Evolving Classifiers to Inform Clinical Assessment of Parkinson's Disease

Michael A. Lones*, Jane E. Alty†, Stuart E. Lacy*, D. R. Stuart Jamieson†, Kate L. Possin‡,
Norbert Schuff§¶ and Stephen L. Smith*
*Intelligent Systems Group, Department of Electronics, University of York, York, UK
†Department of Neurology, Leeds Teaching Hospitals NHS Trust, Leeds, UK
‡Memory and Aging Center, University of California, San Francisco, USA
§Department of Veterans Affairs Medical Center, San Francisco, USA
¶Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA

***Abstract*—We describe the use of a genetic programming system to induce classifiers that can discriminate between Parkinson's disease patients and healthy age-matched controls. The best evolved classifer achieved an AUC of 0.92, which is comparable with clinical diagnosis rates. Compared to previous studies of this nature, we used a relatively large sample of 49 PD patients and 41 controls, allowing us to better capture the wide diversity seen within the Parkinson's population. Classifiers were induced from recordings of these subjects' movements as they carried out repetitive finger tapping, a standard clinical assessment for Parkinson's disease. For ease of interpretability, we used a relatively simple window-based classifier architecture which captures patterns that occur over a single tap cycle. Analysis of window matches suggested the importance of peak closing deceleration as a basis for classification. This was supported by a follow-up analysis of the data set, showing that closing deceleration is more discriminative than features typically used in clinical assessment of finger tapping.**

I. Introduction

Neurodegenerative diseases, such as Parkinson's, Alzheimer's and Huntington's, are caused by the loss or functioning neurones in the brain. These diseases mostly affect the elderly and, due to ageing populations around the world, they represent a growing social and economic problem. Although there are currently no cures for the majority of these diseases, early diagnosis and frequent monitoring are important in order for sufferers to plan their lives and receive appropriate therapeutic treatment. However, these diseases can be challenging to diagnose, particularly in their early stages. Parkinson's, for example, has reported misdiagnosis rates of up to 25%, and is often confused with other neurodegenerative conditions such as progressive supranuclear palsy (PSP) [2, 8]. The diagnosis of idiopathic Parkinson's is further confounded by it often appearing clinically indistinct from other types of parkinsonism due to genetic, vascular, toxic or drug causes [4]. Consequently, diagnosis is made through a subjective clinical assessment of a patient's symptoms, rather than via biochemical markers.

In previous work, we have speculated about how computational techniques could be used to make clinical assessment more objective [1]. In particular, we have proposed the use of computational intelligence methodologies to look for patterns within recordings of conventional clinical assessments [12, 16, 18]. Computational intelligence techniques are particularly desirable, because we often have little idea of which patterns are most discriminative. In this respect, evolutionary algorithms are especially useful, since they are very flexible with regard to how solutions are represented, implying that we can use a variety of representations in order to investigate different types of pattern. For instance, in our previous work, we have evolved both static [18] and dynamical [12] representations.

However, a potential weakness of computational intelligence techniques is that their inferred solutions can be difficult to understand. This is particularly important for medical diagnosis, since it is desirable that both clinicians and patients have confidence in the diagnosis; this is difficult to achieve if a diagnostic classifier's behaviour is not well understood. Nevertheless, this does not mean that an evolved diagnostic classifier can not be used to inform diagnosis—since, even if its behaviour is not wholly understood, it is often possible to identify the features that underlie its predictive power and use these to inform clinical assessment.

In this paper, we describe our use of a genetic programming system to induce classifiers that can recognise diagnostically-relevant patterns in movement data recorded from Parkinson's disease (PD) patients. In particular, we describe how a relatively simple analysis of an evolved classifier can identify the components of movement that underlie the classifier's behaviour, and how these components offer higher diagnostic power than those conventionally measured during clinical assessment.

The paper is organised as follows: Section II describes the clinical data used in this study, Section III outlines how classifiers were evolved, Section IV reports classifier performance, Section V discusses the behaviour of the most discriminative evolved classifier, Section VI discusses the implications for clinical assessment, and Section VII concludes.

II. Clinical Data

We recorded the movements of 49 PD patients and 41 age-matched controls as they carried out a standard clinical finger tapping test. This involved each subject repeatedly tapping together the thumb and index finger of their dominant hand for a period of 30 seconds, with instructions to do this at the highest rate and largest amplitude they could comfortably

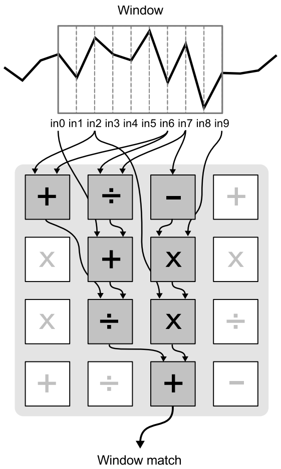


Fig. 1: Example of a window classifier represented using CGP.

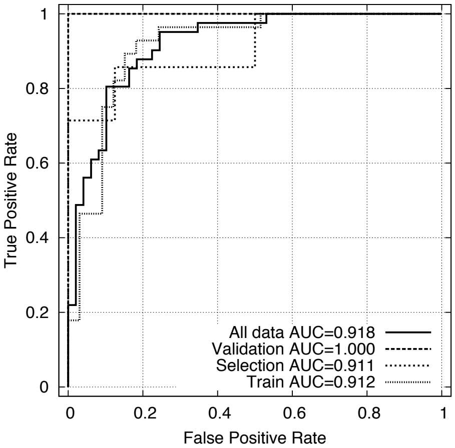


Fig. 2: ROC curves for most discriminative classifier.

achieve. Their movements were recorded using a Polhemus Patriot electromagnetic motion tracking device, whose sensors were attached to the subject's thumb and index finger, and which provided the position and rotation of these sensors with respect to a fixed source at a rate of 60Hz. The study was granted ethical approval, and written informed consent was obtained from each subject.

Prior to being processed by an evolved classifer, each subject's movement data was preprocessed. First, the data was converted to an acceleration time series, since this helps to emphasise small changes that occur in the subject's movements during tapping. Next, the acceleration data was down-sampled by a factor of 2, and a mean average filter of size 2 was applied; this removes noise and emphasises the shape of the acceleration profile. Finally, the acceleration data was truncated to one standard deviation around the mean and scaled uniformly to the unit interval; this transforms the absolute acceleration data into relative acceleration data, which we have found to lead to the evolution of more robust classifiers.

III. Evolving Classifiers

IRCGP [17], a variant of Cartesian genetic programming (CGP) [13], was used to induce classifiers that could distinguish between movement data recorded from PD patients and controls. We have previously used this algorithm to induce diagnostic classifiers for a number of biomedical problems [10, 11]; full details can be found in [11]. As with standard CGP, solutions are represented using an integer Cartesian grid; at each position there is a function that receives its inputs from functions located at other co-ordinates within the grid. Functions may also receive inputs from external inputs, and there is a single designated output function. Hence, a solution is a directed graph describing a mathematical expression that maps inputs to output (see Fig. 1).

Each evolved expression may use up to 20 external inputs, which represent a contiguous time series window within the preprocessed acceleration data. When classifying a tapping sequence, the data is presented as a series of overlapping windows, each wide enough to cover a period of slightly more than a single tap, on average. The classification for a sequence is then the mean of the CGP expression's output for all windows. Hence, an evolved CGP expression is expected to recognise diagnostically-relevant acceleration features that occur within a single tap, and the classification for a subject is the mean occurrence of these features over the whole tapping period.

The evolutionary algorithm is used to find diagnostic classifiers that have high predictive accuracy. This is done using an objective function that attempts to maximise the area under the ROC curve (AUC) when separating PD patients from controls, i.e. a classifier that maximises accuracy across all trade-offs between specificity and sensitivity. AUC is equivalent to the probability that a randomly chosen subject will be assigned to the correct class [7]. An AUC of 1 means that a classifier can achieve 100% specificity and 100% sensitivity. An AUC of 0.5 indicates performance no better than random. An AUC less than 0.5 indicates the same predictive power as one with $1-\mathrm{AUC}$, but with a reversed ordering of the classes within its output range: during selection, these are treated equivalently.

To encourage the evolution of interpretable classifiers, evolved solutions were restricted to a 5x5 Cartesian grid, i.e. a maximum of 25 functions. The evolutionary algorithm used a population size of 200, a generation limit of 100, tournament selection size 4, uniform crossover, and a mutation rate of 6% for functions and 4% for functionality elements.

IV. Classifier Performance

Over the course of 50 independent runs, classifiers were induced using a training set comprising two-thirds of the clinical data. A selection set, comprising another one-sixth of the data, was then used to identify the evolved classifier with the best performance on previously unseen data. This most general classifier was then re-evaluated using a validation set, comprising another one-sixth of the data, in order to provide an unbiased estimate of its performance. Fig. 2 shows the ROC curves for this classifier.

It can be seen that the classifier is fairly stable across the different partitions of the data, achieving an AUC of 0.918 across all subjects, and a perfect classification of the validation

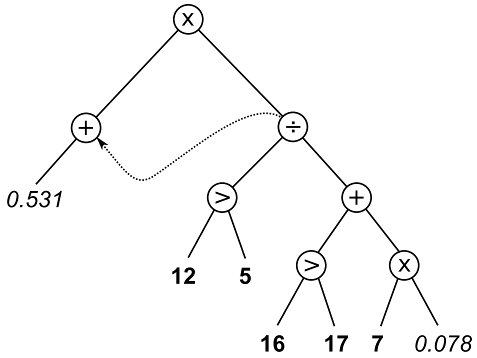


Fig. 3: Most discriminative classifier. Window offsets are in bold, constants are in italics, and $>$ indicates the $\max$ function. The dashed line indicates sub-expression reuse, which is possible because we are using graph-based CGP, rather than a standard tree-based GP.

set. For all subjects, this corresponds to an accuracy of around 85%. By comparison, clinical accuracies for PD diagnosis are around 92-94% for movement disorder specialists, 75% for non-expert secondary carers, and 47% for community care [14]. However, it should be noted that the two classification tasks are only partially comparable. In one sense, the task we are solving is harder, because (for ethical reasons) the patients are on dopaminergic medications, which reduce the signs of PD. In another sense, the task is easier, because we are only distinguishing PD from healthy, rather than distinguishing between different neurodegenerative conditions. Nevertheless, the accuracy of the evolved classifier is relatively good, and it achieves this using only a single source of data, rather than (in the clinical case) a complete assessment and medical history.

This is not the first time that computational intelligence techniques have been applied to PD diagnosis [3]. However, our work is distinct in that we have collected data from a relatively large sample of PD and control patients compared to previous studies. This gives us more confidence in the generality of our results, especially given the high degree of variance known to exist within the PD population. For instance, many of the previous studies have used a dataset hosted at UCI's machine learning repository, which comprises vocal recordings from 23 PD patients and only 8 healthy controls [9]. Our approach also has the advantage of basing diagnosis on a standard clinical test, allowing us to compare against clinical features used to assess PD.

V. Behavioural Analysis

Fig. 3 shows the evolved expression used by the most discriminative classifier. The expression is relatively simple, comprising 7 functional elements (5 of which are used twice) and two constants, and using inputs from 5 of the offsets within an acceleration data window. Despite its simplicity, it is not obvious how the classifier works by looking at this expression alone.

A. Pattern matches

Another way of understanding the classifier's behaviour is to look at the patterns of acceleration which cause it to produce a particularly high or a particularly low output, since these correspond to strong indications of PD or normal tapping behaviour. Fig. 4 shows some examples of data windows: Figs. 4a–c show taps that produced a strong PD response, and Figs. 4d–f show taps that produced a strong normal response.

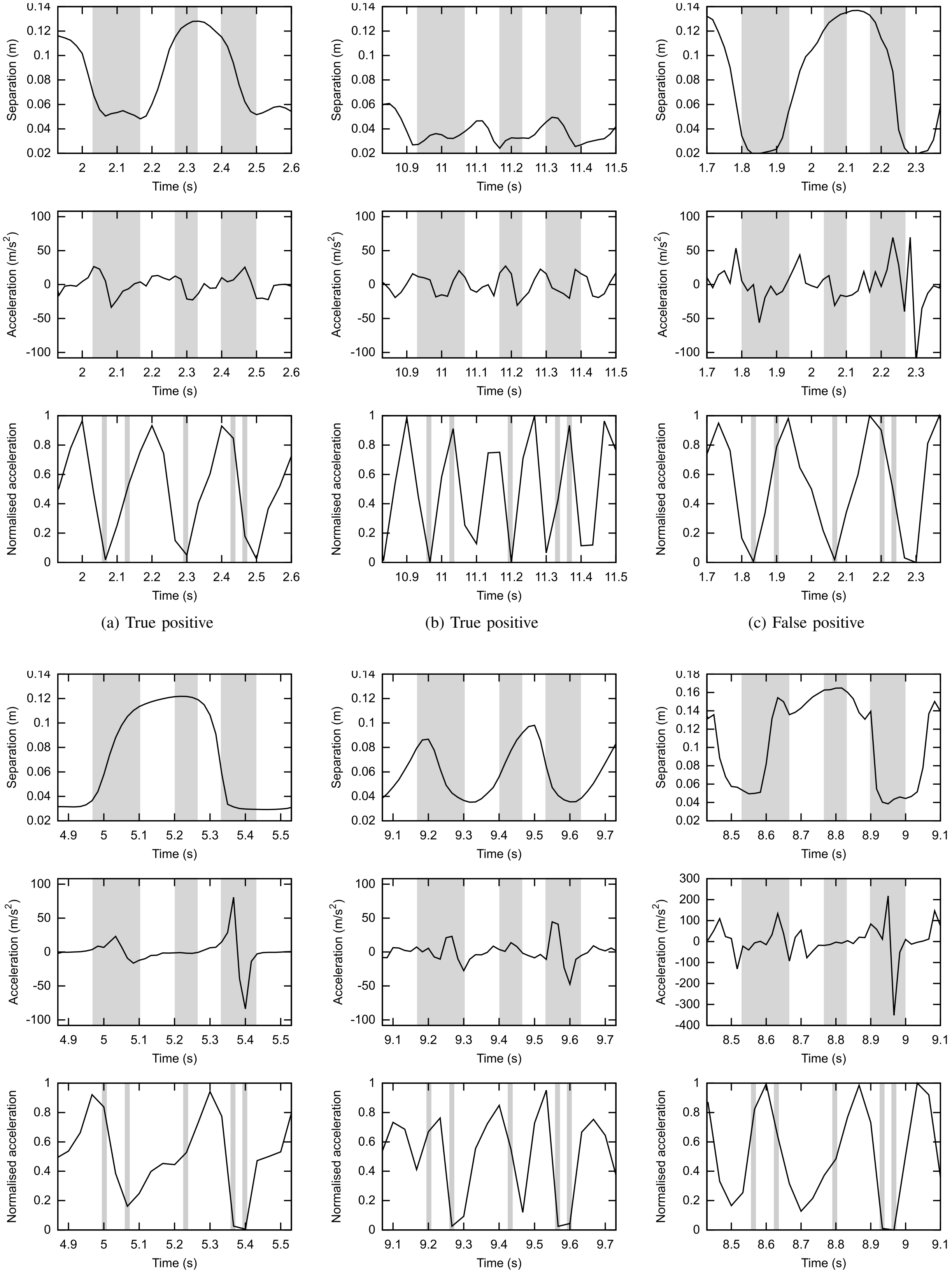
In addition to showing the preprocessed window of accelerations input to the classifier (the bottom plot in each case), Fig. 4 also shows the corresponding raw acceleration and finger separation data. The grey boxes drawn on the latter show the periods of movement that contribute to the window offsets used by the classifier, which are shown as grey lines in the preprocessed windows. This indicates that there are three regions considered by the classifier. For the majority of windows with strong responses (and all the windows shown here), the right-most region corresponds to the closing part of the tap. For windows that give a strong PD response (Figs. 4a–c), the left-most region usually corresponds to the region of least separation between taps, which appears to act as a reference point for aligning the start of a tapping motion. This can be seen in Figs. 4a and 4c. For windows that give a strong non-PD response, by comparison, the left-most region tends to overlay the opening phase (Fig. 4d) when matching against a single tap, rather than the between-tap phase. Although the classifier's mathematical expression appears quite simple, this suggests that it is looking for different patterns when matching PD taps or control taps.

B. Effect of preprocessing

In general, acceleration windows in PD patients tend to display high-frequency components, indicating jerky movement. However, this jerkiness also occurs in control subjects, many of whom have other age-related conditions such as arthritis; as such, it may not be a good indicator of PD. Fig. 4 illustrates how down-sampling and moving average filtering tends to remove these high-frequency components, leaving just the gross shape of the acceleration profile. It also shows how truncation tends to obfuscate small differences in the absolute sizes of acceleration and deceleration peaks, which are unlikely to be diagnostically relevant. There is also a tendency for PD patients to tap at lower amplitudes than other people in their age group, but again this is not a robust indicator for diagnosis. In this case, Fig. 4b shows how scaling removes information about a subject's tapping amplitude. As a result of these preprocessing operations, the signal received by the classifier is significantly different to the raw acceleration profile, and emphasises gross features such as overall patterns of acceleration/deceleration during a tapping motion, and the presence of significant differences in magnitude of accelerations or decelerations.

C. Feature extraction

Considering the evolved expression in Fig. 3, the largest contribution to its output comes from the division node. Our analysis of match windows suggests that, for windows that produce strong non-PD signals, the main determinant of this division is the relative size of the deceleration in the closing phase of the tap. In a normal tapping motion, this closing deceleration should be significantly larger than the closing deceleration in the opening phase of the tap, since the collision between thumb and finger produces a sudden stop in comparison to the more elastic slowing when the thumb



(a) True positive

(b) True positive

(c) False positive

(d) True negative

(e) True negative

(f) False negative

Fig. 4: Example tapping data windows, showing in each case [top] the separation between index finger and thumb, [middle] the pattern of acceleration, and [bottom] the corresponding preprocessed data window that is used by the evolved classifier. Grey regions show the parts of the window that are considered by the evolved classifier when assigning a diagnostic class.

and finger are maximally opposed. This is very evident in Fig. 4d. However, in windows that produce strong PD responses from the classifier, we do not see this relationship. Instead, decelerations during the opening and closing phases are of a similar magnitude, as shown in Fig. 4a. This suggests that the relative size of the closing deceleration is the main feature underlying the classifier's behaviour. This interpretation is supported by examples of misclassified tap windows. In Fig. 4c, a tap from a control sequence is misclassified as PD; this appears to be because the relatively large deceleration in closing has been lost during preprocessing. In Fig. 4f, a tap from a PD sequence is misclassified as non-PD; in this case, this appears to be due to an abnormal pattern of tapping followed by a strong deceleration.

Fig. 5 shows examples of a PD patient and a control subject carrying out finger tapping over a 5s interval. In the control subject, the relationship between opening and closing peak deceleration is easy to see: in each tap, the yellow marker (peak opening deceleration) always appears lower than the green marker (peak closing deceleration). In the PD subject, by comparison, there is no clear relationship between the two: sometimes closing deceleration is greater than opening deceleration, sometimes vice versa.

However, it should be noted that the evolved expression does not always match single taps. Figs. 4b and e show examples where a pair of taps at a higher frequency produce a strong PD and non-PD response, respectively. Whilst these matches may be superfluous, it is interesting to note that the PD match is clearly abnormal, corresponding to two subsequent taps with extra intermediate deceleration phases. The control match, on the other hand, appears to be recognising consistent behaviour between taps—a feature that is sometimes absent in PD patients. It should also be noted that a subject's classification is not based on individual strong matches, but rather on the classifier's mean response to all windows in a sequence. Consequently, even though the evolved expression appears relatively simple, and we have been able to make some insights into how it classifies, it is quite likely that we have not elicited its full behaviour.

D. Closing deceleration as a predictor

Given the prominent role that closing deceleration appears to play in the classifier's diagnostic ability, we looked at the predictive accuracy of this component alone. For each patient and control tapping sequence, this involved identifying the boundaries between taps, measuring the magnitude of the largest deceleration during closing, and taking the mean across all taps. Fig. 6 shows the resulting ROC curve when the mean closing deceleration is used to discriminate between patients and controls, alongside ROC curves for mean amplitude, mean speed, and the evolved classifier. The AUC for mean closing deceleration is fairly high, and significantly higher than those for speed and amplitude, metrics which are used in clinical assessment. Nevertheless, it is still significantly lower than the AUC for the classifier as a whole, again suggesting that the classifier also takes other features into account.

VI. Discussion

According to current clinical practice, a patient's finger tapping is scored according to the MDS-UPDRS scale [5]. In principle, this involves watching the patient carry out ten finger taps, during which the practitioner must estimate deviation from normality with regard to amplitude, speed, rhythm and hesitations. However, there is little theoretical underpinning to this process and there is considerable inter-rater and intra-rater variation within the scores [15]. Furthermore, most raters tend to favour amplitude over the other measures [6]. As we have shown, amplitude is less discriminative than closing deceleration as a predictor of PD, and we would expect better diagnostic accuracy if clinicians were to measure closing deceleration.

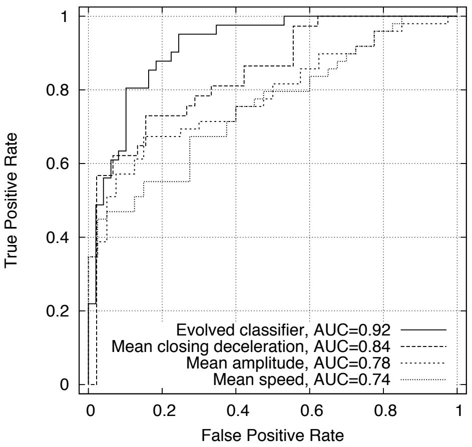
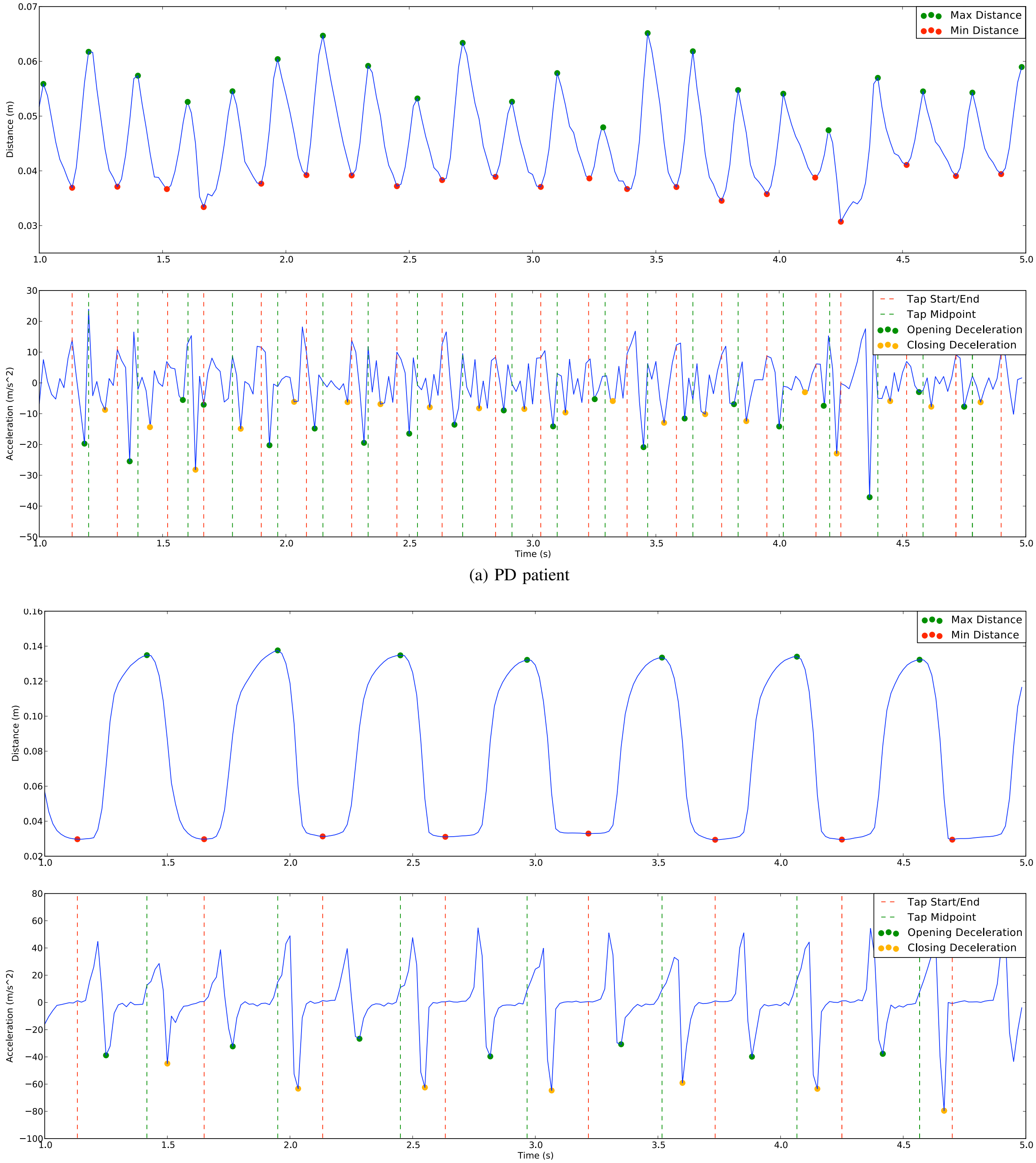


Fig. 6: Comparison of ROC curves for most discriminative classifier, closing deceleration, and standard clinical measures of finger tapping.

The importance of measuring internal components of taps, rather than gross measures of movement, has previously been made in [19]. In a smaller study involving 16 PD patients and 32 controls, the authors noted that the peak velocity during the opening part of a tap was more discriminative than a group of other metrics. By comparison, our results suggest that the closing phase is diagnostically more relevant than the opening phase. This difference may be due to our use of larger samples (49 PD and 41 controls), particularly in regard of the PD population.

Since the MDS-UPDRS scale does not measure internal components of a tap, this feature is unlikely to be detected in a standard clinical assessment. Even if it were considered during clinical assessment, the final deceleration is a difficult component to measure visually. This suggests that a process of recording a patient's movements, followed by suitable analysis, might be more effective for diagnosing and monitoring PD and other movement disorders.

Whilst we have only considered PD in this paper, the methods we have described could be applied more widely to other neurodegenerative diseases. Many of these present with movement disorders, and in most cases there is only a limited understanding of the relationships between disease states and abnormal movement characteristics. Better characterisation could improve the ability of clinicians to distinguish between clinically similar conditions (such as PD and PSP). It could also allow clinicians to more accurately measure how a patient's symptoms change over time, and to prescribe



(a) PD patient

(b) Healthy control

Fig. 5: Examples of a PD patient and a healthy age-matched control carrying out finger tapping over a 5s interval.

medication appropriately. For PD, this is important, since incorrect dosage can lead to unpleasant, difficult to treat, side-effects. In addition to contributing to better diagnosis and monitoring, better characterisation might also lead to a better understanding of the disease processes underlying neurodegenerative conditions, which are often poorly understood.

In this work, we have focused on the use of window-based genetic programming classifiers. These are effective at identifying patterns that occur within single tap cycles, and are relatively easy to analyse. However, window-based architectures such as this are limited in their ability to characterise movement trends that occur over longer time periods. In other work [12], we have looked at whether novel dynamical classifier architectures can be used to identify such patterns. Our initial results are promising, suggesting that dynamical classifiers can achieve higher accuracy than static classifiers. However, it is much harder to identify the basis of their discriminative ability. Hence, to an extent there is a trade-off between using relatively simple models to inform clinical assessment, and using more complex models that can achieve higher diagnostic accuracy. In future work, we aim to develop more complex analytical methods to complement these more complex models.

VII. Conclusions

In this paper, we have described how a genetic programming system was used to induce diagnostic classifiers that can distinguish between recordings of Parkinson's disease patients and healthy age-matched controls carrying out finger tapping, a standard clinical assessment task. The most discriminative classifier achieved a diagnostic accuracy comparable to those achieved by experienced clinicians. A behavioural analysis of this classifier revealed that the most important feature underlying its accuracy was the peak magnitude of deceleration during the closing phase of a finger tap. Measurements of this feature within the clinical data support this observation, showing that closing deceleration is a better predictor of Parkinson's than standard clinical metrics such as amplitude and speed. This information could be used to improve the accuracy of clinical assessment of Parkinson's. However, given the limited ability of humans to measure subtle features of movement, it also supports the use of automated methods within clinical assessment.

References

[1] N. M. Aly, J. R. Playfer, S. L. Smith, and D. M. Halliday. A novel computer-based technique for the assessment of tremor in Parkinson's disease. *Age and Ageing*, 36(4): 395–399, 2007.

[2] N. P. S. Bajaj, V. Gontu, J. Birchall, J. Patterson, D. G. Grosset, and A. J. Lees. Accuracy of clinical diagnosis in tremulous parkinsonian patients: a blinded video study. *J Neurol Neurosurg Psychiatry*, 81(11):1223–1228, Nov 2010.

[3] R. Das. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572, 2010.

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

[5] C. G. Goetz, S. Fahn, P. Martinez-Martin, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. *Movement Disorders*, 22(1):41–47, 2007. ISSN 1531-8257.

[6] D. A. Heldman, J. P. Giuffrida, R. Chen, M. Payne, et al. The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures. *Movement Disorders*, 26(10):1859–1863, 2011. ISSN 1531-8257.

[7] H. C. Kraemer, G. A. Morgan, N. L. Leech, J. A. Gliner, J. J. Vaske, and R. J. Harmon. Measures of clinical significance. *J Am Acad Child Adolesc Psychiatry*, 42 (12):1524–1529, Dec 2003.

[8] C. B. Levine, K. R. Fahrbach, A. D. Siderowf, R. P. Estok, V. M. Ludensky, and S. D. Ross. Diagnosis and treatment of Parkinson's disease: a systematic review of the literature. *Evid Rep Technol Assess (Summ)*, (57):1–4, May 2003.

[9] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 56(4):1015–1022, 2009.

[10] M. A. Lones and S. L. Smith. Discriminating normal and cancerous thyroid cell lines using implicit context representation Cartesian genetic programming. In *Proc. IEEE Congress on Evolutionary Computation (CEC) 2010*. IEEE Press, 2010.

[11] M. A. Lones and S. L. Smith. Objective assessment of visuo-spatial ability using implicit context representation Cartesian genetic programming. In S. L. Smith and S. Cagnoni, editors, *Genetic and Evolutionary Computation: Medical Applications*. John Wiley & Sons, Chichester, UK, 2010.

[12] M. A. Lones, S. L. Smith, A. M. Tyrrell, J. E. Alty, and D. R. S. Jamieson. Characterising neurological time series data using biologically-motivated networks of coupled discrete maps. *BioSystems*, 2012. to appear.

[13] J. F. Miller and P. Thomson. Cartesian genetic programming. In R. Poli, W. Banzhaf, W. B. Langdon, J. F. Miller, P. Nordin, and T. C. Fogarty, editors, *Third European Conf. Genetic Programming*, volume 1802 of *Lecture Notes in Computer Science*, 2000.

[14] National Institute for Health and Clinical Excellence. *Parkinson's disease: diagnosis and management in primary and secondary care*. Royal College of Physicians, 2006. URL `http://www.nice.org.uk/CG035`.

[15] B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, and J. D. Speelman. Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov Disord*, 20(12):1577–1584, Dec 2005.

[16] S. L. Smith and J. Timmis. An immune network inspired evolutionary algorithm for the diagnosis of Parkinson's disease. *BioSystems*, 94(1-2):34 – 46, 2008.

[17] S. L. Smith, S. Leggett, and A. M. Tyrrell. An implicit context representation for evolving image processing filters. In *Proceedings of the 7th Workshop on Evolutionary Computation in Image Analysis and Signal Processing*, volume 3449 of *Lecture Notes in Computer Science*, pages 407–416, 2005.

[18] S. L. Smith, P. Gaughan, D. M. Halliday, Q. Ju, N. M. Aly, and J. R. Playfer. Diagnosis of Parkinson's disease using evolutionary algorithms. *Genetic Programming and Evolvable Machines*, 8(4):433–447, 2007.

[19] M. Yokoe, R. Okuno, T. Hamasaki, Y. Kurachi, K. Akazawa, and S. Sakoda. Opening velocity, a novel parameter, for finger tapping test in patients with parkinson's disease. *Parkinsonism & Related Disorders*, 15(6): 440–444, 2009.