing two Axivity AX3 accelerometers

[1]A link to this dataset is available at http://www.humansensing.cs.cmu.edu/software.

Fig. 2. Experimental setup

TABLE I
SUMMARY OF DATA COLLECTED IN LAB

Subject	# Labeled minutes per hand	% Tremor events	
		Left hand	Right hand
2	74.9	80.0	40.6
3	55.9	55.9	73.7
4	55.2	57.1	37.1
5	88.1	39.0	44.1
7	97.1	26.9	19.3
8	91.3	8.6	37.9
9	96.1	21.9	7.6
10	84.5	7.8	11.7
11	51.5	69.2	26.3
12	74.7	2.0	1.2
Total	**769.2 (∼12.8 hours)**		

throughout the day for two to four weeks. Labels of tremor were provided by the subjects via a cell phone app, which was designed to prevent subjects from submitting many entries within a short time span or backdating entries. The app prompted subjects to submit an entry roughly every hour and subjects were paid per entry. To improve label accuracy, subjects were only asked to record the amount of tremor they experienced within the 5 minutes prior to submitting the entry. Following the recommendation given in previous work [22], we used stratified rather than binary weak labels. That is, rather than asking subjects whether they experienced or did not experience tremor within the previous five minutes, we instead provided three label options (*Almost none*, *Half the time*, and *Almost always*). We chose to use three options, a slight deviation from the four strata used by Zhang et al. [22], because we felt that subjects would be able to more accurately select from a smaller set of options. All subjects made regular entries during the data collection period. Fig. 3 shows the labels provided by the subjects over time. Note that subject 12 only participated for two weeks while all other subjects participated for four weeks. While these labels may be subject to the biases inherent in self-reporting, they are no less accurate than paper diaries, which are the current gold-standard for obtaining wild labels. Furthermore, the time stamps available from the cell phone app are much more precise than what could be gleaned from paper diaries. These stratified weakly labeled data are subsequently referred to as *WILD*.

B. Features

Previous work on automated tremor detection [24]–[28] have generally used very similar features. In this work, we

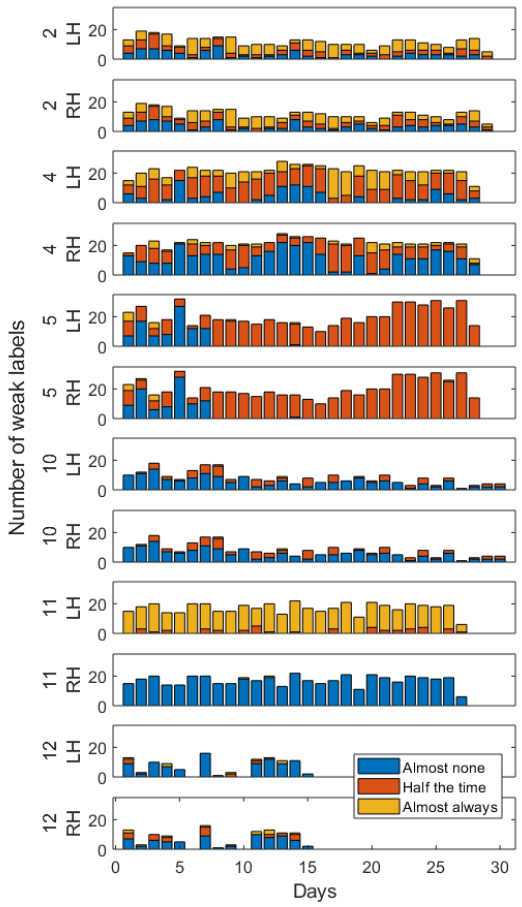

Fig. 3. Labels submitted by each subject over time

used the same features as those described by Patel et al. [28]. We first high-pass filtered the accelerometer signals with a 1 Hz cutoff using an 8th-order elliptic filter. Features were computed over three-second windows of the signal with one-second overlap. Each window was labeled as tremor if at least half of the window was tremor. On each window, the range, root mean square, Shannon entropy, dominant frequency and ratio of energy in the dominant frequency over total energy were computed over each axis. The peak normalized cross-correlation and the time lag associated with the peak were computed over each pair of axes. Our feature vector for each window was therefore 21-dimensional.

C. Algorithms and model selection

The purpose of this paper is to explore how well LAB data represents WILD data and to understand what kinds of training datasets would be most effective for an automated, in-home, tremor monitoring system. Given the small size of our dataset, we chose to use linear Support Vector Machine (SVMs) for all experiments. In particular, we trained SVMs on three different partitions of our dataset (described below). Training data were normalized to have a mean of zero and a standard deviation of one. (Note that different algorithms used different training data sets, therefore normalization parameters differed across the various algorithms.) The SVMs were implemented using the LIBLINEAR library provided by Fan et al. [29]. In all cases, model selection for the hyperparameter C was chosen through 3-fold cross validation. That is, the training dataset was partitioned into three folds, two of which were used for training and one for validation. Thirteen models were trained (one for each C in the range of 2^{-13} to 2^{-1} in powers of 2) on the training folds and evaluated on the validation fold. This procedure was repeated for each of the three possible permutations of selecting training and validation sets from the three folds. The results were averaged across the three trials and the highest performing C was then used to train a new model on the entirety of the training data set (all three folds). Performance of this model was then evaluated on the test dataset, which was distinct from the dataset used for training.

Below we describe the three different partitionings of our dataset for training. A graphical representation is shown in Fig. 4. Note that, because people use their left and right hands in very different ways, data from separate hands were considered to be from separate subjects.

1) Generic SVM from LAB data (Gen-LAB): We evaluated a standard, binary SVM on LAB data using leave-one-subject-out cross validation: *i.e.* training on all subjects excluding one, and testing on that left out subject. For model selection, the three folds were chosen such that each fold contained data from one third of the training subjects. To select C, we chose that which led to the highest average Area Under the Curve (AUC) value across the three possible validation sets. The final model was then tested on the test subject's LAB and WILD data. This experimental protocol represents a typical machine learning pipeline and is subsequently referred to as *Gen-LAB*. See Fig. 4(a,b).

2) Person-specific SVM from LAB data (PS-LAB): We trained a standard, binary SVM on LAB data from the test subject. The LAB dataset was first partitioned into three folds, two of which were used for training/validation and one for testing. Model selection was performed on the training/validation portion. The learned SVM was then tested on the test partition of the LAB data and the entire WILD dataset. Results were then averaged across all three learned SVMs, corresponding to three permutations of selecting the training/validation and test sets from the three folds of the LAB data. Following the recommendation of [30], all folds were chosen to be temporally connected segments, rather than selected from randomized samples. It has been shown that training a person-specific classifier leads to improved performance over a generic classifier [31], and we compared performance on LAB versus WILD data for such a classifier, which is subsequently referred to as *PS-LAB*. See Fig. 4(c,d).

3) Person-specific SVM from WILD data (PS-WILD): Using the WILD data from the test subject, we trained a stratified, Multiple Instance SVM (MI-SVM), as was used by Zhang et al. in [22]. We assigned approximate tremor percentages of [0-33%], [33-66%], and [66-100%] to the labels *Almost none*, *Half the time*, and *Almost always*, respectively. Similarly to the methodology for *PS-LAB*, the WILD dataset was first partitioned into three folds (two for training/validation and one for testing). The learned SVMs

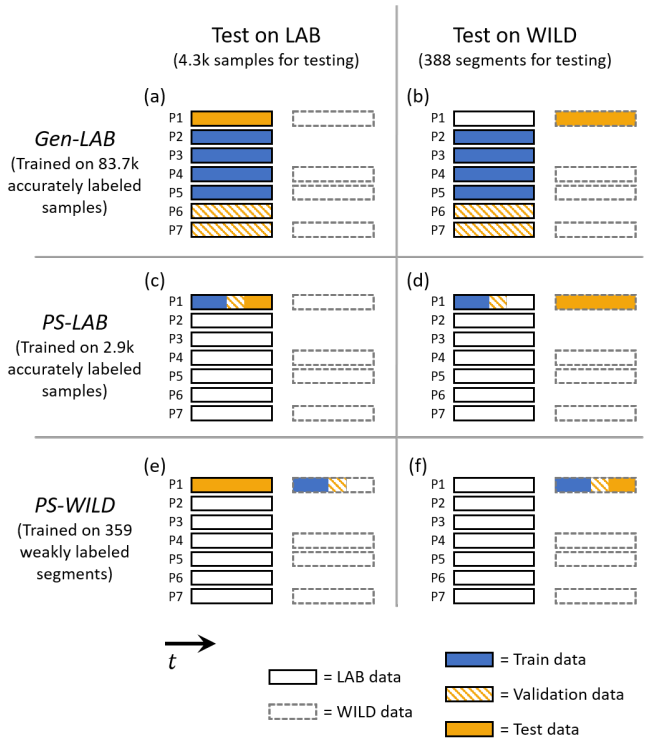

Fig. 4. Schematic representation of training, validation, and test datasets for all experiments in which P1 is the test subject. Models were trained for each given test subject. Note that, because training and test sets differ for a given test subject, the reported number of samples/segments are average values across all subjects. Cross validation (averaging validation results across all folds) was used to do model selection. For *PS-LAB* and *PS-WILD*, the subject's LAB or WILD data were split into three folds, two of which were used for training/validation. Three models were trained, one for each permutation of selecting folds for training and testing. Results were averaged across all three models.

were tested on the test partition of the WILD dataset and the entirety of the LAB dataset. Results were averaged across all three permutations. As with *PS-LAB*, folds were chosen to be temporally connected. As shown in Fig. 3, labeling frequency was relatively consistent across the data collection period. Therefore, these partitions correspond to similar numbers of days. Because accurate labels are not available for WILD data, we could not use AUC for selecting C. Instead, we chose to use mean absolute error of the detected percentage (described below in Sec. III-D) as the performance metric for model selection. Note that, for some partitions, training data would lack segments with "*Almost none*" or "*Almost always*" labels, making it not possible to initialize the stratified MI-SVM algorithm. Such partitions were ignored during model selection.

The purpose of training on WILD data is to explore the relative benefits of accurate labels from other subjects versus weak labels from the test subject. This classifier is subsequently referred to as *PS-WILD*. See Fig. 4(e,f).

D. Performance metrics

All learned models (with no retraining) were tested on both LAB and WILD data, and were compared across three performance metrics: (1) accuracy, (2) area under the (Receiver Operating Characteristic) curve (AUC), and (3) mean absolute error in detected percentage (MAE). In this paper, we define MAE as follows: For each segment, the percentage of tremor within was computed using the trained model. If the detected percentage was within the associated range for the given label, the error on that segment was set to be zero. Otherwise, it was set to be the absolute difference between the detected percentage and the closest bound. For example, if a segment's label was *Almost always* (associated range of [66-100%]) and the detected percentage of tremor in that segment was 50%, then the absolute error on that segment would be 16%. The absolute error was then averaged across all segments to get the mean absolute error. Note that, because accurate labels (exact time points of tremor events) are not available for WILD data, we could not compute accuracy and AUC on WILD data. To compute mean absolute error in detected percentage on LAB data, we converted accurate labels into weak labels: LAB data were broken into 5-minute segments (no overlap), the percentage of tremor within each segment was computed, and the segment was assigned a label of *Almost none*, *Half the time*, or *Almost always* if the percentage of tremor was within [0-33%], [33-66%], or [66-100%], respectively

IV. Results and Discussion

The left side of Table II, which corresponds to the experiments represented by Fig. 4(a,c,e), gives accuracy, AUC and MAE values on LAB data for the *Gen-LAB*, *PS-LAB*, and *PS-WILD* classifiers. Consistent with previous findings on person specific classifiers [31], the *PS-LAB* classifier has the highest performance. The *Gen-LAB* classifier has slightly lower accuracy than *PS-LAB* on average, but very similar AUC values. Meanwhile, *PS-WILD* has much lower performance. These results are unsurprising for two reasons: (1) weak labels are inherently less precise than accurate labels, and (2) LAB and WILD data are poor representations of each other, as discussed below.

On the right half of Table II, we can compare MAE values when testing on LAB versus WILD data, corresponding to partitions shown in Fig. 4(a,c,e) versus Fig. 4(b,d,f), respectively. *Gen-LAB* and *PS-LAB* both exhibit significant drops in performance ($p < 0.01$ for a one-sided paired t-test). The large variations in performance indicate that LAB data may not be very representative of WILD data. Interestingly, the largest performance deviation occurs with the *PS-LAB* classifier, implying that it is likely overfitting to the individual's LAB data. Consistent with findings from Hammerla [32], results show that the *PS-WILD* classifier experiences the least variation in performance between LAB and WILD data, and the difference is not significant ($p > 0.05$). Furthermore, *PS-WILD* demonstrates the highest performance on WILD data.

It is possible that one reason for the improved performance of *PS-WILD* is that it was able to learn the biases of the participants. That is, the participants may have perceived their tremors to occur more or less frequently than reality, and the algorithm learned to concur with these skewed perceptions. Alternatively, they may have interpreted the labels

TABLE II

COMPARING CLASSIFIER PERFORMANCE ON LAB AND WILD DATA

	Test on LAB (Fig. 4a/c/e)									Test on WILD (Fig. 4b/d/f)		
	Accuracy			AUC			MAE			MAE		
Subject	*Gen-LAB* Fig. 4(a)	*PS-LAB* 4(c)	*PS-WILD* 4(e)	*Gen-LAB* 4(a)	*PS-LAB* 4(c)	*PS-WILD* 4(e)	*Gen-LAB* 4(a)	*PS-LAB* 4(c)	*PS-WILD* 4(e)	*Gen-LAB* 4(b)	*PS-LAB* 4(d)	*PS-WILD* 4(f)
2 (L)	71.6	**89.9**	57.2	0.84	**0.86**	0.64	8.90	**0.22**	7.53	9.36	12.55	**7.47**
2 (R)	79.9	**85.4**	29.1	0.85	**0.90**	0.65	1.07	**0.17**	6.84	11.51	13.09	**8.51**
4 (L)	**83.4**	79.5	39.6	**0.89**	0.80	0.72	**1.39**	5.11	2.66	9.75	7.68	**6.05**
4 (R)	**89.5**	88.9	28.4	**0.92**	0.88	0.80	**0.06**	0.58	3.80	10.90	10.56	**5.46**
5 (L)	67.1	**74.6**	22.1	0.77	**0.80**	0.49	12.67	**3.54**	5.93	14.37	11.16	**4.50**
5 (R)	63.4	**69.1**	20.7	0.76	**0.77**	0.54	15.35	10.61	**9.39**	11.75	8.49	**3.80**
10 (L)	86.5	**87.6**	3.2*	0.75	**0.79**	0.56*	**1.71**	2.22	12.72*	6.44	8.93	**3.62***
10 (R)	86.0	**87.4**	2.2*	0.79	**0.88**	0.49*	**1.15**	1.18	1.44*	5.51	7.51	**3.12***
11 (L)	86.4	**88.1**	69.3	0.93	**0.94**	0.80	1.93	**1.50**	-	14.48	**12.72**	-
11 (R)	83.2	**83.3**	0.0*	0.88	**0.89**	0.75*	**3.67**	5.00*	16.58	4.64	3.57	**0.00***
12 (L)	87.3	**97.8**	0.3	**0.87**	0.77	0.53	**0.00**	**0.00**	3.52	3.81	7.51	**3.23**
12 (R)	84.0	**9.7**	0.9	**0.79**	0.75	0.76	0.43	**0.00**	8.12	6.53	14.66	**5.26**
Average	80.7	**85.9**	22.7	**0.84**	**0.84**	0.64	4.03	**2.51**	7.14	9.09	9.87	**4.64**

Bold indicates best performance within the associated performance metric.
MAE, described in Sec. III-D, is the mean absolute error in detected percentage.
Because accurate labels are not available for WILD data, accuracy and AUC could not be computed on WILD data.
* indicates results were averaged over only two partition. In the third partition, stratified MI-SVM for the *PS-WILD* classifier could not be initialized because the training set was lacking either "*Almost none*" or "*Almost always*" segments.
Note: for participant 11 (L), none of the training partitions included any "*Almost none*" segments.

differently, and the algorithm learned each participant's particular interpretation. However, while it would certainly be beneficial to clinicians to have completely unbiased symptom monitoring, the *PS-WILD* is no more biased than the current gold standard of patient self-reports. Furthermore, *PS-WILD* can at least improve monitoring frequency by automating it.

Another possible cause for the high performance of *PS-WILD* on the WILD data is due to overfitting. As described in [30], refraining from randomizing the samples before splitting into folds helps prevent performance estimates from being overly optimistic. However, to truly test generalizability, one would need to examine performance on data collected several months later. Nonetheless, these results indicate that training on ⅔ of weakly labeled WILD data generalizes better to the left out ⅓ than training on accurately labeled LAB data. It is possible that a laboratory dataset with many more participants and a broader set of activities could lead to better performance than the PS-WILD models. However, these findings suggests that, when building a system for automated, passive, continuous symptom monitoring, it may be more beneficial to personally tailor the system to specific users by training on their own, in-home, weakly labeled data than to invest significant resources in building a large, accurately labeled dataset from laboratory recordings of other people.

V. CONCLUSION AND FUTURE WORK

This paper directly analyzes how well data collected in laboratory settings represents data collected in-the-wild for the purpose of continuous, automated PD tremor detection. Previous work has typically trained machine learning algorithms on laboratory data under the assumption that results will generalize to in-the-wild data. Other work has collected data in the wild, but these datasets lack labels for training the algorithms and for assessing symptom detection performance on such data. Three different methods of partitioning the dataset were used to build three models – *Gen-LAB*, *PS-LAB*, and *PS-WILD* – per subject. For every model, performance on laboratory data differs greatly from that on wild data (see Table II). Furthermore, while the person-specific classifier trained on LAB data (*PS-LAB*) has the highest performance on LAB data, it has the lowest performance on WILD data. These findings imply that we should not assume in-lab performance will transfer to the wild.

Another interesting finding is that the person-specific classifier trained on WILD data (*PS-WILD*) performs better on WILD data than either of the classifiers trained on LAB data (*Gen-LAB* and *PS-LAB*). It is expected that training an algorithm on data from a specific user/environment will lead to higher performance on that user/environment. However, it is surprising that training on weak (*i.e.*, less precise) labels from the test subject can outperform training on accurate labels from the test subject.

Together, these findings suggest that when developing a system for continuous, automated symptom detection, higher accuracy can be achieved by asking users to weakly label their own, in-the-wild data for training than to invest significant resources in building a training dataset collected from other people. In this way, machine learning algorithms can be tailored to each user and a person-specific baseline can be established for later comparison during monitoring.

The work in this paper serves as a preliminary exploration into LAB versus WILD data. We envision a system where users might submit labels over the course of one or two weeks, after which a model would be trained and symptom detection would proceed automatically. It would be interesting to explore how many labels users would need to provide for performance to stabilize and whether the distribution of these labels over time affects performance. In this paper, due

to the small size of our dataset, we chose to use a linear SVM as our classifier. However, before building a product requiring users to submit their own labels, future work should investigate whether these findings hold with other feature sets, datasets, and algorithms.

VI. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1602337 and by the Center for Machine Learning and Health Fellowship in Digital Health. We would also like to acknowledge Andrew Whitford, Alex Cebulla, and Stanislav Panev for helping to collect the data, and Katelyn Stebbins labeling the data.

References

[1] Parkinson's Foundation, "Statistics," 2020. [Online]. Available: https://www.parkinson.org/Understanding-Parkinsons/Statistics

[2] J. M. Fisher, N. Y. Hammerla, T. Ploetz, P. Andras, L. Rochester, and R. W. Walker, "Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers," *Parkinsonism & Related Disorders*, vol. 33, pp. 44–50, 2016.

[3] S. Das, B. Amoedo, F. De la Torre, and J. Hodgins, "Detecting Parkinsons' symptoms in uncontrolled home environments: a multiple instance learning approach," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 3688–3691.

[4] R. P. Hubble, G. A. Naughton, P. A. Silburn, and M. H. Cole, "Wearable sensor use for assessing standing balance and walking stability in people with Parkinson's disease: a systematic review," *PloS one*, vol. 10, no. 4, p. e0123705, 2015.

[5] C. Ossig, A. Antonini, C. Buhmann, J. Classen, I. Csoti, B. Falkenburger, M. Schwarz, J. Winkler, and A. Storch, "Wearable sensor-based objective assessment of motor symptoms in Parkinson's disease," *Journal of Neural Transmission*, vol. 123, no. 1, pp. 57–64, 2016.

[6] K. J. Kubota, J. A. Chen, and M. A. Little, "Machine learning for large-scale wearable sensor data in Parkinson's Disease: Concepts, promises, pitfalls, and futures," *Movement Disorders*, vol. 31, no. 9, pp. 1314–1326, 2016.

[7] Á. Sánchez-Ferro, M. Elshehabi, C. Godinho, D. Salkovic, M. A. Hobert, J. Domingos, J. M. Van Uem, J. J. Ferreira, and W. Maetzler, "New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review," *Movement Disorders*, vol. 31, no. 9, pp. 1283–1292, 2016.

[8] T. O. Mera, D. A. Heldman, A. J. Espay, M. Payne, and J. P. Giuffrida, "Feasibility of home-based automated Parkinson's disease motor assessment," *Journal of Neuroscience Methods*, vol. 203, no. 1, pp. 152–156, 2012.

[9] S. Arora, V. Venkataraman, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3641–3644.

[10] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism & Related Disorders*, vol. 21, no. 6, pp. 650–653, 2015.

[11] Sage Bionetworks, "mpower: Mobile Parkinson disease study," 2015. [Online]. Available: http://parkinsonmpower.org/

[12] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey *et al.*, "The mpower study, Parkinson disease mobile data collected using researchkit," *Scientific Data*, vol. 3, p. 160011, 2016.

[13] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. Dorsey, S. Saria, and A. Terzis, "High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection," *arXiv preprint arXiv:1601.00960*, 2016.

[14] L. Chahine, L. Uribe, P. Hogarth, J. McNames, A. Siderowf, K. Marek, and D. Jennings, "Portable objective assessment of upper extremity motor function in Parkinson's disease," *Parkinsonism & Related Disorders*, vol. 43, pp. 61–66, 2017.

[15] F. Lipsmeier, K. I. Taylor, T. Kilchenmann, D. Wolf, A. Scotland, J. Schjodt-Eriksen, W.-Y. Cheng, I. Fernandez-Garcia, J. Siebourg-Polster, L. Jin *et al.*, "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial," *Movement Disorders*, vol. 33, no. 8, pp. 1287–1297, 2018.

[16] A. Weiss, S. Sharifi, M. Plotnik, J. P. van Vugt, N. Giladi, and J. M. Hausdorff, "Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer," *Neurorehabilitation and Neural Repair*, vol. 25, no. 9, pp. 810–818, 2011.

[17] R. I. Griffiths, K. Kotschet, S. Arfon, Z. M. Xu, W. Johnson, J. Drago, A. Evans, P. Kempster, S. Raghav, and M. K. Horne, "Automated assessment of bradykinesia and dyskinesia in Parkinson's disease," *Journal of Parkinson's Disease*, vol. 2, no. 1, pp. 47–55, 2012.

[18] C. Pulliam, S. Eichenseer, C. Goetz, O. Waln, C. Hunter, J. Jankovic, D. Vaillancourt, J. Giuffrida, and D. Heldman, "Continuous in-home monitoring of essential tremor," *Parkinsonism & Related Disorders*, vol. 20, no. 1, pp. 37–40, 2014.

[19] J. P. Giuffrida, D. E. Riley, B. N. Maddux, and D. A. Heldman, "Clinically deployable kinesia™ technology for automated tremor assessment," *Movement Disorders*, vol. 24, no. 5, pp. 723–730, 2009.

[20] A. L. Silva de Lima, T. Hahn, N. M. de Vries, E. Cohen, L. Bataille, M. A. Little, H. Baldus, B. R. Bloem, and M. J. Faber, "Large-scale wearable sensor deployment in Parkinson's patients: the Parkinson@home study protocol," *JMIR Research Protocols*, vol. 5, no. 3, p. e172, 2016.

[21] A. L. Silva de Lima, T. Hahn, L. J. Evers, N. M. de Vries, E. Cohen, M. Afek, L. Bataille, M. Daeschler, K. Claes, B. Boroojerdi *et al.*, "Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease," *PLoS One*, vol. 12, no. 12, p. e0189161, 2017.

[22] A. Zhang, A. Cebulla, S. Panev, J. Hodgins, and F. De la Torre, "Weakly-supervised learning for Parkinson's disease tremor detection," in *Engineering in Medicine and Biology Society (EMBC), 2017 Annual International Conference of the IEEE*. IEEE, 2017, pp. 143–147.

[23] A. Zhang, "Personalized and weakly supervised learning for parkinson's disease symptom detection," Ph.D. dissertation, Pittsburgh, PA, January 2020.

[24] P. Pierleoni, L. Palma, A. Belli, and L. Pernini, "A real-time system to aid clinical classification and quantification of tremor in Parkinson's disease," in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*. IEEE, 2014, pp. 113–116.

[25] B. T. Cole, S. H. Roy, C. J. De Luca, and S. H. Nawab, "Dynamical learning and tracking of tremor and dyskinesia from wearable sensors," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 982–991, 2014.

[26] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 313–322, 2007.

[27] F. M. Khan, M. Barnathan, M. Montgomery, S. Myers, L. Côté, and S. Loftus, "A wearable accelerometer system for unobtrusive monitoring of Parkinson's disease motor symptoms," in *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 120–125.

[28] S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. Growdon, D. Standaert, M. Akay, J. Dy, M. Welsh, and P. Bonato, "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.

[29] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[30] N. Y. Hammerla and T. Plötz, "Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1041–1051.

[31] G. M. Weiss and J. Lockhart, "The impact of personalization on smartphone-based activity recognition," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[32] N. Y. Hammerla, J. M. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plotz, "PD disease state assessment in naturalistic environments using deep learning," in *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI, 2015, pp. 1742–1748.