Less is More: Univariate Modelling to Detect Early Parkinson's Disease from Keystroke Dynamics

Antony Milne[1], Katayoun Farrahi[2], and Mihalis A. Nicolaou[1,3]

[1] Department of Computing, Goldsmiths, University of London, UK
[2] Electronics and Computer Science, University of Southampton, UK
[3] The Cyprus Institute, Cyprus

Abstract. We analyse keystroke hold times from typing logs to detect early signs of Parkinson's disease. We develop a feature that captures the dynamic variation between consecutive keystrokes and demonstrate that it can be be used in a univariate model to perform classification with AUC = 0.85 from only a few hundred keystrokes. This is a substantial improvement on the current baseline. We argue that previously proposed methods are based on overcomplicated models — our simpler method is not only more elegant and transparent but also more effective.

1 Introduction

After Alzheimer's, Parkinson's disease (PD) is the world's second most prevalent neurodegenerative disease [1]. Currently, diagnosis is based on a specialist's interpretation of neurological tests completed by the patient at a clinic [2]. This procedure is time-consuming, expensive, subjective, and rather inaccurate (especially for identifying early stages of PD) [3].

Giancardo et al. [4] suggest that early PD can be detected through the analysis of typing logs, studying data obtained from 85 subjects (42 Parkinson's, 43 control) each transcribing text for around 15 minutes. Subsequent analysis of the keystroke dynamics focusses on the length of time between pressing and releasing each key (hold time), a measure believed to be outside a subject's conscious control and independent of typing skills. The so-called neuroQWERTY index (nQi) method is developed to classify a typing session as that of a Parkinson's sufferer or a control subject.

We regard this as a valuable line of research that demonstrates promising results for detecting early PD. In this paper, however, we present results indicating that the analysis presented in Ref. [4] is opaque and overly complicated for the problem at hand. Following the philosophy that 'less is more', we find that the classification performance of nQi can be equalled, and even surpassed, by a far simpler and more easily reproducible methodology.

We begin in Sec. 2 by outlining the nQi formalism and results. This is followed by an exploration of the basic features of the hold time data (Sec. 3), and a demonstration that a univariate model can be used to straightforwardly achieve classification performance equal to nQi (Sec. 4). In Sec. 5 and 6, we develop more sophisticated dynamic features of the data that can be used, again in a univariate model, to substantially outperform nQi. Sec. 7 discusses a recent contribution to the literature [5], which also suffers from significant overengineering and, more importantly, reports results we believe to be invalid.

2 Classification with neuroQWERTY index

Let us briefly describe nQi and the datasets involved. These are labelled early PD (those within five years of confirmed diagnosis: 18 Parkinson's, 13 control) and de novo PD (newly diagnosed and untreated: 24 Parkinson's, 30 control). Each dataset consists of a set of typing sessions. We use h_n to denote the hold time of the n^{th} keystroke during a typing session, which has N keystrokes in total. Both nQi and our proposed classification methods are concerned with the one-dimensional time series $\boldsymbol{h}$.

Ref. [4] begins by partitioning each time series into non-overlapping windows of length 90 s. We write $\boldsymbol{h} \equiv (\boldsymbol{h}^1, \boldsymbol{h}^2, \ldots, \boldsymbol{h}^I)$ to indicate this partitioning, where I gives the total number of windows. Any $\boldsymbol{h}^i$ with fewer than 30 elements is removed. Then, for each window i, a 7-dimensional feature vector $\boldsymbol{x}^i$ is calculated for $\boldsymbol{h}^i$. Let q^i_j be the j^{th} quartile of the elements of $\boldsymbol{h}^i$, and denote the interquartile range as $\Delta q^i \equiv q^i_3 - q^i_1$. Then $\boldsymbol{x}^i$ consists of the following features:

- The proportion of elements that are outliers, defined as $h^i_n < q^i_1 - \frac{3}{2}\Delta q^i$ or $h^i_n > q^i_3 + \frac{3}{2}\Delta q^i$.
- The skewness, given by $(q^i_2 - q^i_1)/\Delta q^i$.
- The flight time between consecutive keystrokes.[4]
- The proportion of elements in $\boldsymbol{h}^i$ that are in each of four equally-spaced bins between 0 and 500 ms.

Training is performed with an ensemble of 200 Linear ϵ-Support Vector Regression models, where hyperparameters are selected using a grid search approach on an external dataset. During testing, a value of nQi for each $\boldsymbol{x}^i$ is calculated by applying all 200 regression models to $\boldsymbol{x}^i$ and then finding the median score, nQi^i. To arrive at a single nQi score for the typing session, these median scores are then averaged over the I windows: $\text{nQi} = \frac{1}{I}\sum_{i=1}^{I} \text{nQi}^i$.

To evaluate nQi, Giancardo et al. [4] perform cross-validation by training on the early PD dataset and testing on the de novo PD dataset, and then vice-versa. This yields a single prediction of nQi for each of the 85 subjects in the combined dataset.[5] Area under the Receiving Operating Characteristic curve (AUC) is

[4] As given, this will yield a number for each keystroke; it is not explained in Ref. [4] how this measure is then aggregated over the window. Moreover, we note that, contrary to the principles promoted by Giancardo et al., this measure appears to use more than purely hold time data.

[5] In fact, each subject in the early PD dataset produced two typing sessions. While training or testing, each typing session is handled independently. If a subject has produced multiple typing sessions then the average nQi is computed to produce a single score.

used to evaluate the binary classification of each subject as either Parkinson's sufferer or control subject.

In our work, we follow precisely the same evaluation strategy, so that our classification results can be directly compared with those given in Ref. [4]. We are able to reproduce AUC = 0.81 reported by Giancardo et al. for classification using nQi.

3 Exploratory analysis

We begin by performing initial analysis of the early PD and de novo PD datasets, something that has not previously been presented in the literature. Fig. 1 shows the distribution of all the hold times in each dataset, split between Parkinson's and control subjects. Unsurprisingly, there is a clear shift towards longer hold times for Parkinson's sufferers, especially for the early PD dataset. The plots also suggest that there is a greater variance in hold time for Parkinson's sufferers compared to control.

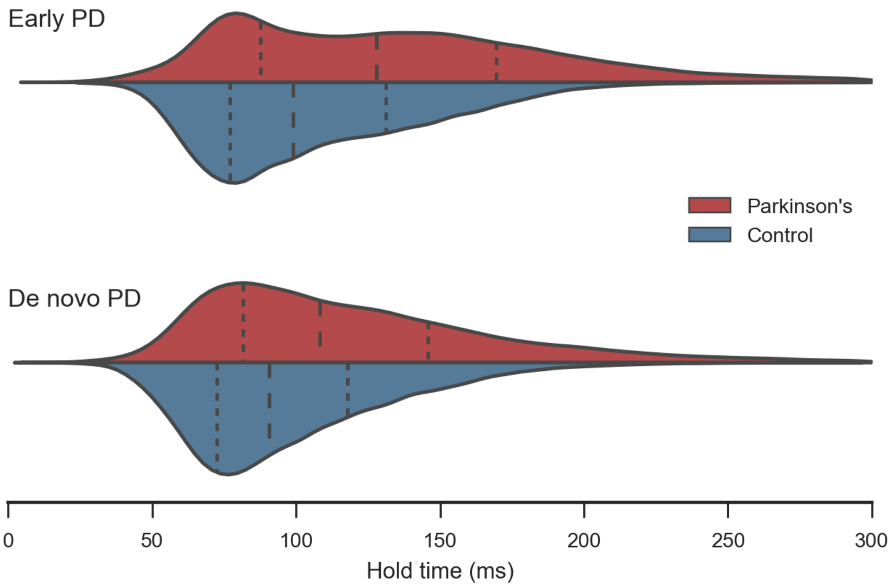

Fig. 1. The distribution of hold times for each of the two datasets used, distinguishing Parkinson's sufferers from control subjects. Each half of the violins are normalised to the same area. Dashed lines indicate the position of the lower quartile, median and upper quartile. Hold times above 300 ms are not shown here (corresponding to about 0.85% of the total data, and overwhelmingly from Parkinson's sufferers).

However, we are interested in classifying individual subjects rather than groups as a whole. To probe the difference in distributions suggested by Fig. 1, we calculate the hold time mean $\langle h \rangle \equiv \frac{1}{N}\sum_{n=1}^{N} h_n$ and standard deviation $\sigma(h) \equiv \sqrt{\langle h^2 \rangle - \langle h \rangle^2}$ for each subject. Fig. 2 suggests that these statistics could be used to classify at the level of individual subjects. There is a clear trend towards Parkinson's sufferers having higher keystroke hold time mean and standard deviation. In particular, standard deviation appears to be a promising candidate for a discriminatory statistic.

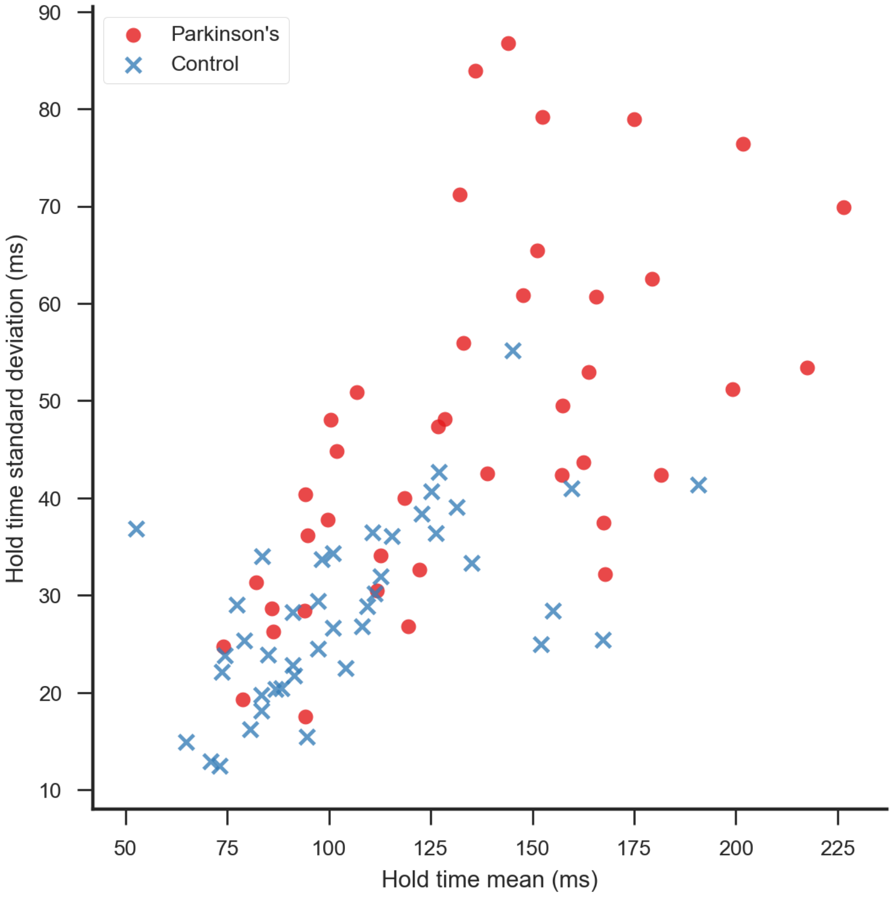

Fig. 2. The mean hold time $\langle h \rangle$ and standard deviation $\sigma(h)$ for all users in the study. Data from the early PD and de novo PD datasets are shown the same way. The average (std) of $\sigma(h)$ is 47 (18) for Parkinson's and 29 (9) for control, suggesting the power of this statistic as a discriminatory feature.

4 Classification with elementary statistics

One might well wonder whether these basic statistics alone are sufficient to effectively discriminate between Parkinson's and control subjects. We perform Logistic Regression using the features $\langle h \rangle$ and $\sigma(h)$ for each subject using scikit-learn's default parameters [6] and immediately obtain a classification performance comparable to nQi. In fact, we obtain AUC = 0.82 using standard deviation alone as a single feature (compared to AUC = 0.81 for nQi).[6] Fig. 3 and Table 1 show the performance of this univariate method with standard deviation feature, which we refer to as the Stdev model, along with the performance of nQi and two other models which will be discussed in later sections. The classification performance achieved using a single elementary statistical feature is very similar to that obtained using nQi.

It is for this reason that we believe nQi is a contrived method for performing the classification task. Let us highlight the differences between nQi and our Stdev model:

- nQi splits the time series $\boldsymbol{h}$ for each user into several windows, calculates features for each window separately, and then recombines statistics at the end; we use a feature that uses the time series as a whole.
- nQi uses seven features that capture, in various ways, properties of the distribution of hold times;[7] we use one feature. Furthermore, standard deviation is an extremely well-known and transparent statistic.
- nQi uses an ensemble of 200 classifiers, with hyperparameters optimised using an external dataset; we use a single Logistic Regression algorithm with no optimisation of hyperparameters required.

Clearly the seven features of nQI capture more of the typing behaviour, and these features could be used to paint a more complete picture of a subject. However, for the purposes of classification on the datasets provided, there is no evidence to suggest that nQi outperforms the considerably simpler and more elegant Stdev method. One can achieve strong classification performance without the need to engineer particular statistical features, use anything beyond hold times, or perform carefully optimised ensemble models. We emphasise that the method we propose here has been evaluated using exactly the same cross-validation strategy on the same data as nQi (as are all models discussed in this paper).

Of course, this is not to say that performing a Logistic Regression with default hyperparameters on a single feature is the best possible method. Indeed, we will later formulate a method which substantially outperforms both the Stdev model and nQi. We present the Stdev model in order to show that one may immediately

[6] This classification performance is very similar to that obtained using using both $\langle h \rangle$ and $\sigma(h)$ as features, whilst the performance using just $\langle h \rangle$ as a feature is substantially lower.

[7] We note again that, unlike the Stdev method, nQi actually appears to use information about the flight time in addition to purely hold time data.

and very straightforwardly obtain a baseline classification performance that is comparable to the convoluted methods of nQi. We note that in a related paper on smartphone typing data [7], a univariate model using an elementary statistical feature (sum of covariances) was in fact found to outperform all of the more complicated multivariate methods studied.

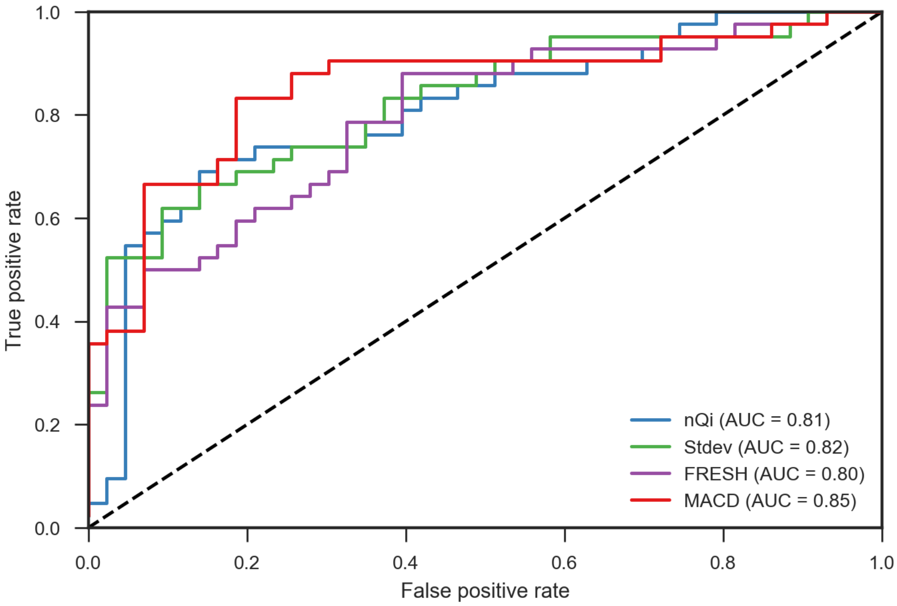

Fig. 3. The ROC curves for all the models evaluated in this paper. nQi values are taken from Ref. [4] (and reproduced by us). The other three methods use a Logistic Regression algorithm with different features. Stdev and MACD correspond to the univariate models with features $\sigma(h)$ and $\langle|\Delta|\rangle$ respectively. FRESH refers to the multivariate model with the five most relevant time series features automatically extracted from each training set. All models were evaluated using the same cross-validation strategy as that used in Ref. [4] (training on the early PD dataset and testing on the de novo PD dataset, and then vice-versa).

5 Feature extraction

We now consider what features might be the most relevant for detecting early PD. We have already seen that using a univariate method based on the standard deviation yields strong classification performance, but can we do better by using more sophisticated features and a multivariate model?

Recall that the data we are working with is a one-dimensional set $\boldsymbol{h}$, whose elements h_n $(n = 1, 2, \ldots, N)$ are ordered according to the order of keystrokes

Table 1. The performance of all the models evaluated in this paper, labelled as in Fig. 3. We follow the same evaluation strategy as Ref. [4] by reporting values of the confusion matrix and accuracy at the cut-off point determined by maximising Youden's J Statistic [8].

Model	**TP**	**FN**	**TN**	**FP**	**Accuracy**	**AUC**
nQi	30	12	36	7	0.76	0.81
Stdev	27	15	37	6	0.75	0.82
FRESH	36	6	26	17	0.73	0.80
MACD	34	8	35	8	0.81	0.85

recorded. Simple statistical measures such as standard deviation discard information encoded in the ordering of the elements h_n; typing behaviour might be captured more effectively by measures that take into account the actual dynamics of $\boldsymbol{h}$.

There are countless features that one could extract from a time series, but not all will be relevant for identifying discriminatory behaviour. We use the Feature Extraction based on Scalable Hypothesis (FRESH) algorithm and associated library `tsfresh` [9, 10]. This characterises time series using a comprehensive set of well-established features, including those that are 'static' (e.g. standard deviation) and truly 'dynamic' (e.g. Fourier transform coefficients). The relevance of each feature is evaluated by quantifying its significance for predicting the target label (for us, Parkinson's or control).

We perform a classification of the time series data with FRESH using the following procedure. The training data is analysed to find the m most relevant features for predicting whether the user has PD. These m features are then extracted on the test data and used to perform classification using Logistic Regression. Features are standardised by scaling to vanishing mean and unit variance. By running this model on $m = 1, 2, \ldots, 10$, we find that the best performance is achieved by $m = 5$. The AUC for this is again comparable to nQi and our univariate standard deviation method (see Fig. 3 for the ROC curve and Table 1 for evaluation metrics).

Let us look at the features extracted by FRESH on the time series $\boldsymbol{h}$. We perform cross-validation based on two datasets (early PD and de novo PD), and hence two different sets of $m = 5$ features are found as being the most relevant during training. These are given in full in Table 2.

For both the early PD and the de novo PD datasets, FRESH finds that several features given by the function `change_quantiles` are highly relevant. This function aggregates consecutive differences between elements of $\boldsymbol{h}$. More precisely, we fix a corridor set by the quantiles `ql` and `qh` and take only those elements for which both `ql` $\leq h_n \leq$ `qh` and `ql` $\leq h_{n+1} \leq$ `qh`. Define $\Delta_n \equiv h_{n+1} - h_n$; then the feature found by `change_quantiles` is given by the aggregator function `f_agg` applied to the set of all Δ_n ($|\Delta_n|$ when `isabs` is set). In other words, we are analysing (a subset of) the differences in hold time between consecutive keystrokes. This captures a more complex element of variance that 'static'

Table 2. The five most relevant features found by FRESH on the early PD and de novo PD datasets. Features are given by the functions and parameters used to calculate them with the `tsfresh` package [10].

Early PD
`change_quantiles(ql=0.8,qh=1.0,isabs=True,f_agg=mean)`
`change_quantiles(ql=0.0,qh=1.0,isabs=True,f_agg=var)`
`spkt_welch_density(coeff=5)`
`variance`
`standard_deviation`
De novo PD
`change_quantiles(ql=0.6,qh=0.8,isabs=True,f_agg=var)`
`change_quantiles(ql=0.4,qh=1.0,isabs=True,f_agg=mean)`
`change_quantiles(ql=0.6,qh=1.0,isabs=True,f_agg=mean)`
`change_quantiles(ql=0.6,qh=0.8,isabs=False,f_agg=var)`
`max_langevin_fixed_point(r=30, m=3)`

measures such as standard deviation do not (although it is worth noting that standard deviation is in fact identified as a highly relevant feature for at least the early PD dataset).

Given the thoroughness of the FRESH algorithm, which extracts several hundred features, it is perhaps at first surprising that this multivariate method does not significantly outperform the univariate method using standard deviation. However, note that none of the most relevant features are common between the two datasets. We are effectively suffering from overfitting: FRESH identifies some rather obscure features that fit the training data very well but do not generalise to the test data. Take, for example, the feature discovered using `spkt_welch_density`, which is present in the early PD but not the de novo PD dataset. This corresponds to the cross power spectral density at a particular frequency after $\boldsymbol{h}$ has been transformed to the frequency domain. This is a feature that happens to correlate strongly with the binary classification targets on the early PD data, but that should clearly not be taken as a feature that truly captures a genuine difference between the typing behaviours of Parkinson's sufferers compared to control subjects.

6 Classification with mean absolute consecutive difference

Using the analysis produced by FRESH, we believe that features based on `change_quantiles` are suitable for capturing the intricate dynamic behaviour of our time series without overfitting. In particular, we take `ql` $= 0.6$ and `qh` $= 1.0$ to mark the corridor of hold times, i.e. we take only the elements of $\boldsymbol{h}$ for which both h_n and h_{n+1} are in the 60^{th} percentile. We then take the mean of the absolute difference in hold time between these consecutive keystrokes to give the feature $\langle|\varDelta|\rangle \equiv \frac{1}{N}\sum_{n=1}^{N}|\varDelta_n|$, where we recall that $\varDelta_n \equiv h_{n+1} - h_n$. We refer to this as the mean absolute consecutive difference (MACD).

Ref. [4] notes that in order to identify Parkinson's sufferers effectively, it is necessary to capture transient bradykinesia effects that prevent the subject from lifting their fingers from keys in a consistent manner. However, static features that describe the distribution of hold times do not yield such information. In contrast, MACD captures precisely the dynamic variation in hold time between one keystroke and the next. We restrict MACD to analysing hold times in the 60^{th} percentile as typing patterns involving longer hold times appear to be particularly discriminatory.

Using MACD as a univariate feature and classifying with Logistic Regression, we obtain the ROC curve and evaluation scores shown in Fig. 3 and Table 1. Crucially, we find AUC $= 0.85$, significantly outperforming all the models previously considered. In fact, using MACD, one can obtain effective classification without needing to analyse every element of the hold time series $\boldsymbol{h}$. In Fig. 4 we demonstrate how classification performance depends on the number of keystrokes analysed. We truncate $\boldsymbol{h}$ after a certain number of elements and perform classification according to the same scheme outlined above, using the MACD model. Fig. 4 demonstrates that one may achieve very good performance (AUC > 0.80) from analysing only 200 keystrokes in a typing session.

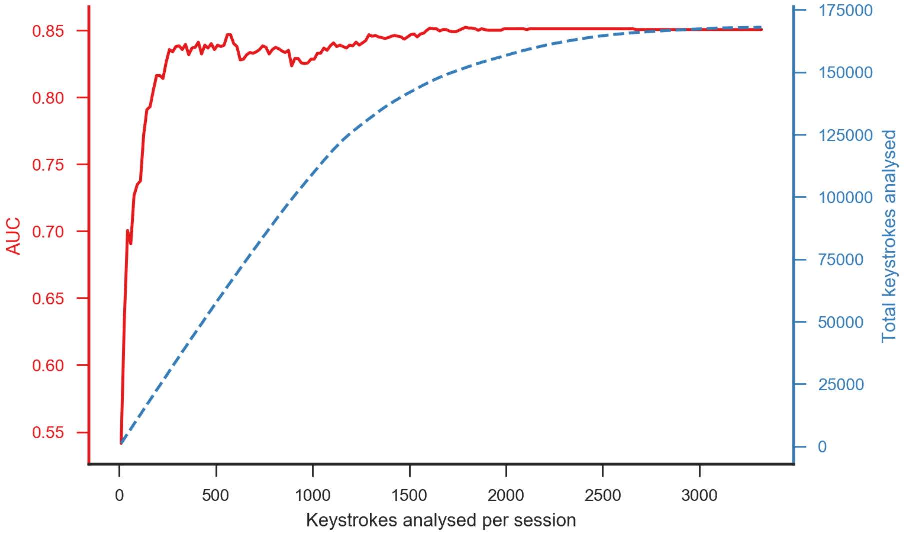

Fig. 4. The dependence of classification performance on the number of keystrokes analysed. The x axis gives the length of the truncated time series $\boldsymbol{h}$. In red (left y axis) we show the AUC achieved by the MACD model operating on the truncated time series; in blue (right y axis) we show the total number of keystrokes that are analysed across all typing sessions in the whole dataset of 85 users.

7 Tappy study

Finally, we make some important remarks regarding the 'Tappy' dataset and associated analysis performed in a recent study by Adams [5]. Some concern peculiarities with the data; some concern the methods used during the analysis; and some concern the validity of the results. Although we believe that Adams' work should be of considerable interest to researchers, we were not able to replicate the perfect evaluation results claimed. Other researchers have similarly struggled to achieve the performance claimed by Adams [11]. Here we suggest where there may be flaws in the analysis presented in Ref. [5]. Moreover, we see once again the use of severely overcomplicated methods.

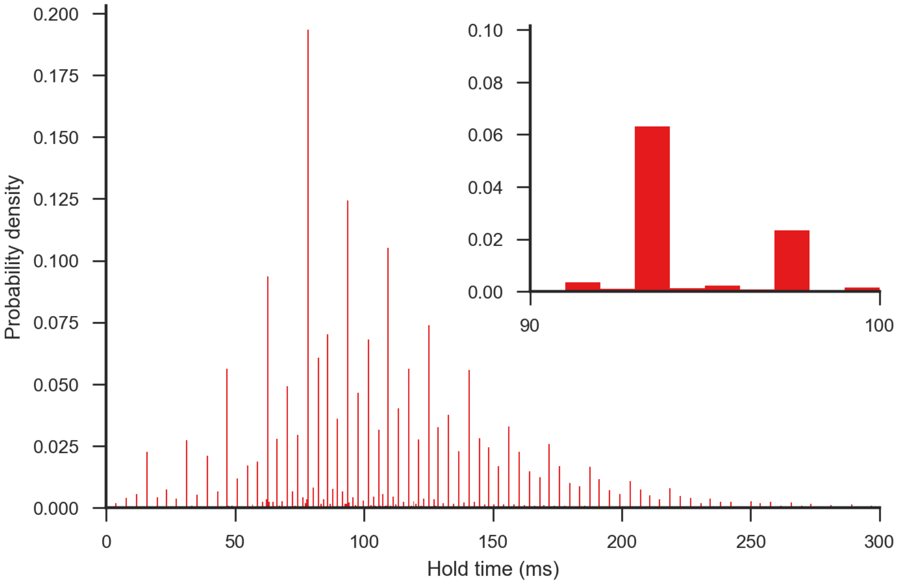

Fig. 5. Hold times for every keystroke used in the Tappy study, with a bin size of 1 ms, indicating a peculiar form of noise affecting the data. Hold times greater than 300 ms are not shown (corresponding to about 0.25% of the data). The inset plot zooms in on hold times between 90 ms and 100 ms.

Again, we begin by simply plotting the distribution of hold times analysed in the study (Fig. 5). As with the datasets associated with Ref. [4], keystroke timing is recorded to an accuracy of 3 ms. However, there appears to be some artefact affecting the recorded times, so that certain hold times are very much more likely than others. For example, a hold time of precisely 78.1 ms accounts for 9.5% of all the hold times recorded; overall, the 13 most common hold times recorded account for more than 50% of the data. Adams uses features that should not be unduly affected by the unnatural spikiness of the hold time distribution; we highlight these peculiarities for two reasons: firstly, to demonstrate the value of performing data exploration, and secondly, as a caution to researchers that future studies on similar problems may benefit from smoothing the data prior to analysis.

Ref. [5] performs the classification task of distinguishing Parkinson's sufferers from control based on both hold time and latency (the interval between pressing one key and the next). These are analysed using elementary statistical features describing the distributions, e.g. mean, standard deviation, skewness and kurtosis, giving a total of 9 features for hold time and 18 for latency. As Adams notes, given the dataset of 53 subjects (20 Parkinson's, 33 control), this large selection of features could easily lead to overfitting. As such, Linear Discriminant Analysis (LDA) is performed on each set of features as a means of dimensionality reduction to produce a single combined feature for hold time and a single combined feature for latency. Each single combined feature is then classified using an ensemble of eight separate models (Support Vector Machine, Decision Tree Classifier, K-Nearest Neighbours, etc.), the results of which are aggregated using a weighted average to produce an overall classification prediction.

We believe that, much like Ref. [4], this is an overengineered approach. The space produced by LDA is limited to one dimension (as constrained by the rank of the between-classes scatter matrix in a binary classification problem). Therefore the optimal decision criterion requires a single threshold value to be established. The use of ensemble techniques to perform such a task is unnecessary and overcomplicated.

Most importantly, however, we believe that the classification results of the study are not reproducible. Adams reports a perfect cross-validated performance, with every subject correctly classified as Parkinson's or control (AUC = 1.00). Based on our efforts to replicate the results, we find this to be wholly implausible and suspect it is an error resulting from flaws in the data acquisition or analysis. In particular, we speculate that the claimed perfect performance is the result of erroneously performing the supervised dimensionality reduction method of LDA on both the training and test data. This flaw is suggested by the description of the pre-processing stage given in Ref. [5]. If this is indeed the case then it would lead to gross overfitting of the data and hence an exaggerated AUC score for the classification task.

8 Conclusion

We have presented a critical analysis of methods proposed in Ref. [4, 5] for detecting early signs of Parkinson's disease from typing data. Whilst we believe that such work offers exciting possibilities for improved healthcare, we find the proposed methods to be overengineered and opaque. Moreover, the complexity of the neuroQWERTY index model [4] is demonstrably unnecessary: we achieve equal classification performance (AUC = 0.82) using the standard deviation as the single feature in a Logistic Regression. By performing a thorough investigation of more sophisticated time series features, we formulate the concept of mean absolute consecutive difference (MACD), which can be used as a single feature to classify the data with AUC = 0.85. Importantly, we demonstrate that such performance can be obtained from only a few hundred keystrokes, thereby achieving state of the art results while using significantly fewer samples than previous techniques. We select relevant features from a huge range of complicated time series features and find that multivariate models using up to such 10 features do not outperform the univariate model using MACD by itself — sometimes the simplest method is indeed the best.

References

1. Elbaz, A., Carcaillon, L., Kab, S., Moisan, F.: Epidemiology of Parkinson's disease. Rev. Neurol. **172**(1), 14–26 (2016)
2. Martínez-Martín, P., Gil-Nagel A., Gracia L., Gómez J., Martínez-Sarriés, J., Bermejo, F.: Unified Parkinson's disease rating scale characteristics and structure. Mov. Disord. **9**(1), 76–83 (1994)
3. Pagan, F.L.: Improving outcomes through early diagnosis of Parkinson's disease. Am. J. Manag. Care **18**(7 Suppl), S176–82 (2012)
4. Giancardo, L., Sánchez-Ferro, A., Arroyo-Gallego, T., Butterworth, I., Mendoza, C. S., Montero, P., Matarazzo, M., Obeso, J. A., Gray, M. L., Estépar, R.: Computer keyboard interaction as an indicator of early Parkinson's disease. Sci. Rep. **6**, 34468, (2016)
5. Adams, W. R.: High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing. PLoS One **12**(11), e0188226 (2017)
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
7. Arroyo-Gallego, T., Ledesma-Carbayo, M., Sánchez-Ferro, A., Butterworth, I., Sanchez-Mendoza, C., Matarazzo, M., Montero- Escribano, P., Lopez-Blanco, R., Puertas-Martín, V., Trincado, R., Giancardo, L.: Detection of Motor Impairment in Parkinson's Disease Via Mobile Touchscreen Typing. IEEE Trans. Biomed. Eng. **64**(9), 1994–2002 (2017)
8. Youden, W. J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
9. Christ, M., Kempa-Liehr, A. W. , Feindt, M.: Distributed and parallel time series feature extraction for industrial big data applications. arXiv:1610.07717 (2016)
10. tsfresh, https://github.com/blue-yonder/tsfresh. Last accessed 12 Feb 2018
11. Kaggle: raw data used to predict the onset of Parkinson's from typing tendencies, https://www.kaggle.com/valkling/tappy-keystroke-data-with-parkinsons-patients. Last accessed 1 Aug 2018