2023 19th International Conference on Mobility, Sensing and Networking (MSN) | 979-8-3503-5826-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/MSN60784.2023.00119

A Weakly Supervised Learning Framework for Parkinson's Disease Assessment Using Wearable Sensor

Ziheng Li
National Pilot School of Software
Yunnan University
Kunming, China
liziheng9050@mail.ynu.edu.cn

Xiyang Peng
Department of Computer Science
University of Sheffield
Sheffield, UK
xpeng24@sheffield.ac.uk

Yuting Zhao
National Pilot School of Software
Yunnan University
Kunming, China
zhaoyuting@mail.ynu.edu.cn

Xulong Wang
Department of Computer Science
University of Sheffield
Sheffield, UK
xl.wang@sheffield.ac.uk

Yun Yang
National Pilot School of Software
Yunnan University
Kunming, China
yangyun@ynu.edu.cn

Po Yang*
Department of Computer Science
University of Sheffield
Sheffield, UK
po.yang@sheffield.ac.uk

***Abstract*—Wearable technology has played a crucial role in computer-aided diagnosis and long-term monitoring of Parkinson's disease (PD). How to efficiently and accurately assess the severity of Parkinson's disease using wearable devices remains the essential problem. However, in the real free-living environment, we have encountered two issues: weak annotation and class imbalance, which could potentially impede the automatic assessment of Parkinson's disease. To overcome these challenges, we propose a novel Parkinson's disease assessment framework in free-living environment. Specifically, clustering methods are used to learn latent categories from the same activities, and use Latent Dirichlet allocation (LDA) topic models to capture latent features of multiple activities. Then, to mitigate the impact of data imbalance, we augment bag-level data while retaining key instance prototypes. The new framework is applied to a PD dataset collected by wearable sensors in the wild. It achieves an impressive 73.49% accuracy in the fine-grained (normal, mild, moderate, severe) classification of PD severity based on hand movements. Overall, this study contributes to more accurate PD self-diagnosis in the wild, enabling remote guidance for drug intervention from doctors.**

***Index Terms*—Data Augmentation, Parkinson's disease, Wearable sensor, Weak annotation.**

I. Introduction

Parkinson's disease (PD) has been ranked as the second most common disease worldwide, affecting a significant portion of the elderly population [1]. It is estimated that by 2030, approximately 9 million people in the ten most populous countries will suffer from this disease [2]. PD is characterized by a severe loss of dopamine in the forebrain, resulting in motor symptoms such as tremors, muscle stiffness, bradykinesia, postural instability, as well as non-motor symptoms including hyposmia, sleep disturbances, and autonomic dysfunction [3]. To effectively assess these motor symptoms, rating scales have been widely adopted, such as the MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [4]. However, these assessments typically occur in clinical settings with infrequent annual visits, and the UPDRS evaluation is time-consuming, requiring at least 30 minutes and specialized training [5]. These factors contribute to the challenge of monitoring Parkinson's disease effectively. Therefore, there is a need for convenient and objective PD assessment tools to better assist patients.

With the popularization of IoT devices and the advancement of machine learning technology [6]–[8], objective assessment of PD symptom severity through wearable inertial sensors has seen a substantial amount of research. Previous studies [9] employed inertial sensors to precisely measure PD symptoms in a controlled laboratory environment. These research demonstrates the effectiveness of wearable technology in monitoring PD symptoms, providing valuable insights for doctors to gain a better understanding of the patient's PD symptom.

Although wearable technology has demonstrated significant potential in monitoring PD symptoms, it is still very difficult to use wearable devices to assess PD in a free-living environment. In real situations, we found the following two problems:

- Weak Annotation: In real cases, it is very time-consuming to obtain detailed symptom annotations [10]. Expert raters usually score the PD stage of the patient's motor function for a long time, but arbitrary segmentation into fixed-length windows may not reflect disease-related features [11] (merely with longtime annotation). **This is generally considered a weakly supervised problem, which inspired us to develop a assessment framework in the weakly supervised setting.**
- Class Imbalance: most patients will be concentrated in mild Parkinson's disease, and the proportion of severe patients will be small (in the data set we collected, severe

979-8-3503-5826-1/23/$31.00 ©2023 IEEE
DOI 10.1109/MSN60784.2023.00119

patients are usually only 1/4 of mild patients). This leads to class imbalance problem. At the same time, there is often high variability in motor performance between different PD patients at the same stage. **This makes us need to find effective ways to make full use of data to solve this problem.**

The mentioned factors pose challenges in assessing the stage of PD. Building upon these challenges, we propose a novel weakly supervised framework for PD assessment, aiming at solving the problem of PD diagnosis in free-living environments under weakly-annotated settings and with data imbalance.

Our method is divided into two parts, one is PD learning framework under weak supervision: We first use fixed-window segments to extract features from the patient's different activities sensor signal data, Then k-means clustering is used to cluster the different segments of the same activity of all patients. Latent Dirichlet Allocation (LDA) topic models to generate global features from these clustering labels in order to discover the hidden topic information between different activities of PD patients. These two features are then fused as a re-representation of the PD patient features. Then we use the data augmentation method, first find similar patient pairs through similarity comparison to mix, At the same time, by disrupting the order of different segments in the bag to weaken the relationship between the position and time of the segments, more diverse samples are generated, and use the generated pseudo data for training. The intuition behind our method is that the fine-grained features of short-term fixed windows may not reflect the overall disease stage of the patient, we hope to discover the implicit association information between different activities of the patient through the unsupervised topic model. Finally, we performed a PD stage classification test on the real PD free-living environment dataset of 83 subjects, and achieved an accuracy rate of 73.49%(normal, mild, moderate, severe), 11% higher results than segment-based PD stage classification. Proving the effectiveness of our method. Specifically, our main contributions are as follows:

- We propose a PD stage assessment framework under the weak annotation setting, which classifies by combining local features of multiple segments and global topic features of multiple activities.
- We propose a simple yet effective data augmentation method to generate more data to enrich the original data and improve the classification performance of minority classes.
- Validation and evaluation on a dataset of 83 people in a real free-living environment demonstrates the effectiveness of the method.

The rest of the paper is organized as follows. Related works are discussed in section II. Parkinson's disease stage assessment framework is presented in section III, evaluated and results presented in section IV. Section V concludes this paper.

II. Relate work

We will review related work from three aspects: 1)Wearable technology PD severity assessment. 2)Weakly supervised learning. 3)Time series data augmentation.

A. Wearable technology for PD severity assesment

There have been considerable studies on the monitoring and management of Parkinson's disease motor symptoms through wearable technology, and it has been proven to be as effective as clinical scores [12]. Previous research [13], [14], Monitoring single PD symptoms via wearable sensors. However, most studies only focus on a single symptons. A reason is that the evaluation of PD staging often requires comprehensive evaluation of multiple PD motor symptoms, which brings difficulties to this work. At the same time, most of their work did not consider the problem of weak annotation.

B. Weakly supervised learning

In practical situations, it is very difficult for PD patients to accurately record the onset time of each symptom, but PD stage labels based on long-term observation are available. This is often considered a weakly supervised learning setting [15]. To address this issue, there are currently many studies using multiple-instance learning(MIL) methods. Such as instance-level methods [16] and bag-level methods [17]. At present, MIL has been widely used in various fields. But as far as I know, MIL is less considered in the field of human activity recognition, and there are only a few studies [18], especially in the evaluation of Parkinson's disease. Therefore, in this study, we propose a MIL-based framework for PD diagnosis to address the weak annotation problem of real situations.

C. Time series data augmentation

Data augmentation is an effective method to increase the number and diversity of samples under limited data. There have been many studies on data augmentation, especially in the field of medical data, because medical data collection is very expensive and often has severe class imbalance.

One of the methods is to generate more diverse samples through small transformations in the original signal to enrich the feature expression, such as [19], They defined a series of data augmentation methods for wearable sensor signals, including Permutation, magnitude-warping, cropping, jittering, Rotation. And it has improved in the three status classifications of PD bradykinesia, OFF, and dyskinesia. However, this method may not be suitable for the evaluation of PD stage, because the magnitude-warping, jittering and other methods may change the severity of patients' symptoms and do not consider the differences between patients. Other data augmentation work such as [20]. These methods are all based on instance-level fixed-window data augmentation, which loses the holistic information of multiple activities and ignores the differences among subjects.

III. Methodology

A. Framework overview

The comprehensive framework is visually represented in Figure 2. This framework is organized into five distinct components. The initial part encompasses data preprocessing, sliding window segmentation. The second component extracted the features from the original signal. In the third part, these segments features will be used for k-means clustering and aggregation, We then fit the distribution of these cluster labels through the LDA model to generate new global features. In the fourth part, we use data augmentation methods to generate more data and alleviate the negative effects of data imbalance. The final segment pertains to the training and testing phase, where machine learning models are trained and assessed using these features. We define the research problem as a four-class classification problem. Input the features x of the participants' activities and output the patient's PD stage y (Healthy, mild, moderate, and severe PD). This approach aims to develop a machine learning model which is capable of evaluating the disease severity of PD patients. Detailed descriptions of each step will follow in the subsequent sections.

B. Data collection

The data for this research study were collected by our team at the hospital from January 15, 2021, to July 30, 2022. A total of 70 PD patients and 15 healthy volunteers participated in the study. In this study, all of the participants signed the informed. During the activities, all participants wore Shimmer3 IMUs on their left wrist, right wrist, left ankle, right ankle, and waist to collect acceleration and gyroscope signal data. To minimize the burden on patients, in subsequent experiments, we only utilized data from the right wrist sensor. And only uses part of the activities.

The Shimmer3 IMU is connected to the computer through wireless Bluetooth. On the computer, we use the ConsensysPRO software to collect the signal data at a high sampling frequency of 200HZ. The participant will perform 12 kinds of activities, with a 1-minute rest between each activity. Before the experiment begins, the researchers will instruct the PD patients to complete the requirements of the activity. While after the experiment began, there is no any guidance or interference from the investigator. Videos will be recorded during the collection and all the data will be scored according to the H-Y scale by a neurologist. The label is individual-level based on the patient's performance in multiple activities. Table I and Figure 1 present the activities performed in the experimental setup. After excluding abnormal subjects, 83 subjects were finally used in the study.

C. Data Preprocessing and Segmentation

We first use a 4th-order Butterworth filter with a bandpass range of 0.3Hz-20Hz filter out the gravity component. Z-score normalization will be applied to the signal, after which the data will be sliced at 300 data points (1.5 seconds) with 50% overlap. Finally, We will compute time and frequency domain-related features (standard deviation, variance, skewness, kurtosis, rms, energy, median, range, correlation.....).

Fig. 1. Overview activity: (a)Finger taps (b)Clench and open alternately (c)Rapid alternating movements of hands (d)Hand rotation-right/left (e)Finger to nose-left/right (f)Standing with arms hold (h)Walk back and forth (i)Arising from chair (j)Drinking water (k)Picking things

TABLE I
DEMOGRAPHIC DATA OF STUDY POPULATION. (MEAN AND STD.)

	Healthy	**PD**	**Total**
Num.patient	15	70	85
Age	23.56(2.24)	67.57(7.84)	49.78(20,47)
Weight	62.42(.5.52)	58.51(9.60)	62.48(10.24)
Height	171.66(11.39)	160.51(7.73)	169.25(9.05)
UPDRS Level	Healthy:15	Mild:41 Moderate:17 Severe::12	Healthy:15 Mild:41 Moderate:17 Severe::12
Gender Ratio	Male:13 Female:2	Male:37 Female:33	Male:50 Female:35
Num.instance	1618	7821	9439

D. Feature extract and fusion

Clustering to create documents: As shown in Fig 2, the multiple activity signals are divided into segments, $i_1^1, i_1^2, i_1^3, i_1^4 ..... i_2^1, i_2^2, i_2^3, i_2^4, i_m^n$ m represents the m-th activity, n represents the n-th window segment of the activity. i represents the feature set of the segment. Subsequently, the feature set of the same activity segment of different patients will be clustered using k-means, and the clustered labels will be used as words to form documents:

$$\text{Document}_p = \left[F_1^1, F_1^1 \ldots\ldots F_a^b\right] \tag{1}$$

F_a stands for k-means clustering label, b stands for the b-th activity, and p stands for the p-th subject. Through this method, we aggregate the words of multiple activities of the patient to generate a document, and then use the topic model to generate a global feature.

LDA topic model generate global features: The document-word matrix is the input to LDA, and LDA outputs the document-topic distribution as global features. The detailed calculation process can be found in [21]. The advantage of this

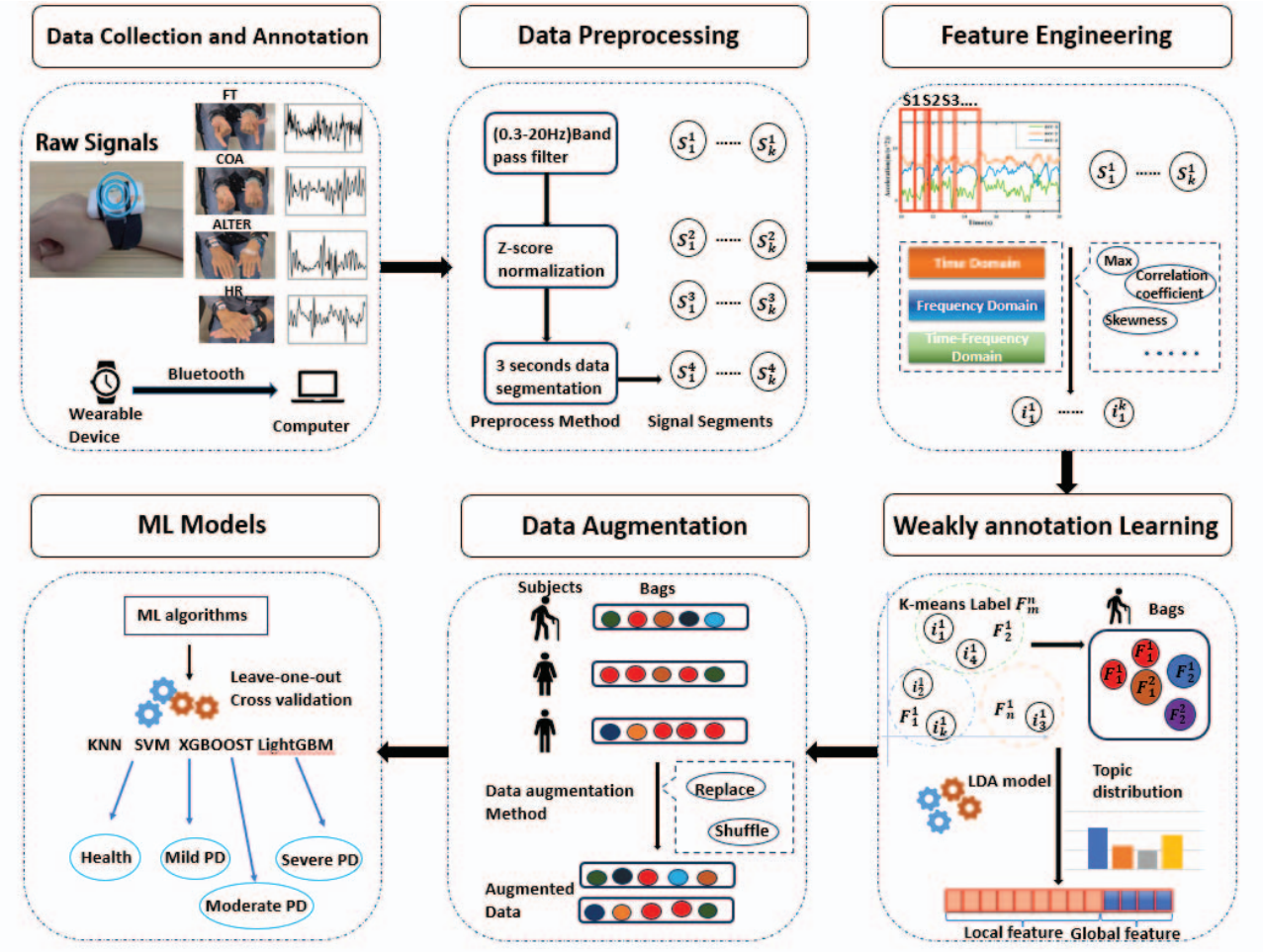


Fig. 2. Overview of the framework: The original wearable device signal is decomposed into various segments via slicing and feature extraction, and these segments are clustered to generate cluster labels. The document will be composed of multiple active clustering labels and the topic probability distribution will be generated by the topic model as the extracted global features for classification. We then used the package's data augmentation method to generate more samples.

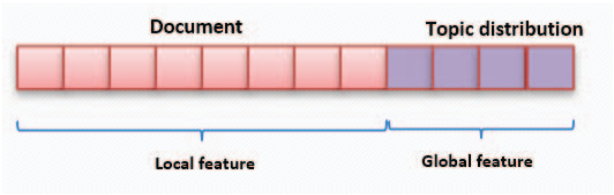


Fig. 3. Feature vector

is that each subject's different activities will generate a topic distribution feature, which is generated based on the global information of multiple activities, with more information and better feature expression. Fig 3 shows the final feature vector.

E. Data Augmentation

Data augmentation has two purposes, one is to improve the prediction accuracy of the minority class to solve the problem of class imbalance, and the second is to reduce the variability between patients and improve the robustness of the model. Fig 4 shows how the data augmentation method works. First, we assemble the instance clustering labels into vectors A and B according to the original chronological order. Then, we calculate the distances between different patients using the formula 2 and select pairs with close distances. Afterward, we use two methods to mix the samples: a) We re-randomly mix similar sample pairs to generate new samples. These Bag pairs are derived from the same PD stage patients, so the labels remain unchanged, resulting in a greater diversity of samples and alleviating intra-class differences. b) We shuffle the order of instances within each bag. The purpose of this is to handle the uncertainty of when patients develop symptoms. Through shuffling, we generate samples that are independent of time.

$$\text{Distance(H)} = \sum_{i=1}^{n} (A_i \neq B_i) \tag{2}$$

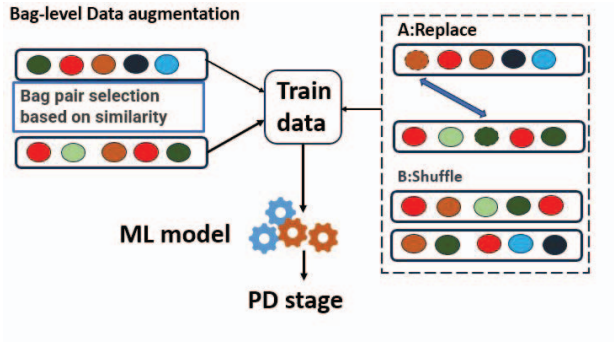
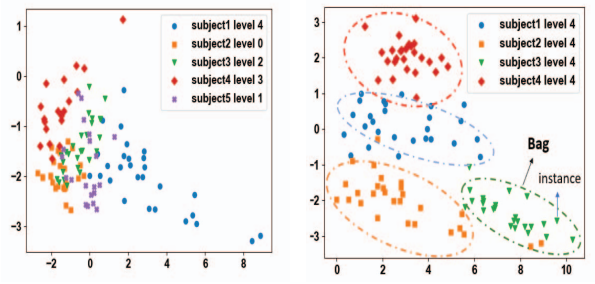


Fig. 4. Bag-level augmentation method



(a) (b)

Fig. 5. Data distribution after PCA dimension reduction. Each point represents the activity signal characteristics of the subject in the 1.5s window. (a)Data distribution of different PD severity levels. (b)Individual differences in the same category.

IV. RESULTS OF PROPOSED FRAMEWORK

A. Experimental setup and evaluation methods

We selected different basic machine learning algorithms to verify our proposed framework, including KNN, SVM, XGBOOST, Lightgbm, and used the leave-one-out cross-validation method in our 83-person data set. Finally, we used Accuracy(ACC), precision(P), recall(R), F1-score as evaluation indicators.

B. Weak annotation

Figure 5(a) shows the instance distribution of 5 subjects in different PD stages. It can be inferred that not all instances of PD patients are in the ON state, and there may be instances similar to level 0, but the instance-level labels are difficult to obtain. This leads to the emergence of the weak labeling problem.

Figure 5(b) shows four subjects in PD stage 4. Although they are all performing the same activity and labeled as PD stage 4, the instance distributions of different subjects show differences due to the Subject 1 developed tremors while subject 4 performed bradykinesia during the performance of the activity. This shows that different subjects of the same category will produce individual differences, resulting in more difficult classification, so more abundant data is needed to enhance sample diversity.

TABLE II
DATA AUGMENTAION RESULT

Method	Accuracy	Precision	Recall
No augmentation	68.67	68.23	68.67
Rotation	65.06	64.86	65.06
Scaling	66.27	66.07	66.27
Random Oversample	71.08	70.55	71.08
SMOTE	71.08	70.72	71.08
Propose Framework	73.49	73.77	73.49

C. Parameter setting

In order to find the optimal number of cluster centers to determine the type of words, we set the number of cluster centers $k \in [4,5,6,7,8,9,10,11,12,13,14,15,16,17]$. At the same time, another parameter, the number of topics in the LDA model, is also an important hyperparameter, because different numbers of topics will affect the interpretability of the LDA model on documents and feature extraction. We set the number of topics $t \in [4,5,6,7,8,9,10]$ to verify the impact of different numbers of topics on the model. We use accuracy to search for optimal parameters and finally use the parameter setting of cluster center k=8 topic number t=4 to achieve an accuracy 68.67%.

D. Results

Experiment 1 :In this experiment, we use 4 machine learning models to report the classification results of 4 stages of PD based on multiple patient activities. Fig 6 shows the performance of our framework and the baseline based on instance recognition in terms of precision, recall, and accuracy. It can be seen that XGB shows better classification performance, and in our proposed framework, LDA features are added After that, it is obviously improved(The highest accuracy rate reaches 68.67%, which is 12% higher than the instance-based method and 6% higher than the MIL method without adding LDA features), which proves the effectiveness of our proposed framework, and in the follow-up data augmentation research, we choose XGB as our classification model to compare with other sample generation methods.

Experiment 2 :Table II shows the Xgboost classification performance after adding data augmentation components in our framework. At the same time, we compared random oversampling, smote, rotation and scaling. It can be seen that among all methods, the similar pair hybrid method we use achieves the best classification performance. Reached 73.49%, which is 4.82% higher than before data augmentation, and the average Recall is improved, which shows that our proposed method can increase the diversity of data very well, keep the features unchanged, and improve the classification performance of minority classes.

V. CONCLUSION

In this work, we attempt to assess Parkinson's disease stages using a single wearable sensor worn on the right hand. And it was verified on the real 83-person PD dataset, but we found two problems: weak labeling and data imbalance. In order to

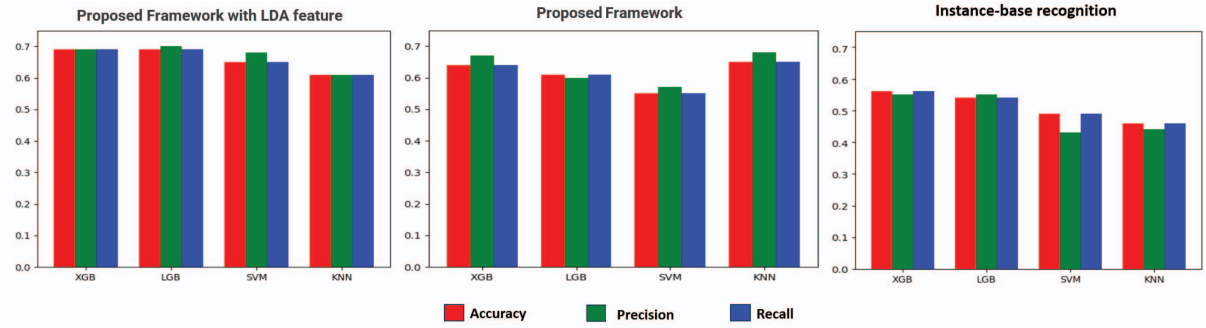


Fig. 6. Performance results of real PD dataset

solve these two problems, we proposed a PD stage evaluation framework represented by symbols, and used The topic model was developed to further obtain better feature expression. At the same time, We proposed a similarity-based pattern mixing method and the final result achieved an accuracy rate of 73.49%, which proves that the framework can overcome the influence of weak annotations and enrich data diversity. In future work, we will use more sensor data and more advanced methods to improve our classification accuracy. Overall, this study contributes to more accurate PD self-diagnosis in the wild, enabling remote guidance for drug intervention from doctors.

VI. ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No. 62061050). We thank Yunnan First People's Hospital for its strong support.

REFERENCES

[1] M. McHenry, "Symptoms and possible causes cures for parkinsons disease," *Brain Matters*, vol. 3, no. 1, pp. 8–10, 2021.

[2] E. a. Dorsey, R. Constantinescu, J. Thompson, K. Biglan, R. Holloway, K. Kieburtz, F. Marshall, B. Ravina, G. Schifitto, A. Siderowf *et al.*, "Projected number of people with parkinson disease in the most populous nations, 2005 through 2030," *Neurology*, vol. 68, no. 5, pp. 384–386, 2007.

[3] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for parkinson disease," *Archives of neurology*, vol. 56, no. 1, pp. 33–39, 1999.

[4] 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: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[5] G. AlMahadin, A. Lotfi, E. Zysk, F. L. Siena, M. M. Carthy, and P. Breedon, "Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms-a patient and healthcare professional perspective," *BMC neurology*, vol. 20, no. 1, pp. 1–13, 2020.

[6] E. G. Spanakis, D. Kafetzopoulos, P. Yang, K. Marias, Z. Deng, M. Tsiknakis, V. Sakkalis, and F. Dong, "Myhealthavatar: Personalized and empowerment health services through internet of things technologies," in *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 2014, pp. 331–334.

[7] J. Qi, P. Yang, M. Hanneghan, D. Fan, Z. Deng, and F. Dong, "Ellipse fitting model for improving the effectiveness of life-logging physical activity measures in an internet of things environment," *Iet Networks*, vol. 5, no. 5, pp. 107–113, 2016.

[8] P. Yang, G. Yang, J. Liu, J. Qi, Y. Yang, X. Wang, and T. Wang, "Duapm: An effective dynamic micro-blogging user activity prediction model towards cyber-physical-social systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5317–5326, 2019.

[9] D. Rodríguez-Martín, A. Samà, C. Pérez-López, A. Català, J. M. Moreno Arostegui, J. Cabestany, À. Bayés, S. Alcaine, B. Mestre, A. Prats *et al.*, "Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer," *PloS one*, vol. 12, no. 2, p. e0171764, 2017.

[10] L. Lonini, A. Dai, N. Shawen, T. Simuni, C. Poon, L. Shimanovich, M. Daeschler, R. Ghaffari, J. A. Rogers, and A. Jayaraman, "Wearable sensors for parkinson's disease: which data are worth collecting for training symptom detection models," *NPJ digital medicine*, vol. 1, no. 1, p. 64, 2018.

[11] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 216–220.

[12] S. Del Din, A. Godfrey, C. Mazzà, S. Lord, and L. Rochester, "Free-living monitoring of parkinson's disease: Lessons from the field," *Movement Disorders*, vol. 31, no. 9, pp. 1293–1313, 2016.

[13] K. Shima, T. Tsuji, A. Kandori, M. Yokoe, and S. Sakoda, "Measurement and evaluation of finger tapping movements using log-linearized gaussian mixture networks," *Sensors*, vol. 9, no. 3, pp. 2187–2201, 2009.

[14] A. J. Espay, J. P. Giuffrida, R. Chen, M. Payne, F. Mazzella, E. Dunn, J. E. Vaughan, A. P. Duker, A. Sahay, S. J. Kim *et al.*, "Differential response of speed, amplitude, and rhythm to dopaminergic medications in parkinson's disease," *Movement Disorders*, vol. 26, no. 14, pp. 2504–2508, 2011.

[15] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

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

[17] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1713–1721.

[18] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

[19] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 216–220.

[20] J. F. A. Otero, K. López-de Ipina, O. S. Caballer, P. Marti-Puig, J. I. Sánchez-Méndez, J. Iradi, A. Bergareche, and J. Solé-Casals, "Emd-based data augmentation method applied to handwriting data for the diagnosis of essential tremor using lstm networks," *Scientific Reports*, vol. 12, no. 1, p. 12819, 2022.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.