Multiple-Instance Learning for In-The-Wild Parkinsonian Tremor Detection

Alexandros Papadopoulos[1], Konstantinos Kyritsis[1], Sevasti Bostanjopoulou[2],
Lisa Klingelhoefer[3], Ray K. Chaudhuri[4], Anastasios Delopoulos[1],

***Abstract*— Parkinson's Disease (PD) is a neurodegenerative disorder that manifests through slowly progressing symptoms, such as tremor, voice degradation and bradykinesia. Automated detection of such symptoms has recently received much attention by the research community, owing to the clinical benefits associated with the early diagnosis of the disease. Unfortunately, most of the approaches proposed so far, operate under a strictly laboratory setting, thus limiting their potential applicability in real world conditions. In this work, we present a method for automatically detecting tremorous episodes related to PD, based on acceleration signals. We propose to address the problem at hand, as a case of *Multiple-Instance Learning*, wherein a subject is represented as an unordered bag of signal segments and a single, expert-provided, ground-truth. We employ a deep learning approach that combines feature learning and a learnable pooling stage and is trainable end-to-end. Results on a newly introduced dataset of accelerometer signals collected in-the-wild confirm the validity of the proposed approach.**

I. Introduction

Parkinson's Disease is characterized by a set of motor symptoms, including tremor, bradykinesia and rigidity, as well as non-motor symptoms, such as depression, constipation and sleep disorders. In particular, tremor, bradykinesia and rigidity have been characterized as cardinal symptoms, meaning that their co-existence is sufficient for an accurate diagnosis of the disease [1]. Early diagnosis of PD can prove beneficial to the patient as it enables a more efficient treatment of the symptoms at the earlier stages of the disease, thus ensuring a higher quality of life in the coming years [2]. Therefore, automatic detection of PD-related symptoms is a research direction that holds much promise.

One particular symptom that has received much attention from the research community is that of tremor. Parkinsonian tremor consists of rhythmic shaking in hands and other body extremities with a typical frequency in the range of $3-7$Hz. PD tremor can be classified into two main categories: *Resting tremor* which occurs when the muscles are relaxed and *action tremor* which occurs when voluntary muscle movement takes place.

In recent years, many methodologies have been proposed to automatically detect the presence of either type of tremor in cases associated with PD. Most works employ *Inertial Measurement Unit (IMU)* sensors, either as standalone devices or embedded in consumer electronics like smartphones and smartwatches, owing to the wide availability of such devices. The authors of [3], for example, use a commercial accelerometer and propose the use of Empirical Mode Decomposition coupled with Support Vector Machines, to differentiate between Parkinsonian and essential tremor. A different approach was proposed in [4], where the authors use gyroscope data and perform regression modelling to estimate the severity of tremor, while [5] highlights the potential of using the accelerometer sensors embedded in modern smartphones to record and evaluate tremor episodes.

More recently, the idea of using a smartphone as a measuring device for PD was widely adopted by *i-PROGNOSIS* [6], a European Horizon 2020 project, whose aim is to develop tools that will detect the early onset of PD, based on data collected via smartphone. More specifically, [6] introduces a multi-modal approach, where many data sources, including IMU, voice and typing patterns, are collected in-the-wild and subsequently used to detect a variety of PD symptoms, such as tremor, bradykinesia, rigidity and voice degradation. An important benefit of the approach is that data collection takes place unobtrusively, meaning that users need only install the data collection application.

In this work, we deal with the problem of detecting PD tremor in an in-the-wild setting. Tremor detection in that setting has not been thoroughly addressed by the research community, mainly due to the lack of appropriate datasets. To that end, we introduce a new dataset that contains accelerometer recordings, captured under entirely unscripted and unsupervised conditions, via the smartphone application of [6].

In contrast to other PD symtoms, tremor has an intermittent nature, meaning that it exhibits unpredictable on and off periods. This implies that we cannot associate the tremor ground truth of a subject, provided by a domain expert, with all the sessions (i.e. recordings) that subject contributed, as that would lead to extreme amounts of label noise and would mislead the training procedure of our models. To mitigate this issue, we view the problem at hand through the prism of *Multiple-Instance Learning (MIL)*.

Multiple-instance learning is a type of supervised learning, where the learner is presented with objects called *bags*. Each bag is essentially a set that contains multiple data instances. In contrast to the standard supervised setting, where annotations for all the instances available, ground-truth in the MIL setting is provided at the bag level. Although each individual instance may in fact admit to an annotation of its own, that is assumed to be unknown.

We can address the problem of detecting PD tremor

[1]Multimedia Understanding Group, Information Processing Laboratory, Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece [2]Department of Neurology, Hippokration Hospital, Thessaloniki, Greece [3]Department of Neurology, Technical University of Dresden, Dresden, Germany [4]International Parkinson Excellence Research Centre, King's College Hospital NHS Foundation Trust, London, UK

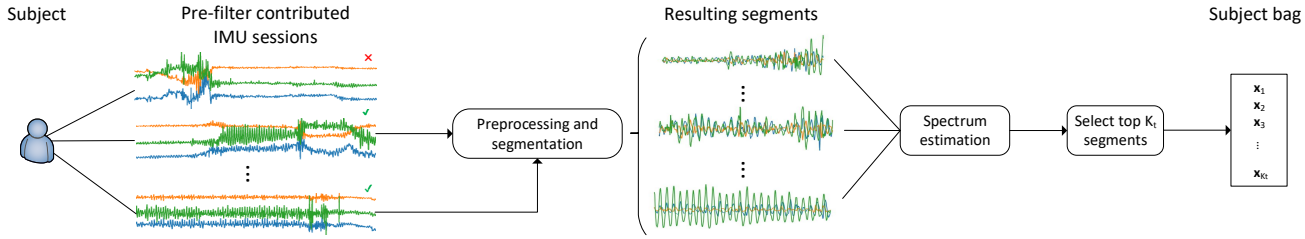

Fig. 1: Overview of the bag creation process. A single bag, that contains segments of contributed accelerometer signals, is created for each subject. The resulting bag is associated with a single tremor label, provided by domain experts.

in the-wild, as a case of multiple-instance learning. MIL naturally accommodates the lack of annotations for the individual sessions contributed by the subjects and can be used to properly model the intermittent nature of tremor: each subject can be described by a bag containing segments from their accelerometer recordings and a single tremor ground truth, provided by experts. We propose to model the tremor probability of a subject, given their bag of segments, by adopting a recently proposed deep multi-instance learning approach based on the attention mechanism [7].

Training and evaluation of the proposed method was performed on our newly introduced tremor dataset, which contains in-the-wild accelerometer data from 37 subjects, leading to encouraging early results.

II. Methodology

In the standard supervised learning setting, we aim at finding a mapping $f : \mathbb{R}^N \to \mathcal{Y}$. Depending on the nature of $\mathcal{Y}$, we either deal with a regression ($\mathcal{Y} = \mathbb{R}$) or a classification problem ($\mathcal{Y} = \mathbb{Z}$). The mapping is learned through a training procedure that minimises some appropriate cost function, on a given set of instances $\mathbf{x}_k \in \mathbb{R}^N$ and their corresponding targets $y_k \in \mathcal{Y}$. In the multiple-instance learning setting, instead of individual instances $\mathbf{x}_k$, we are presented with bags of instances $X_j = \{\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jK_j}\}$. The goal here is to find a mapping that associates a bag with a single label, that is, a mapping $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = 2^{\mathbb{R}^N}$ is the power set of $\mathbb{R}^N$. We assume that although the individual instances $\mathbf{x}_{jk}$ of a bag have labels $y_{jk} \in \mathcal{Y}$ of their own, they are unknown and only a single label y_j, which characterizes the whole bag, is available. In the following, we will limit our discussion to the case where $\mathcal{Y} = \{0, 1\}$, i.e. the case of binary classification of non-tremor vs tremor.

A recent approach, put forward in [8], is to model the bag label probability $p(y|X)$, using a three-step transformation:

$$p_{model}(y|X) = g\left(\underset{\mathbf{x}\in X}{\sigma}(f(\mathbf{x}))\right) \tag{1}$$

where

i) $f : \mathbb{R}^N \to \mathbb{R}^M$ transforms each instance $\mathbf{x}_i$, to a low-dimensional embedding of size M.

ii) $\sigma : 2^{\mathbb{R}^M} \to \mathbb{R}^M$ is a permutation-invariant pooling function that produces a fixed-length representation.

iii) $g : \mathbb{R}^M \to \mathcal{Y}$ transforms the pooled representation to the final bag label probability.

The choice of functions f, σ, g leads to different approaches to the MIL problem. For instance, if f is the identity function and σ is the histogram of codebook assignments, equation (1) reduces to a Bag of Features approach [9]. In this work, we use neural networks to parameterise the transformations f, g and the attention-based pooling mechanism proposed in [7], to model the pooling operator σ. More specifically, let $H = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K\}$ be a bag of K embeddings resulting from the elementwise application of the embedding function f to the original bag X (the j index is ommited for simplicity). The function σ is defined as a non-linear combination of the instance embeddings:

$$\mathbf{z} = \sigma(H) = \sum_{k=1}^{K} a_k \mathbf{h}_k \tag{2}$$

where

$$a_k = \frac{\exp\left(\mathbf{w}^T \tanh\left(\mathbf{V}\mathbf{h}_k^T\right)\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{w}^T \tanh\left(\mathbf{V}\mathbf{h}_k^T\right)\right)} \tag{3}$$

The non-linearity of the pooling operator, not apparent at first sight, arises from the way the quantities a_k are computed: the weight assigned to each $\mathbf{h}_k$ depends on its value as well as the learnable parameters $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$. The construction of equation (3) draws inspiration from the attention mechanism, that is widely used in the context of machine translation tasks [10]. In the context of MIL, it can be used to discover key instances within a bag, thus resulting in a bag representation that is useful to the final classifier g.

We propose to use the attention-based pooling mechanism described above, in order to perform Parkinsonian tremor detection in-the-wild. The rest of this section describes the process of creating bags of instances and the training procedure used for bag classification. Specific details about the model architecture are given in section IV.

A. Bag creation

Each subject in the dataset contributed one tri-axial accelerometer session for each phone call they had realised during the data collection period. Each such session is expected to be of variable length depending on the duration of the call. To overcome this issue while keeping things as simple as possible, we segment each session into non-overlapping windows of fixed-length W, and construct a bag

out of the resulting set of windows. In doing so, we explicitly choose not to model any intra-session dependencies between neighbouring segments. Instead, we represent the subject as an unordered bag of signal segments.

Let $S = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ denote the set of windows contributed by a subject, where $\mathbf{w}_i \in \mathbb{R}^{W\times 3}$ and N denotes the total number of windows contributed by that subject. In order to limit the size of the bag without sacrificing its descriptiveness, we perform a ranking of the windows based on their energy in the band of $[3, 7]$ Hz (the tremor band) and keep the top K_t windows. In addition, due to the periodic nature of tremor, we further transform the signal segments from the time to the frequency domain, using Welch's method for spectral density estimation, and keep the spectral coefficients for the frequency band $[0, 25]$ Hz. The overall process leads to a bag of constant-size, that for all subjects has the form $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{K_t}\}$ where $\mathbf{x}_k \in \mathbb{R}^{F\times 1}$ and F denotes the number of spectral coefficients kept. A schematic overview of the process is given in Fig 1.

B. Model training

The bags acquired through the above procedure, with one bag corresponding to one subject, can now be associated with the available tremor annotations, provided by the clinical experts. The resulting set of tuples (X, y), with y denoting the label of the bag, defines the empirical data distribution $\hat{p}_{data}$ on which we will train the model of equation (1).

We use fully-connected neural networks (described in detail in section IV) to model the functions f, g, σ. The whole model, consisting of the composition of these 3 functions as defined in equation 1, is trained in an end-to-end manner for E epochs, using the standard cross entropy loss:

$$\mathcal{L} = \underset{X,y\sim\hat{p}_{data}}{-\mathbb{E}} [\hat{p}_{data}(y|X) \log(p_{model}(y|X))] \qquad (4)$$

After training, the model can be applied to a new, previously unseen subject in order to estimate its tremor probability. The probability estimate can then be binarised using a threshold T, to acquire a final tremor prediction.

III. Dataset

A. Collection

In this work, we use a dataset of IMU recordings that was collected[1] outside laboratory conditions, in the context of [6]. Collection was performed through an unobtrusive smartphone application that initiated capturing of the smartphone's IMU sensors whenever a phone call took place. Sensor recording lasted at most for 75 seconds, due to battery constraints. The application was distributed to a pool of users who used it throughout their daily lives for a period ranging from a few days to several months. Each user thus contributed a varying number of IMU sessions to the dataset, depending on the number of phone calls they realised during the time the application was installed on their phone. In the following, we focus only on signals collected via the widely-available accelerometer sensor.

[1]Data collection was approved by the Institutions Ethical Review Board and subjects provided electronic consent

B. Pre-processing

A common pre-processing pipeline was applied to all the collected signals to remove any discrepancies caused by the recording conditions. First, we performed a pre-filtering step that discarded any sessions deemed as problematic. The criteria for rejecting a session where i) Short duration ii) Low sampling frequency iii) Low signal energy iv) Existence of extreme values. Each surviving session was subsequently resampled to a common sampling frequency of 100 Hz. A segment of 5 seconds was then trimmed from the start and the end of the signal, to remove the transition phenomena from moments when the user either picks up or hangs up the phone. Finally, a high-pass filter with a cutoff frequency of 1 Hz was applied to remove the gravitational component from the acceleration signal.

Due to the intermittent nature of tremor, we imposed an additional restriction on the quantity of data a user must have contributed in order to be used in our experiments. More specifically, we stipulated that each user contribute IMU data of at least 2.30 minutes in total, after the pre-filtering step described above. Subsequently, users whose data contribution did not meet that minimum requirement were not considered. This ultimately resulted in a dataset of 37 subjects with a total contribution of ~ 35 hours of acceptable data.

C. Annotation

The *Unified Parkinson's disease rating scale (UPDRS)* is an established rating scale used by clinicians to quantify the intensity of PD. UPDRS evaluations at successive points in time, allow for a longitudinal perspective of the disease's progress. The standard UPDRS exam contains a self-reported questionnaire, where the subject provides an estimation of their symptom severity under daily living conditions, and a motor examination part, where the physician quantifies the intensity of their symptoms at the moment of the examination.

Each user in our dataset underwent a thorough clinical examination, including a full UPDRS evaluation as well as an aggregated PD symptom history from past evaluations. This resulted in two potential sources of annotation for our data: the self-report provided by the patient and the medical history provided by the doctor. However, both the self-report and the tremor history provide only a rough indication of whether the patient exhibits tremor in *any body extremity*. This may lead to extreme label noise, since a user may have tremor only in the right hand but use the left hand when making a phone call. Therefore, all their contributed sessions would be tremor-less but we still would consider them as tremorous, owing to the lack of more detailed annotation. To mitigate this issue, we additionally performed manual annotation of each subject. To that end, a group of signal processing experts used the subjects whose medical history indicated high tremor intensity, to acquire a sense of how tremor exhibits in an IMU signal. Having done that, the rest

of the subjects were annotated by individually inspecting the raw accelerometer signal of each session as well as its power spectrum, and taking into account both the self-report and the tremor history.

IV. Experiments & Results

Training and evaluation of the proposed method was performed by employing a *Leave One Subject Out (LOSO)* scheme. We performed 3 different experiments, one for each available type of annotation. In each experiment, we used the signal processing experts' annotations for training and evaluated the resulting model using each of the available annotations (self-report, medical history, signal experts).

We used a window length W of 5 seconds for the signal segmentation procedure. Each resulting bag contained at most $K_t = 1500$ segments. For the instance embedding function f, we used a network with 3 fully-connected layers of $256, 128, 64$ units respectively, coupled with the Leaky-ReLU non-linearity with 0.2 negative slope and dropout with drop probability $p = 0.5$. Similarly, for the final classifier g, we use 3 fully-connected layers with $32, 16, 2$ units respectively, along with the Leaky-ReLU activation with 0.2 negative slope and dropout with $p = 0.2$, in all but the final layer. The attention pooling function σ, was parameterised by a 2-layer network that implemented equation (3), with the attention dimension, L, set to 16. The model was trained end-to-end for $E = 500$ epochs using the Adam optimizer with learning rate $\epsilon = 0.0005$ and exponential decay after the first 250 epochs by a factor of 0.9. Finally, the decision threshold, T, was set to 0.5.

To compare against other methods, we used two standard pooling algorithms under the same experimental setup. More specifically, we used the *Bag of Features (BoF)* algorithm, with a codebook size of 128, and the *Fisher Vector (FV)* encoding scheme [9] with 32 modes. Each algorithm was used to encode the bag of segments that characterised each subject. A *Support Vector Machine (SVM)* was then used to classify the resulting bag encodings. The chi-square kernel was used for the BoF encoding (due to its histogram nature), while the linear kernel was employed for the FV encoding. The C hyperparameter of the SVM was set to 1 and balanced weights were used for each class.

Each experiment was repeated 5 times to account for random initialisation issues. The average performance of each method over the 5 trials is given in Tables I, II, III.

TABLE I: Evaluation results using annotations provided by signal processing experts

Model	Precision	Sensitivity	Specificity	F-score
Deep MIL	0.893	0.763	0.961	0.851
BoF + SVM	0.888	0.291	0.984	0.438
FV + SVM	0.552	0.763	0.738	0.750

Based on these results, we can see that the deep MIL approach leads to the best performance under almost all evaluation schemes. In particular, when evaluating on the annotations provided by the signal processing experts (which can be considered as the most reliable due to reasons discussed in section III-C), the MIL approach outperforms the alternatives by a large margin. This trend persists on the other evaluation schemes as well, suggesting that the attention-based pooling method can accurately identify tremor-related instances within the subject bag, even at the presence of label noise (as indicated by tables II, III).

TABLE II: Evaluation results using annotations provided by medical experts

Model	Precision	Sensitivity	Specificity	F-score
Deep MIL	0.863	0.422	0.936	0.582
BoF + SVM	0.863	0.211	0.968	0.339
FV + SVM	0.584	0.500	0.663	0.570

TABLE III: Evaluation results using annotations provided by the subject self-report

Model	Precision	Sensitivity	Specificity	F-score
Deep MIL	0.791	0.399	0.888	0.551
BoF + SVM	0.937	0.158	0.988	0.270
FV + SVM	0.612	0.515	0.655	0.577

V. Conclusions

We have presented a method for automatically detecting Parkinsonian tremor from accelerometer data using a deep multiple-instance learning approach. Early results are promising and indicate that the method can adequately handle the noisy and unpredictable nature of signals obtained in-the-wild, thus highlighting its potential in detecting tremor in a general population of users.

VI. Acknowledgments

The work leading to these results received funding from the EU Commission under Grant Agreement No. 690494 (http://www.i-prognosis.eu, H2020).

References

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] G. Becker *et al.*, "Early diagnosis of parkinson's disease," *Journal of Neurology*, vol. 249, no. 3, pp. iii40–iii48, Oct 2002.

[3] L. Ai *et al.*, "Classification of parkinsonian and essential tremor using empirical mode decomposition and support vector machine," *Digit. Signal Process.*, vol. 21, no. 4, pp. 543–550, July 2011.

[4] J. A. Gallego *et al.*, "Real-time estimation of pathological tremor parameters from gyroscope data," *Sensors*, vol. 10, no. 3, pp. 2129–2149, 2010.

[5] J.-F. Daneault *et al.*, "Using a smart phone as a standalone platform for detection and monitoring of pathological tremors," *Frontiers in human neuroscience*, vol. 6, p. 357, 12 2012.

[6] *i-PROGNOSIS: Towards an early detection of Parkinson's disease via a smartphone application*. Zenodo, Sept. 2017.

[7] M. Ilse *et al.*, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[8] M. Zaheer *et al.*, "Deep sets," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3391–3401.

[9] J. Sánchez *et al.*, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[10] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.