Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

PD Disease State Assessment in Naturalistic Environments Using Deep Learning

Nils Y. Hammerla
Culture Lab, Digital Interaction Group
Newcastle University, UK
nils.hammerla@ncl.ac.uk

James M. Fisher
Health Education North East, UK

Peter Andras
School of Computing and Mathematics
Keele University, UK

Lynn Rochester
Institute of Neuroscience
Newcastle University, UK

Richard Walker
Northumbria Healthcare
NHS Foundation Trust, UK

Thomas Plötz
Culture Lab, Digital Interaction Group
Newcastle University, UK

Abstract

Management of Parkinson's Disease (PD) could be improved significantly if reliable, objective information about fluctuations in disease severity can be obtained in ecologically valid surroundings such as the private home. Although automatic assessment in PD has been studied extensively, so far no approach has been devised that is useful for clinical practice. Analysis approaches common for the field lack the capability of exploiting data from realistic environments, which represents a major barrier towards practical assessment systems. The very unreliable and infrequent labelling of ambiguous, low resolution movement data collected in such environments represents a very challenging analysis setting, where advances would have significant societal impact in our ageing population. In this work we propose an assessment system that abides practical usability constraints and applies deep learning to differentiate disease state in data collected in naturalistic settings. Based on a large data-set collected from 34 people with PD we illustrate that deep learning outperforms other approaches in generalisation performance, despite the unreliable labelling characteristic for this problem setting, and how such systems could improve current clinical practice.

Introduction

Parkinson's Disease (PD) is a degenerative disorder of the central nervous system that affects around 1% of people over 60 in industrialised countries (de Lau and Breteler 2006). People affected by PD show a variety of motor features that gain in severity with the progression of the disease, which include rigidity, slowness of motion, shaking and problems with gait [among others]. The severity and nature of these motor features vary over the course of the day, which has a significant impact on the quality of life of people with PD. Management of the condition relies on tailored treatment plans that provide a specific schedule for the type and dosage of a multitude of medications taken by each individual. Devising such treatment plans is a challenge as clinical consultations may be infrequent and only provide a snapshot of the condition, which may not give an adequate picture of the daily fluctuations beyond recall by the individual. Objective, automated means to assess PD in people's daily lives are therefore much desired.

In order to become a useful clinical tool, such automated assessment systems have to be deployed in naturalistic, ecologically valid surroundings such as the private home. Systems based on, or evaluated in, such naturalistic settings are, however, very rare. The reason for this apparent shortcoming is clear: while capturing data in naturalistic environments is straight-forward using e.g. body-worn movement sensors, obtaining reliable labels useful for system development is practically difficult, if not impossible, as even trained annotators would show only modest agreement with experts (Palmer et al. 2010). In practice only unreliable and infrequent labels can be obtained in such surroundings, for example using symptom diaries kept by each participant. Instead of addressing this issue, current systems for the assessment of PD rely on data captured in the laboratory, where daily life is just simulated (Hoff and others 2001), or attempt to recreate the laboratory in the private home using e.g. movement tasks under remote supervision by a clinician (Giuffrida et al. 2009). Research on PD is missing adequate tools that would allow data from ecologically valid surroundings to be exploited in system development, as this problem with its unique challenges, has received little attention from the machine learning community. This represents a significant barrier for practical assessment systems, overcoming which may dramatically improve the quality of life of people affected by PD.

In this paper we investigate the problem of predicting the disease state in PD patients in naturalistic surroundings, i.e. the daily life of individuals affected by PD. We illustrate that assessment systems have to overcome significant challenges in analysing large quantities of ambiguous, low-resolution multi-variate time-series data for which only infrequent and unreliable labels can be obtained due to practical usability constraints. Labels are subject to various sources of noise such as recall bias, class confusion and boundary issues and do not capture the main source of variance in the data, as people engage in many (unknown) physical activities that have significant effect on the recorded sensor signals. Based

on a large data-set that contains approx. $5,500$ hours of movement data collected from 34 participants in realistic, naturalistic settings, we compare how deep learning and other methods are able to cope with the characteristic label noise. We find that deep learning significantly outperforms other approaches in generalisation performance, despite the unreliability of the labelling in the training set. We show how such systems could improve clinical practice and argue that such a setting could serve as a novel test-bed for unsupervised or semi-supervised learning, where improvements would have significant societal impact.

Assessing disease state in naturalistic surroundings

The quality of life of people with PD is significantly affected by fluctuations in the severity of the disease. Periods where motor symptoms (such as tremor or bradykinesia) are more prominent are typically referred to by clinicians and patients as "*off* time". Conversely, periods where motor symptoms are well controlled are referred to as "*on* time". As the condition progresses, motor fluctuations between these differing *disease states* become more frequent and less predictable. Furthermore, prolonged medication usage is associated with the development of additional involuntary movements known as *dyskinesia*. Tailored treatment plans aim to reduce the severity of these fluctuations. In this work we focus on the assessment of disease state in PD, as it represents a crucial component for improved management of the condition in clinical practice.

In order to be useful for clinical practice, such assessment systems have to be applicable in naturalistic, ecologically valid surroundings such as the private home. Yet research on automated assessment in PD generally does not address this issue. Systems are based on laboratory environments, where participants engage in a series of movement tasks that are part of the clinical assessment procedure in PD (Goetz et al. 2008; Patel et al. 2009). With extensive instrumentation of the participants those system achieve very good results in e.g. detecting dyskinesia with more than 90% accuracy (Tsipouras et al. 2010). However, such scripted movement tasks, even if extended to include activities of daily living to simulate daily live (Hoff and others 2001), are only a very poor model of naturalistic behaviour. Some systems aim to re-create the clinical assessment in the daily life of a subject while being supervised by a clinician (remotely), effectively simulating such laboratory conditions in naturalistic surroundings (Mera et al. 2012; Giuffrida et al. 2009). Whether individual assessment systems generalise to naturalistic environments is rarely explored, where a recent review just found 3 out of 36 studies to include data recorded in naturalistic settings, although the authors specifically focussed on this aspect (Maetzler et al. 2013). Even where naturalistic data is gathered it is not utilised during system development, but instead being used to gain some insight into the performance of systems based on medical prior knowledge (e.g. (Griffiths et al. 2012; Hoff, Van Der Meer, and Van Hilten 2004)). This laboratory-driven research represents a significant barrier towards practical assessment systems.

The reliance on controlled laboratory environments stems from the difficulties encountered when collecting and exploiting data from in naturalistic surroundings. This issue is split into two aspects. The most pressing concern from a machine learning perspective relates to obtaining ground-truth information about disease state in PD in naturalistic environments. Even if e.g. video recordings can be obtained, which is unlikely, it can be difficult for annotators to assess the disease state with high reliability (Palmer et al. 2010). Instead, labels have to be obtained in cooperation with the patients, where the common best practice are disease state diaries (Reimer et al. 2004). Such diaries just provide an infrequent (e.g. one sample per hour) and unreliable impression of the disease state. Participants may have trouble identifying their own disease state, or fill out the diary retrospectively (recall bias). Additionally, the disease characteristics evolve gradually and are unlikely to change exactly on the hour, leading to issues at the boundaries of the provided labelling.

The second aspect relates to usability aspects of the sensing system. Recording (unlabelled) data in naturalistic settings is straight-forward if sensing solutions require little cooperation by the patient, do not rely on external infrastructure, abide by privacy constraints, and follow a suitable physical design of the devices (McNaney et al. 2011). Any practical sensing system will necessarily be a compromise between the obtainable sensing resolution (e.g. degrees of freedom, number of sensors, ambiguity of recordings) and abiding usability constraints of the target population to maximise compliance. The most suitable sensing approach for naturalistic deployments are small body-worn movement sensors (Maetzler et al. 2013), which capture multi-variate time-series data that give an impression of the participant's physical activity and overall behaviour with a large amount of noise and inherent ambiguity. The main sources of variance in this movement data are the physical activities that participants engage in, such as walking, and not the overall disease state. The disease state rather has an effect on how activities are performed (e.g. "slower" while *off*). However, the activities that participants engage in are unknown, as collecting additional activity logs to gain an impression of the physical activities of participants would be too burdensome for longitudinal settings, particularly if participants suffer from cognitive decline. This also renders approaches such as *active learning* (e.g. (Stikic, Van Laerhoven, and Schiele 2008)) difficult to apply for this population, as they also require significant cooperation by the individual.

We can summarise the challenges for exploiting naturalistic data in this setting as follows: *i)* There is a significant disparity between the frequency at which data is collected (e.g. 100Hz) and the accessible labelling (e.g. one per hour); *ii)* The participant-provided labelling is inherently unreliable, subject to recall bias, class confusion and boundary issues; *iii)* The recorded data mostly reflects unknown activities, across which an assessment system has to generalise to obtain an impression of disease state in PD. Addressing these challenges through methodological advances would have significant impact on clinical practice for PD and other degenerative conditions where assessment faces

similar issues.

System overview

In response to these challenges we develop a novel approach to the assessment of disease state in PD. Instead of basing the development of our approach on data collected in a laboratory setting, we exploit large amounts of data gathered in naturalistic surroundings, the daily life of people affected by PD. In many applications of machine learning it is easy to obtain large amounts of unlabelled data, and devising systems capable of exploiting such data to improve recognition performance has become a popular field in machine learning. One approach that has been shown to be effective for e.g. phoneme recognition (Deng, Hinton, and Kingsbury 2013) and object recognition (Lee et al. 2009) is *deep learning*, where unlabelled data is used to greedily initialise multiple layers of feature extractors. In this work we apply deep learning to the problem of disease state assessment in PD to explore if these methods can cope with the unreliable labelling that results from naturalistic recording environments.

Our system comprises a typical analysis pipeline common for activity recognition in ubiquitous computing (Bulling, Blanke, and Schiele 2014). First the captured data is segmented using a sliding window procedure, after which a hand-crafted set of features is extracted from each frame. In cross-validation experiments these features are then used to train a sequence of Restricted Boltzmann Machines (RBMs) (Hinton, Osindero, and Teh 2006). A softmax top-layer is added to the trained generative model which is further fine-tuned using conjugate gradients to maximise classification performance.

Wearable sensing system

Our sensing setup consists of two movement sensors, one worn on each wrist of the participant, which have been used in previous applications such as in Autism research (Plötz et al. 2012), and sports (Ladha et al. 2013). The movement sensors contain a tri-axial accelerometer that measures acceleration along three perpendicular axes with high temporal resolution (100 Hz). On a single charge, such devices are able to capture acceleration data for up to 12 days. The sensors are attached using comfortable velcro straps and are waterproof. Colour coding ensured that the sensor location and orientation remained constant throughout the study. This sensing system represents a compromise between usability and signal quality. The small number of sensors in a convenient location along with their high usability allow data capture in the daily life of the participants with very high compliance. However, for the sake of prolonged battery life no further modalities beyond accelerometers were included (e.g. gyroscopes, magnetometer).

Data collection

Overall 34 participants were recruited who exhibited mild to severe level Parkinson's Disease (Hoehn and Yahr stages I-IV (Hoehn and Yahr 1998)), were not significantly cognitively impaired and were taking immediate-release levodopa medication. All participants provided informed consent for involvement and ethical approval was obtained from the relevant authorities. The study was split into two subsequent phases:

Phase 1 (LAB) consists of lab-based recordings. Participants attended a movement research laboratory without having taken their early morning dose of medication (where possible) and spent on average 4 hours in the facility while wearing the sensing system. At regular intervals (e.g. once per hour or more), the current state of the disease was assessed by a clinician. Based on video recordings, a second clinician rated the disease state for each examination. Assessments where the two clinicians disagreed where discarded (overall agreement > 0.95). Data is extracted surrounding each of the 141 remaining assessments. The assessment itself is removed as participants engage in a series of movements selected to assist within clinical evaluation but are highly unlikely to be representative of naturalistic behaviour. Data from phase 1 is denoted as LAB throughout the rest of this work.

Phase 2 (HOME) corresponds to longitudinal recordings in the participant's private homes. After completing phase 1, participants wore the sensing system continuously over the course of a week, including at night. Each participant filled out a disease state diary, a pre-formatted document where ticks indicate disease state for each hour, to the best of their abilities. The diary included: *asleep*, *off*, *on*, and (troublesome) *dyskinesia*. A total of approx. $5,500$ hours of accelerometer data was collected, for which approx. $4,500$ hourly labels were provided by the participants (80% diary compliance). The labels are inherently unreliable, as symptom characteristics are very unlikely to change exactly on the hour, participants may have trouble classifying their own disease state, and diaries may be filled out retrospectively at the end of the day. Data collected in phase 2 is denoted as HOME throughout the rest of this paper.[1]

Pre-processing and feature extraction

Each disease state is characterised by different expressions of the common motor features in PD. During the *off* state, people with PD feel slow, stiff and may show increased tremor. In the *on* state, symptoms are less severe and tremor may disappear completely. Bouts of dyskinesia present as somewhat repetitive involuntary movements that may involve the wrists. Crucially the recorded data does not just contain the expression of the disease states but includes (unknown) naturalistic physical activities that have significant effect on the recorded signal. We extract features from segmented accelerometer data, where each segment spans one minute in duration. In these relatively long segments we aim to even out the impact of physical activities and try to capture the underlying characteristics, expressed as differences in the distribution of the acceleration measurements.

From the raw recordings contained in each frame $f^t = (f_L^t, f_R^t) \in R^{n \times 6}$ the acceleration magnitudes for each sensor m_L, m_R are estimated which are subsequently filtered

[1] Data-set is available at **http://di.ncl.ac.uk/naturalisticPD**.

using a high-pass filter with a cut-off frequency of 0.5Hz to remove the gravitational component. The filtered magnitudes are used to obtain their first derivatives (jerk) j_L, j_R. The magnitude of orientation change c_L, c_R is calculated from the raw recordings of each sensor as follows:

$$c_L \;=\; \left\{\cos^{-1}\left(f_{L,i}\cdot f_{L,i+1}\right)\right\}_{i=1\cdots(n-1)}, \qquad (1)$$

where $(\cdot)$ denotes the vector dot-product, $f_{L,i} \in \mathbf{R}^3$ are the recordings of sensor L at position i (relative time within frame). c_R is calculated accordingly for the sensor on the right wrist. Based on m_L and m_R we estimate the power spectral density p_L, p_R using a periodogram on 10 frequency bands between 1 and 8 Hz to capture repetitive movements typical for motor features in PD.

We capture the statistical characteristics of the movement within a frame using the ECDF representation introduced in (Plötz, Hammerla, and Olivier 2011; Hammerla et al. 2013), which corresponds to concatenated quantile functions along with their mean. For each frame we obtain its feature representation x^t by concatenating the ECDF representations of the acceleration magnitudes m_L, m_R, jerk j_L, j_R, orientation change c_L, c_R and power spectral density p_L, p_R. We further include the *time spent not moving* (threshold on $c_L + c_R$) as in (Griffiths et al. 2012), energy, minimum, maximum, standard deviation of m_L, m_R and binary PD phenotype (*tremor-dominant*). Using 10 coefficients in the ECDF representation we extract a total of 91 features from each minute of sensor recordings. In the future, this hand-crafted feature extraction will be substituted with a convolutional architecture alleviating the need for medical prior knowledge.

Training procedure

The training procedure comprises of two steps. First the real-valued features are normalised to have zero mean and unit variance (per fold in cross validation). We then apply RBMs to learn a generative model of the input features (Hinton, Osindero, and Teh 2006). After training the first RBM, the activation probabilities of its feature detectors are used as input data for the next RBM. This way, RBMs can be used to greedily initialise deep neural networks by adding more and more layers (Hinton and Salakhutdinov 2006). We learn at most two consecutive RBMs, where the first one contains gaussian visible units (gaussian-binary) to model the real-valued input features and the next one just contains binary units (binary-binary). Learning rates were set to 10^{-4} for the gaussian-binary RBM, and 10^{-3} for the binary-binary RBM, with a momentum of 0.9 and a weight-cost of 10^{-5}. Each RBM is trained for 500 epochs with batches containing 500 samples. Crucially, this first phase of training does not rely on any labels of the input data and is solely driven by the objective to learn a generative model of the training data.

In the subsequent fine-tuning phase we add a top-layer (randomly initialised, $\sigma = 0.01$) to the generative model. This top-layer contains 4 units in a softmax group that correspond to our 4 classes of interest: *asleep*, *off*, *on*, and *dyskinetic*. Using the labels for each input frame we perform 250 epochs of conjugate gradients with batches that gradually increase in size from 256 up to 2,048 (stratified) samples. In the first epoch the weights in all but the top layer remain fixed. Training time averages to around one day per fold on a GPU.

Experimental evaluation

Two scenarios are investigated in this work. In the first setting, a variety of approaches and network architectures are trained on the HOME data-set. To minimise the effect of large pairwise similarity of subsequent minutes of recording we follow a leave-one-*day*-out cross validation approach, where e.g. the first day of recording from all patients constitutes a fold. This represents a compromise between realistic assessment of generalisation performance and the required computational effort for training (which is extensive). The second setting simulates best practice for assessment systems in PD, where the smaller but clinician validated LAB data-set is used for training in a stratified 7-fold cross validation which is subsequently applied to the HOME data-set to assess generalisation performance.

In total the HOME data-set contains approx. $270,000$ samples (minutes) and the LAB data-set contains $1,410$ samples extracted from the recordings surrounding 141 individual disease-state assessments. Additional minutes are extracted for networks that span more than one minute in their input, such that the overall number of samples is retained. Since e.g. the HOME data-set is highly skewed towards *asleep* (31%) and *on* (41%) we chose the mean F1 score as primary performance metric:

$$\frac{2}{c}\sum_{i=1}^{c}\frac{\texttt{prec}_\texttt{i}\;\times\;\texttt{recall}_\texttt{i}}{\texttt{prec}_\texttt{i}\;+\;\texttt{recall}_\texttt{i}}, \qquad (2)$$

where $\texttt{prec}_\texttt{i}$ corresponds to the precision, $\texttt{recall}_i$ to the recall observed for class i and c to the number of classes. The LAB data-set does not contain any instances of *asleep* and the performance is evaluated just using the remaining three classes, even though false positives for *asleep* are included in the calculation of sensitivity and precision.

To illustrate the difficulty of the problem we compare the approach proposed in this work with standard classification methods typical for automated assessment systems in PD. We apply decision trees (C4.5), Naive Bayes (NB), and nearest neighbour classification(1-NN). We further apply support vector machines (SVM) with an rbf-kernel for training on the LAB data-set. On the HOME data-set we failed to achieve convergence to non-trivial solutions in SVMs. In order to investigate the impact of the layout of the deep ANN proposed in this work we evaluate a number of different network topologies the results of which are discussed below.

Results

Recognition results are illustrated in Figure 1. The left plot shows the results for approaches trained on the HOME data-set, while the right plot shows results for those trained on the LAB data-set. Labels indicate the method or network topology, e.g. "$5m-2\times1024$" translates to 5 minutes of input and two hidden layers with 1,024 units each. For each approach, three results are reported: i) the performance on individual

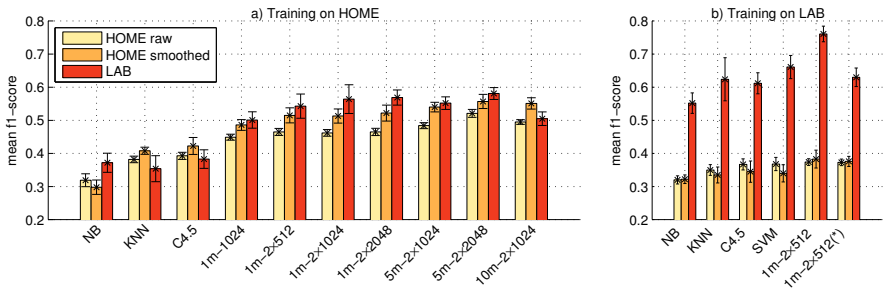

Figure 1: The left plot shows the performance of models trained on the HOME data-set, the right plot shows the performance for models trained on the LAB data-set. Errorbars indicate one standard deviation, estimated based on the performance of individual folds. Colour indicates the data-set used for evaluation. (*) indicates networks pre-trained on HOME and fine-tuned on LAB.

frames in the HOME data-set ("raw"), ii) smoothed predictions using a sliding window of 60m duration ("smoothed") and a step size of 20 minutes, and iii) the performance on the LAB data-set. The rationale behind smoothing the predictions over time is that in clinical applications the fluctuations would be assessed over longer time-frames, instead of being based on individual minute-by-minute predictions.

We first discuss systems trained on the HOME data-set. Overall the traditional approaches perform rather poorly in this setting. While smoothing slightly improves results, KNN, and C4.5 show a drop in performance on the LAB validation set. However, the various deep network topologies investigated here not only show significantly better performance than e.g. C4.5, but also (mostly) show a performance on the LAB validation set that exceeds the performance on the HOME data-set. We infer that the apparent increase in performance on the validation set illustrates the poor quality of the participant-provided labelling in the training-set, rather than an unexpected generalisation ability. Nevertheless it indicates that deep NNs are able to capture disease characteristics that remain inaccessible to more traditional methods, which may be reasoned in the gradient-based optimisation approach that implicitly placed a degree of weight on each sample. Normalised confusion matrices for the best performing network are illustrated in Figure 2. The class with the lowest performance on the HOME data-set is *dyskinesia*. Interestingly that class shows high specificity in the validated LAB data-set, indicating particularly unreliable labels for this disease state in the training-set. Overall adding a second layer and adding more units to the hidden layers improves the results, which is in line with previous results on this type of model. The best results are obtained for networks that span 5 minutes of input. If the input span is increased further to 10 subsequent minutes we see a drop in the performance on the validation set.

The results for systems trained on the LAB data-set differ strongly to those above. While the recognition performance on the LAB data-set (in cross-validation) is very good with peak mean f1-score of 0.76, the generalisation performance when applied to the HOME data-set is very disappointing. We further found no evidence that pre-training a model on the HOME data-set with subsequent fine-tuning on the validated LAB data-set provided significant improvement in generalisation performance (see "(*)" in Figure 1). Instead we see a decline in the cross validation performance, which supports our initial assumption that laboratory-based data is just an incredibly poor model for naturalistic behaviour.

Comparison to related approaches

When trained on the HOME data-set, we see a peak mean f1-score of 0.581 on the LAB data-set, which corresponds to an overall accuracy of 59.4%. On average the classes are differentiated with a sensitivity of 0.57 and a specificity of 0.88. It is difficult to compare these results to prior art, as no systems exist that follow a similar training and evaluation methodology. Hoff et al. (Hoff, Van Der Meer, and Van Hilten 2004) report very similar performance figures with sens. and spec. around 0.7 for *on* and *off* states over a 24h period on 15 participants with PD (compared to 4 states in this work). However, their sensing approach was based on a network of 7 sensors placed across the body and their prediction relied on thresholds set for each individual to maximise performance, effectively limiting the practicality of their approach. For a gold standard, consider that trained nurses may show relatively low accuracy of 0.65 when assessing the severity of motor complications (Palmer et al. 2010).

Systems trained on the LAB data-set show good performance in cross validation experiments up to a mean f1-score of 0.76. These results are comparable with other systems applied in laboratory settings (Maetzler et al. 2013). However, our results indicate that the poor generalisation to realistic behaviour of this artificial setting may also affect other systems based on similar laboratory environments, which has so far not been demonstrated.

Discussion

The quality of life of people affected by PD depends on the management of their condition in the form of tailored treatment plans. Devising such plans is a challenge, as objective information about fluctuations in disease state is not accessible in clinical practice beyond recall by the individual. Current best practice in automated assessment of PD is to obtain data in laboratory conditions, where small amounts of clinician-validated behaviour can be observed. While such systems show good performance in this setting, it is unlikely that they generalise to naturalistic behaviour in people's daily lives. In order to address this issue, assessment

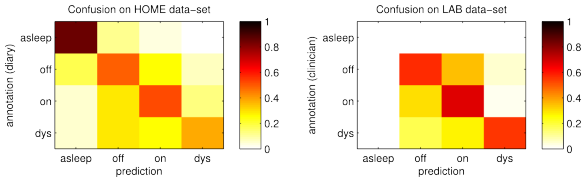

Figure 2: Normalised confusion matrices on both the (smoothed) HOME data-set (left) and on the Lab data-set (right) for a model with 5 minutes as input and two hidden layers with 2,048 units each. While the performance on the class *dys* is relatively low in the HOME data-set, there are just few false positives for this class in the laboratory setting.

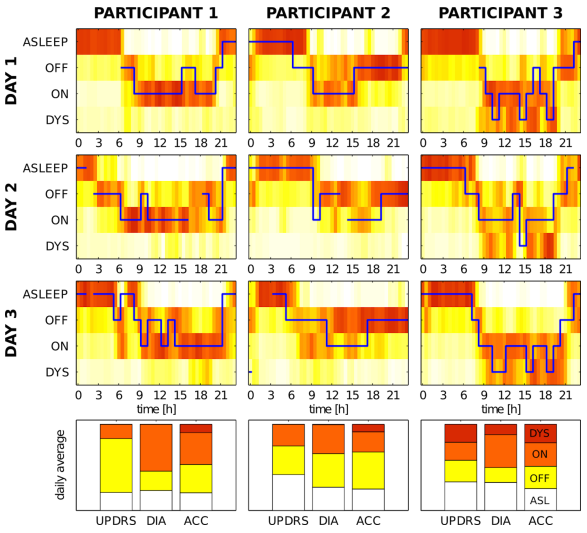

Figure 3: Predictions of the best performing network for three consecutive days and the mean prediction for all 7 days of three participants. Each subplot shows the colour-coded predictions of the network over time (white=0 to red=1). The blue lines indicate the diary entries recorded by each participant (line omitted for missing entries). The bottom row indicates the distribution of disease state according to patient recall (UPDRS), diary (DIA) and our system (ACC).

systems have to be based on data collected in naturalistic, ecologically valid surroundings such as the private home.

In this work we investigated the problem of the assessment of disease state in PD based on a large data-set of many weeks worth of movement data collected from 34 individuals. We developed a novel methodology for research on this problem, which is based on large quantities of naturalistic behaviour collected in the daily life of people affected by PD. Such naturalistic environments pose significant challenges for data acquisition in the form of usability constraints as well as challenges that stem from unreliable labelling obtainable in such settings. Labels that are accessible are infrequent with respect to the data sampling rates and inherently unreliable, subject to recall bias, class confusion, and boundary issues.

In our experiments we showed that deep learning seems particularly suitable to discover disease characteristics despite unreliable labelling of training-data, a setting in which other methods such as decision trees provide poor (generalisation) performance. Deep learning has been applied in similar settings, such as speech (Deng, Hinton, and Kingsbury 2013) or object recognition (Lee et al. 2009), where unlabelled data is easily accessible. However, our results indicate that the common approach to pre-train deep architectures on unlabelled data with subsequent fine-tuning based on a (smaller) set of labelled instances does not improve results in this problem setting. The behaviour observed in laboratory conditions just appears to be a very poor model for naturalistic behaviour, as systems trained on that data show disappointing generalisation performance.

The performance of even the best model does not exceed a mean f1-score of 0.6. To an extent such low results are explained by the poor quality labelling. However, even this relatively low performance is useful for clinical practice. Illustrated in Figure 3 are the predictions for three consecutive days for three participants of the best performing network, where the predicted disease states clearly show very similar patterns of fluctuation compared to the diary entries for each participant. Beyond the assessment of fluctuations in disease state there are other clinical applications. A common measure for the efficacy of interventions in PD is an overall reduction in e.g. "off time", where the average activation of output-units of our system (ACC) only shows little difference to the current best practice for this assessment (DIA) (see Figure 3).

We have not observed any over-fitting to the naturalistic behaviour in the HOME data-set. The unreliable labelling leads to many inconsistencies, which naturally prevent overfitting. A more pressing concern is under-fitting, where automatically adapting or omitting episodes with low confidence may provide significant improvements over the current results. We found that it is crucial to utilise large mini-batches during training (up to 2,048 samples), which may also stem from the unreliable labelling. Another issue surrounds the feature extraction. Effectively the disease state has little impact on the movement data, whose primary source of variance are the physical activities the participants engage in, such as walking. For systems to generalise across those activities it is crucial to tailor a feature representation towards the underlying movement characteristics. These should be accessible to data-driven approaches that avoid manual feature engineering, such as convolutional architectures or techniques like sparse coding (Bhattacharya et al. 2014).

In summary, the problem of disease state assessment in PD is far from being solved. It appears that current challenges may be overcome if novel methodologies are employed in research on PD, where suitable machine learning methods play a key role. Advances that address the unique challenges of this problem setting will have significant societal impact, as not only individuals with PD but also many other degenerative conditions would benefit from practical assessment systems. Beyond possible impact, the characteristic challenges of naturalistic settings make for a unique machine learning problem, which could serve as a novel testbed for the development and evaluation of unsupervised or

semi-supervised learning approaches.

Acknowledgments

We would like to acknowledge the help and support of the staff and patients at the Parkinson's Department of the Northumbria Healthcare NHS Foundation Trust. Parts of this work have been funded by the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE).

References

Bhattacharya, S.; Nurmi, P.; Hammerla, N.; and Plötz, T. 2014. Using unlabeled data in a sparse-coding framework for human activity recognition. *PMC*.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46(3):33.

de Lau, L., and Breteler, M. 2006. Epidemiology of parkinson's disease. *The Lancet Neurology* 5(6):525–535.

Deng, L.; Hinton, G.; and Kingsbury, B. 2013. New types of deep neural network learning for speech recognition and related applications: An overview. In *ICASSP*, 8599–8603. IEEE.

Giuffrida, J. P.; Riley, D. E.; Maddux, B. N.; and Heldman, D. A. 2009. Clinically deployable kinesia™ technology for automated tremor assessment. *Movement Disorders* 24(5):723–730.

Goetz, C. G.; Tilley, B. C.; Shaftman, S. R.; Stebbins, G. T.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stern, M. B.; Dodel, R.; et al. 2008. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement disorders* 23(15):2129–2170.

Griffiths, R. I.; Kotschet, K.; Arfon, S.; Xu, Z. M.; Johnson, W.; Drago, J.; Evans, A.; Kempster, P.; Raghav, S.; and Horne, M. K. 2012. Automated assessment of bradykinesia and dyskinesia in parkinson's disease. *Journal of Parkinson's disease* 2(1):47–55.

Hammerla, N.; Kirkham, R.; Andras, P.; and Plötz, T. 2013. On Preserving Statistical Characteristics of Accelerometry Data using their Empirical Cumulative Distribution. In *Proc. Int. Symp. Wearable Computing (ISWC)*.

Hinton, G., and Salakhutdinov, R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Hinton, G. E.; Osindero, S.; and Teh, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.

Hoehn, M. M., and Yahr, M. D. 1998. Parkinsonism: onset, progression, and mortality. *Neurology* 50(2):318–318.

Hoff, J., et al. 2001. Accelerometric assessment of levodopa-induced dyskinesias in parkinson's disease. *Movement disorders* 16(1):58–61.

Hoff, J.; Van Der Meer, V.; and Van Hilten, J. 2004. Accuracy of objective ambulatory accelerometry in detecting motor complications in patients with parkinson disease. *Clinical neuropharmacology* 27(2):53–57.

Ladha, C.; Hammerla, N. Y.; Olivier, P.; and Plötz, T. 2013. Climbax: skill assessment for climbing enthusiasts. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 235–244. ACM.

Lee, H.; Grosse, R.; Ranganath, R.; and Ng, A. Y. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 609–616. ACM.

Maetzler, W.; Domingos, J.; Srulijes, K.; Ferreira, J. J.; and Bloem, B. R. 2013. Quantitative wearable sensors for objective assessment of parkinson's disease. *Movement Disorders* 28(12):1628–1637.

McNaney, R.; Lindsay, S.; Ladha, K.; Ladha, C.; Schofield, G.; Plötz, T.; Hammerla, N.; Jackson, D.; Walker, R.; Miller, N.; and Olivier, P. 2011. Cueing Swallowing in Parkinson's Disease. In *Proc. ACM CHI Conference on Human Factors in Computing Systems*.

Mera, T. O.; Heldman, D. A.; Espay, A. J.; Payne, M.; and Giuffrida, J. P. 2012. Feasibility of home-based automated parkinson's disease motor assessment. *Journal of neuroscience methods* 203(1):152–156.

Palmer, J.; Coats, M.; Roe, C.; Hanko, S.; Xiong, C.; and Morris, J. 2010. Unified parkinson's disease rating scale-motor exam: inter-rater reliability of advanced practice nurse and neurologist assessments. *Journal of advanced nursing* 66(6):1382–1387.

Patel, S.; Lorincz, K.; Hughes, R.; Huggins, N.; Growdon, J.; Standaert, D.; Akay, M.; Dy, J.; Welsh, M.; and Bonato, P. 2009. Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. *Information Technology in Biomedicine, IEEE Transactions on* 13(6):864–873.

Plötz, T.; Hammerla, N.; Rozga, A.; and Reavis, A. 2012. Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities. In *Proc. Int. Conf. Ubiquitous Comp. (UbiComp)*.

Plötz, T.; Hammerla, N. Y.; and Olivier, P. 2011. Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, 1729–1734. AAAI Press.

Reimer, J.; Grabowski, M.; Lindvall, O.; and Hagell, P. 2004. Use and interpretation of on/off diaries in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* 75(3):396–400.

Stikic, M.; Van Laerhoven, K.; and Schiele, B. 2008. Exploring semi-supervised and active learning for activity recognition. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, 81–88. IEEE.

Tsipouras, M.; Tzallas, A.; Rigas, G.; Bougia, P.; Fotiadis, D.; and Konitsiotis, S. 2010. Automated levodopa-induced dyskinesia assessment. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2411–2414. IEEE.