Weakly-supervised Learning for Parkinson's Disease Tremor Detection

Ada Zhang[1], Alexander Cebulla[2], Stanislav Panev[1], Jessica Hodgins[1], and Fernando De la Torre[1]
[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2] ETH Zurich, Switzerland

***Abstract*— Continuous, automated monitoring of Parkinsons Disease (PD) symptoms would provide clinicians with more information to understand their patients' disease progression and adjust treatment protocols, thereby improving PD care. Collecting precisely labeled data for Parkinson's symptoms, such as tremor, is difficult. Therefore, algorithms for monitoring should only require weakly-labeled training data. In this paper, we evaluate five standard weakly-supervised algorithms and propose a "stratified" version of three of the algorithms, which take advantage of knowing the approximate amount of tremor within each segment. In particular, we analyze PD tremor detection performance as training segments increase in length from 30 seconds to 10 minutes, and labels thereby become less precise. As segment length increases to 10 minutes, standard algorithms are not able to discriminate tremor from non-tremor. However, our stratified algorithms, which can make use of more nuanced labels, show little decrease in performance as segment length increases.**

I. Introduction

Chronic neurodegenerative movement disorders like Parkinson's Disease (PD) pose a serious threat to the elderly population. As many as one million Americans (mostly aged 65 or older) live with PD [1], and there is no cure. Medication can provide symptomatic relief, but patients require larger and more frequent doses as sensitivity to these drugs decreases over time. The current standard of care is as follows: Patients meet with their doctor every three to six months, self-reporting on symptoms and response to medication. Doctors then perform quick motor function assessments by examining patient performance on motor tasks selected from Part III of the Unified Parkinson's Disease Rating Scale (UPDRS) [2]. Finally, doctors adjust medication dosages as necessary.

There are several shortcomings to the current state-of-the-art in PD management: (1) Patient self-reports are often inaccurate and 15-20 minute clinic visits do not provide enough information for doctors to accurately assess their patients; (2) motor function assessments are not only subjective, but also dependent on the time of day and time since the last medication intake; and (3) clinic visits are insufficiently frequent due to their high cost and inconvenience for the patients. All of these factors make it difficult to monitor disease progression and adjust treatment protocols.

Continuous PD motor symptom monitoring should lead to better adjustment of medication and therefore improvements in patient quality of life. Our eventual goal is to build a system for continuous monitoring of PD motor symptoms through wearable sensors, such as accelerometers. The first step in building such a system is to develop machine learning algorithms that can handle the type of data that they will see during constant home use while users perform activities of normal daily living. We anticipate that the training data for such a system will be weakly-labeled because precise labels (exact start and end of every PD symptom) are not possible to obtain "in the wild" and laboratory data will not contain a rich enough set of activities.

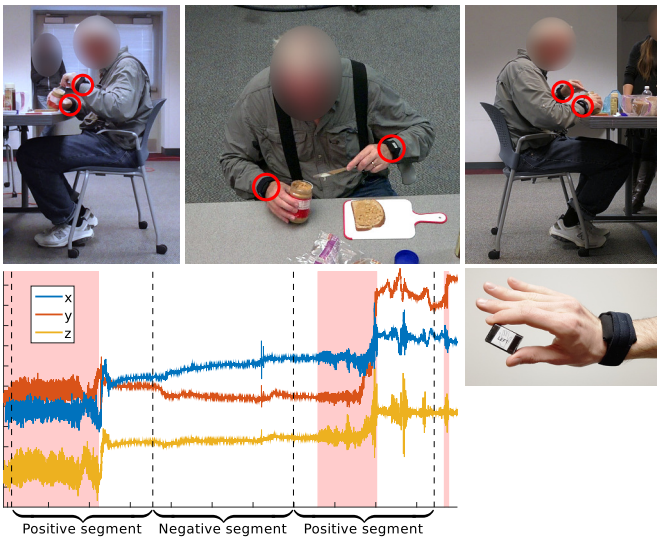

Fig. 1. Top row depicts the three camera views (frontal and two sides) from our experimental setup. Accelerometer locations are circled in red. In the bottom left, we see a sample signal from the accelerometer with ground truth tremor events highlighted in pink. Example segments and their labels are shown below. Note that positive segments indicate the presence of tremor, but not the exact location. The bottom right image depicts the Axivity AX3 accelerometer that we are using.

This paper offers two main contributions for the detection of PD tremor, one of the symptoms of Parkinson's: (1) As a first step toward designing a data collection of PD motor symptoms in the wild, we simulate the types of labels that would be available in such environments and analyze how the performance of several weakly-supervised learning algorithms degrades as labels become less precise. (2) We provide a simple modification to existing weakly-supervised learning algorithms that allow them to take advantage of labels containing the approximate percentage of tremor (*e.g.*, 0-24%, 25-49%, 50-74%, 75-100%) within a segment.

Our analysis finds that once time segments are 10 minutes long, the standard algorithms are not able to discriminate tremor from non-tremor. However, our modified algorithms, which make use of more nuanced labels, show little de-

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

crease in performance as the segment length increases to 10 minutes. This result implies that, for future in-home studies where subjects will be asked to label their own data, 10-minute time segments are sufficiently short (provided subjects can accurately label them) for the learning algorithms, and that subjects should try to label the approximate amounts of tremor within those segments.

II. Related Work

Although many researchers have explored the use of wearables in detecting or evaluating PD motor symptoms, little work has been done on detecting symptoms continuously in the wild. Experiments that included unscripted activities were generally restricted to a maximum of 4-hour long data collections. Some experiments were conducted in research participants' homes over periods of one month [3], [4], or even longer, as in the mPower Mobile Parkinson Disease Study [5]. In these cases, however, participants were only evaluated when they performed specific tasks, rather than continuously throughout the day.

We believe the main limitation for detecting symptoms in the wild is the difficulty of obtaining precise labels (*i.e.*, the exact start and end of tremor events). Coarse labels, such as whether tremor events occur within a larger time segment or not, are a feasible solution. Data with these imprecise labels are called weakly-labeled. While weakly-supervised learning methods are relatively well-explored (see [6] for a review), few have applied these methods to PD symptom detection. In [7], Fisher et al. used a neural network to detect dyskinesia in weakly-labeled data collected in the homes of subjects with PD. However, the authors applied the labels of their one-hour time intervals to every sub-interval. That is, if a subject reported the occurrence of dyskinesia events during a one-hour segment, the authors assumed that dyskinesia was constantly present during the *entire* segment. This naive approach introduces many false positive labels. In this work, we compare the naive approach to several algorithms that explicitly account for the fact that only a portion of a positive segment is the event of interest.

Most similar to our work here is that of Das et al., who compared several weakly-supervised learning techniques on in-home data collected from two subjects [8]. One subject experienced dyskinesia and the other had tremor. Four days of data were collected and subjects labeled their data using paper diaries. While this work served as a proof-of-concept for the utility and necessity of weakly-supervised learning for data collected in the wild, the number of subjects was limited (only one per symptom type) and there was no way to verify that symptoms detected in the wild were true symptom occurrences because there were no ground truth labels.

In this work, we extend that of Das et al. [8], providing a more thorough exploration into the effect of weak-labels on algorithm performance. We have collected a larger dataset in a laboratory environment of subjects who all experience tremor. This dataset allows us to perform leave-one-subject-out validation to test the generalizability of the algorithms. Using the associated video data, we have annotated the start and end of each tremor event to serve as the ground truth. We can thus *simulate* different levels of weak-supervision by controlling the length of the time segment for which labels are provided, and analyze how algorithm performance drops as label uncertainty increases. Furthermore, since ground truth labels exist, we can confirm whether algorithms trained on weakly-labeled data are able to discriminate tremor from non-tremor.

TABLE I
Set of actions performed during data collection

Action	Approximate time (minutes)
Sit and talk	5
Rest tremor* (UPDRS 3.17)	3
Postural tremor* (UPDRS 3.15)	6
Kinetic tremor (UPDRS 3.16)	2
Finger tapping (UPDRS 3.4)	1
Open/close hands (UPDRS 3.5)	1
Pronation/supination of the hands (UPDRS 3.6)	1
Writing	4
Typing	4
Playing chess	10
Playing cards	10
Making a sandwich	5
Eating a sandwich	10
Drinking from a cup	1
Walking	2

* denotes the inclusion of a cognitive distractor, which consisted of counting backwards by 7's from 100.
Note: UPDRS item numbers correspond to those given in [2]

III. Methods

In this section, we describe our experimental setup for comparing several different algorithms and their performance on labeled time segments of varying lengths. We first describe our methods for computing feature vectors from the data and assigning labels to the segments. We then describe the algorithms we compared and our modification, which allows the algorithms to use labels describing the approximate amount of tremor within each segment. Finally, we describe our protocol for splitting the data into training and test sets.

A. Data Collection

Data were collected from four male and one female subject, who had been diagnosed with Parkinson's disease two to four years prior. Each subject experienced 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 (see Table I for a complete list). Data were collected at 100 Hz. Three cameras were used to record the subjects so as to minimize occlusion. These video data were used to annotate tremor events, thereby providing ground truth data. Figure 1 depicts our experimental setup. Table II shows a summary of the collected, labeled data. This research was approved by the Carnegie Mellon University Institutional Review Board.

TABLE II

SUMMARY OF COLLECTED DATA

Subject	# Labeled Seconds	% Tremor Events
1	6552	70.3
2	6626	52.1
3	10560	45.4
4	11158	18.5
5	12698	14.6
Total	**47594 ($\sim$ 13.2 hours)**	**40.2**

Note: Each hand is treated as a separate signal so the number of labeled seconds is twice the time of the recording

TABLE III

PERCENTAGE OF EACH TYPE OF SEGMENT LABEL FOR VARYING SEGMENT LENGTHS

Segment length	Standard labels		Stratified labels			
	Positive	Negative	0-24%	25-49%	50-74%	75-100%
30 s	65.4	34.6	53.0	13.5	10.4	23.1
1 min	76.0	24.0	51.0	16.5	13.6	18.9
3 min	91.9	8.1	47.9	19.7	16.2	16.2
5 min	94.2	5.8	43.9	25.8	16.8	13.6
10 min	97.3	2.7	41.9	29.7	12.2	16.2

B. Features and Labels

Previous work on automated tremor detection [9]–[13] generally use very similar features. In this work, we used the same features as those described by Patel et al. [13]. Features were computed over two-second windows of the signal with one-second overlap. Each window was considered to be tremor if at least half of the window was tremor.

Consecutive windows were collected into segments of varying lengths (from 30 seconds to 10 minutes). Segments were labeled with two different methods:

- Standard method – positive if the segment contained at least one "tremor" window (event), negative otherwise.
- Stratified method – approximate percentage of tremor (*e.g.*, 0-24%, 25-49%, 50-74%, 75-100%) within the segment.

C. Algorithms

For our experiments, we compare the performance of several different algorithms. In particular, we chose to compare the top three performing algorithms reported by Das et al. [8] – Multiple Instance Support Vector Machine (MI-SVM) [14], Iterative Discriminative Axis Parallel Rectangle (ID-APR) [15] and Expectation Maximization Diverse Density (EM-DD) [16]. Given the recent successes of neural networks in machine learning, we also included the Multiple Instance Neural Network (MI-NN) algorithm [17]. We compare these weakly-supervised techniques to a naive Support Vector Machine (Naive-SVM) baseline, which applies that segment-level labels to every subsegment within. That is, if a segment is labeled as containing some tremor, Naive-SVM assumes the entire segment consists of tremor.

We also propose a modification to the MI-SVM, ID-APR and MI-NN algorithms that allows them to take advantage of knowing the approximate amount of tremor within a segment given by stratified segment labels. In our data collection, we found that the incidence of tremor varied greatly across subjects (from 14.5% to 70.3% occurrence), as shown in Table II. The MI-SVM, ID-APR and MI-NN algorithms are biased towards finding rare positive examples among many negative examples – an assumption that holds well for subject 5, but which fails for subject 1. For subjects with frequent tremor, these algorithms may have low recall because they will classify very few events as tremor.

Performance of these algorithms could be improved if they had access to slightly more nuanced labels, such as our stratified labels, which contain the approximate percentage of tremor within a time segment (*e.g.*, 0-24%, 25-49%, 50-74%, 75-100%). In the scenario where subjects are labeling their own in-home data, these stratified segment labels could correspond to labels of "almost no tremor," "not much tremor," "a lot of tremor," and "almost constant tremor."

In addition to containing more information, these stratified labels also have the advantage that they are symmetric between tremor and non-tremor. In contrast, because increased segment length is associated with a higher likelihood of tremor occurring within, labeling segments with the standard method results in a lower incidence of negatively labeled segments as segment length increases. This lack of negative examples can make it difficult for an algorithm to discriminate between tremor and non-tremor. The proportion of each type of stratified label, however, remains relatively constant as segment length increases (see Table III).

MI-SVM, ID-APR, and MI-NN are all solved by iterating between choosing a single event from each positive segment to serve as a positive selector variable (all events in negative segments are considered negative selector variables) and solving the decision boundary given the selector variables. Modifying these algorithms to use stratified labels simply involves changing the number of selector variables pulled from each bag. The number of positive and negative selector variables is adjusted according to the segment label. For example, given a segment with a label of 50-74%, we know that *at least* 50% of the segment must be tremor events, and *at least* 25% of the segment must be non-tremor events. Therefore, the highest scoring 50% of the events are chosen to be positive selector variables, and the lowest scoring 25% are chosen to be negative selector variables. The algorithm is then trained on these selector variables as normal. For a naive baseline using these labels, we assign positive labels to segments with at least 50% tremor, negative labels to the rest, and then apply the naive assumption to the events in these segments and train a linear SVM.

We call these modified algorithms "stratified" and they are closely related to work on learning from label proportions (LLP) [18]–[20]. Much of the work in LLP involves new, probabilistic models that assume the exact proportions of positive to negative events are known. Kück and de Freitas allow for uncertainty in the proportions, but do so by introducing a user-specified parameter to represent this uncertainty [19]. In contrast, our solution is a simple modification of

existing algorithms given *approximate* proportions of labels with no additional parameter.

D. Algorithm Training and Testing Protocol

Data were split into training and testing sets using leave-one-subject-out cross-validation: Data from one subject were left out and the algorithm was trained on data from the remaining subjects. This process was repeated for each subject. Results were averaged across subjects.

IV. Results and Discussion

The goal of this work is to explore how the performance of weakly-supervised algorithms degrades as labels become less precise (time segments become longer). We plan to use the results presented here to inform a future data collection in the wild, where we envision subjects assigning coarse labels (presence/absence of tremor, or the approximate amount of tremor) to short time segments throughout the day.

A. Standard Metrics

We compared the performance of the algorithms on five standard metrics:

$$\text{Accuracy} = \frac{\text{\# correctly classified}}{\text{\# events}},$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$\text{Precision} = \frac{\text{\# true positives}}{\text{\# classified as positive}},$$

$$\text{Recall} = \frac{\text{\# true positives}}{\text{\# ground truth positive events}},$$

$$\text{Specificity} = \frac{\text{\# true negative}}{\text{\# ground truth negative events}}.$$

These performance metrics were computed on event-level labels (as opposed to segment-level labels) to measure whether the algorithms were able to discriminate tremor from non-tremor after being trained on the weakly-labeled data.

Algorithm performance is summarized in Figure 2. As expected, performance generally falls as the segment length increases because the labels become less precise. We can see that the Naive-SVM algorithm fails once input segments become three minutes long: specificity falls to zero, implying that the algorithm has converged to classifying everything as positive (tremor). This failure demonstrates the necessity of weakly-supervised algorithms for accommodating weakly-labeled data. We can also see that, contrary to what Das et al. found [8], the performance of ID-APR is generally worse that MI-SVM and EM-DD. In fact, at 3-minute-length segments, the recall of ID-APR has fallen to nearly zero, implying that it is classifying nearly everything as non-tremor.

For 10-minute-length segments, we see that the non-stratified algorithms have nearly converged to classifying everything as positive (specificity $= 0$) or negative (recall $= 0$). In contrast, the stratified algorithms (excluding the naive version) avoid this behavior. We see the usual trade-off between recall and specificity. However, it is interesting to note that while all four have similar precision (proportion of the events classified as tremor that were true tremor), the recall of the stratified Naive-SVM algorithm is much lower (it is less able to find tremor). Using these metrics it is not clear which of the stratified methods should be preferred, although MI-NN does show better performance in four of the five metrics. However, our proposed performance metric (described below) is able to shed light on which is the superior algorithm.

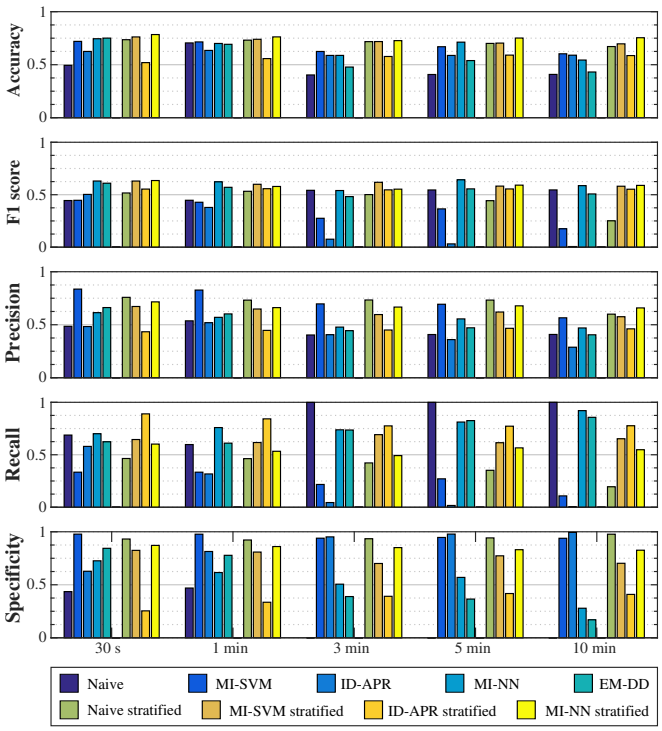

Fig. 2. Standard performance metrics computed on all seven algorithms over varying lengths of training time segments

B. Proposed Performance Metric

We believe that clinicians may find the amount of tremor that occurred within a given period of time to be more informative than whether a particular instant is tremor or not. Therefore, the error in the detected percentage of tremor within a window versus the true percentage of tremor may be a more clinically relevant performance metric. We chose 15-minute windows because this resolution is short enough to give information about the effect of a patient's medication while remaining easily interpretable by the clinician. Table IV shows the mean absolute error in detected tremor for all nine algorithms as they are trained on segments of varying length. We can see that the stratified MI-NN algorithm generally shows the best performance in this metric, although the stratified MI-SVM algorithm is a close second.

V. Conclusion and Future Work

In this work, we took the first step toward developing a system for automated, continuous monitoring of PD motor symptoms. We simulated the types of labels that will be

TABLE IV

MEAN ABSOLUTE ERROR OF DETECTED TREMOR PERCENTAGE WITHIN 15 MINUTE WINDOWS

Algorithm	Training segment length				
	30 s	**1 min**	**3 min**	**5 min**	**10 min**
Naive-SVM	45.5	23.8	64.3	64.3	63.7
MI-SVM	22.9	23.4	28.7	25.1	30.6
ID-APR	18.0	20.8	33.0	34.4	35.9
MI-NN	14.7	17.4	24.3	19.6	30.2
EM-DD	16.8	19.8	45.2	41.7	53.0
Naive-SVM stratified	19.3	18.8	20.1	22.8	28.4
MI-SVM stratified	13.9	16.1	**16.2**	17.6	19.0
ID-APR stratified	42.2	36.3	31.7	30.4	31.0
MI-NN stratified	**13.2**	**14.7**	18.0	**16.3**	**14.2**

available in the second, in-home phase of this study and compared the ability of several algorithms to learn from these weak labels. Using our finely-labeled dataset from five subjects, we could control the length of the training segments and thereby analyze how algorithm performance degraded as labels were made less precise. Furthermore, because we have ground truth labels for every time point, we were able to evaluate the ability of the algorithms to discriminate tremor from non-tremor events. This work thus serves to extend that of Das et al. [8], where only two subjects were used (each with a different motor symptom) and where ground truth labels were not available. We also developed a novel modification to MI-SVM, ID-APR, and MI-NN, which allow them to take advantage of knowing the approximate amount of tremor within a given time segment.

Our analysis resulted in two main findings. (1) Contrary to that reported in Das et al. [8], ID-APR did not have the best performance. We believe our results differ due to our more thorough analysis of multiple performance metrics and our larger dataset. (2) Our novel, stratified algorithms generally showed improved performance over their standard counterparts, particularly as the segment length increased. We found that the stratified version of MI-NN gave the best performance overall.

In future work, we plan to test these learning algorithms on data collected in the wild, which would be labeled by the research participants themselves. Our findings suggest 10-minute length segments are short enough for algorithms to learn from, provided participants label the approximate amount of tremor within.

VI. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1602337. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Parkinson's Disease Foundation, "Statistics on parkinson's," 2013. [Online]. Available: http://www.pdf.org/en/parkinson_statistics

[2] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[3] 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.

[4] 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.

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

[6] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Advances in neural information processing systems*, pp. 577–584, 2003.

[15] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[16] Q. Zhang, S. A. Goldman *et al.*, "Em-dd: An improved multiple-instance learning technique," in *NIPS*, vol. 1, 2001, pp. 1073–1080.

[17] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multi-instance learning," in *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*, 2002, pp. 455–459.

[18] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions," *Journal of Machine Learning Research*, vol. 10, no. Oct, pp. 2349–2374, 2009.

[19] H. Kuck and N. de Freitas, "Learning about individuals from group statistics," *arXiv preprint arXiv:1207.1393*, 2012.

[20] J. Hernández-González, I. Inza, and J. A. Lozano, "Learning bayesian network classifiers from label proportions," *Pattern Recognition*, vol. 46, no. 12, pp. 3425–3440, 2013.