White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/127251/

Version: Accepted Version

Article:
Lacy, Stuart Edward, Smith, Stephen Leslie and Lones, Michael Adam (2018) Using Echo State Networks for Classification: A Case Study in Parkinson's Disease Diagnosis. Artificial intelligence in medicine. pp. 53-59. ISSN: 0933-3657

https://doi.org/10.1016/j.artmed.2018.02.002

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Using Echo State Networks for Classification: A Case Study in Parkinson's Disease Diagnosis

Stuart E. Lacy[a], Stephen L. Smith[a], Michael A. Lones[b]

[a] *University of York, York, United Kingdom*
[b] *Heriot-Watt University, Edinburgh, United Kingdom*

Abstract

Despite having notable advantages over established machine learning methods for time series analysis, reservoir computing methods, such as echo state networks (ESNs), have yet to be widely used for practical data mining applications. In this paper, we address this deficit with a case study that demonstrates how ESNs can be trained to predict disease labels when stimulated with movement data. Since there has been relatively little prior research into using ESNs for classification, we also consider a number of different approaches for realising input-output mappings. Our results show that ESNs can carry out effective classification and are competitive with existing approaches that have significantly longer training times, in addition to performing similarly with models employing conventional feature extraction strategies that require expert domain knowledge. This suggests that ESNs may prove beneficial in situations where predictive models must be trained rapidly and without the benefit of domain knowledge, for example on high-dimensional data produced by wearable medical technologies. This application area is emphasized with a case study of Parkinson's Disease patients who have been recorded by wearable sensors while performing basic movement tasks.

Keywords: Parkinson's Disease, Echo State Networks, neurodegenerative disease

1. Introduction

Reservoir computing is a general approach for modelling complex dynamical systems using a large Recurrent Neural Network (RNN), with only the network output weights being trained [1]. Echo State Networks (ESNs) are

a well known implementation where the output connections are fitted using simple ordinary least-squares regression [2]. Owing to this, ESNs are significantly faster and more scalable than many existing more complex machine learning approaches and are ideally suited for time series analysis. Nevertheless, despite this significant advantage, they have yet to be widely used in data mining applications.

Many medical applications have a need for predictive models that can capture the complexity of biological disease pathways to facilitate personalised healthcare. A good example is the Parkinson's disease (PD) case study considered in this paper. PD is a debilitating progressive neurodegenerative disease that presents with a broad spectrum of movement disorders, which even expert clinicians can find challenging to characterise and discriminate from other related diseases [3]. Wearable sensors can provide significant benefits to patient care by objectively measuring movement disorders in high resolution and therefore help monitor disease progress and their use is becoming increasingly widespread [4].

ESNs, with their ability to model dynamical processes, would seem like a sensible candidate for modelling such data and provide two primary benefits. The first is that they can directly model the raw time series to identify any patterns in the underlying dynamics of the signal that conventional feature extraction techniques may miss. Their second advantage is their rapid training speed, resulting from having a closed form solution. This is an important consideration for applied predictive modelling, owing to the need to train and evaluate candidate models on a range of data sets when performing model selection and evaluation.

In this paper, we consider how ESNs can be applied to the problem of diagnosing PD from movement data of the kind that might be collected using wearable accelerometers. Since there has been little existing work in this area, we focus on exploring the key issue of how inputs and outputs can be mapped to the ESN methodology, and how this affects the predictive accuracy of the model. One aspect that is investigated is whether to segment the data before inputting into the model. This has ramifications for subsequent work on analysis of data recorded from wearable sensors by facilitating simpler processing and analysis at the cost of adding more design time [5]. To evaluate the practicality of the resulting network, a two-fold comparison is performed. First, ESN classifiers are compared to models built on summary features derived using the guidance of an expert in movement disorders, in order to establish whether ESNs offer a more flexible alternative without

compromising on accuracy. The second comparison is against previous attempts on this data set, highlighting the ability of ESNs to rapidly fit an accurate model comparable with those produced from complex optimisation routines requiring significantly longer computational time.

The rest of the paper is organised as follows: Section 2 provides details of ESNs and previous applications to both classification tasks and medical problems in general, and Section 3 details the data collection process of the Parkinson's Disease movement data. The experimental methodology is laid out in Section 4, while Section 5 presents the results. Finally, Section 6 concludes.

2. Echo State Networks

2.1. Background

In recent years, a new and increasingly well researched dynamical systems approach to modelling complex time series has been developed, termed *reservoir computing*. As the name implies, the model is focused around what is known as a *reservoir*: a coupled system of non-linear functional elements in which dynamical behaviour can be modelled. Data is passed directly into the reservoir through a set of input nodes, while the output at each time step is determined by a linear readout. The functional elements are typically computational models of neurons as used by Artificial Neural Networks (ANNs); for this reason reservoir computing is considered a sub-field of ANN research. An additional defining characteristic of reservoir computing approaches is that only the reservoir readout mechanism is trained—typically using ordinary least squares— allowing for a much simpler and less computationally intensive training pipeline than found in other dynamical system modelling approaches. Two common implementations of reservoir computing are Echo State Networks (ESNs) [2] and Liquid State Machines (LSMs) [6]. ESNs typically employ a sparsely connected reservoir of sigmoidal nodes, while LSMs use a more biologically plausible neuron model by incorporating spiking neurons in the reservoir [7]. In this paper, only ESNs are considered owing to their more efficient construction and training method.

2.2. Network Configuration

ESNs comprise three distinct sets of neurons: inputs, the reservoir itself, and the output readout nodes, as shown in Figure 1. When constructing a network, the three weight matrices $\boldsymbol{W^{in}}$, $\boldsymbol{W}$, and $\boldsymbol{W^{out}}$ are initialised

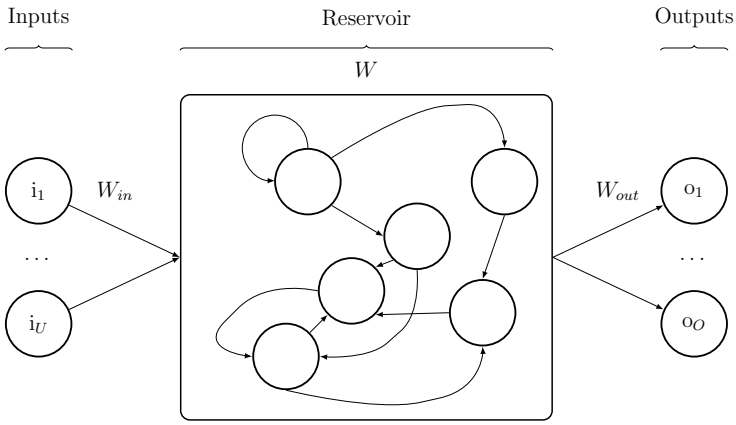

Figure 1: The reservoir of an ESN is sparsely connected with recurrent connections providing a means of maintaining state. Only the weight matrix $\boldsymbol{W^{out}}$ is optimised during training. The number of nodes in the reservoir is much larger than shown

randomly, with weights typically drawn uniformly from $[-1, 1]$. The reservoir weights are further scaled by a parameter called the spectral radius, in order to fulfil what is known as the *echo state property* [8], whereby traces of previous inputs are visible as *echoes* in each following reservoir state with a diminishing presence. The reservoir is sparsely connected to allow for a co-existing range of diverse dynamics, with only around 1% of all connections non-zero. Execution of the network is governed by two state-space equations. Equation 1 defines how the reservoir states $x(n)$ are updated at each time step, where $u(n)$ represents network inputs and $f(.)$ is a predefined activation function, typically the logistic function. By allowing a term corresponding to previous activation states, the network maintains a memory of past inputs, and so can also be used as a generative model.

$$x(n) = f(\boldsymbol{W^{in}}u(n) + \boldsymbol{W}x(n-1)) \tag{1}$$

$$y(n) = g(\boldsymbol{W^{out}}[x(n); u(n)]) \tag{2}$$

The reservoir output at a given time step $y(n)$ is governed by Equation 2, where the concatenated vector $[x(n); u(n)]$ is often referred to as the *extended state vector* $e(n)$, and a second activation function $g(.)$ is employed. Stacking the vectors across all the time-points produces matrices $\boldsymbol{Y}$ and $\boldsymbol{E}$, as in Equation 3. Estimates for $\boldsymbol{W^{out}}$ can be determined by solving ordinary

least squares regression of the target signal $\boldsymbol{D}$ as a function of $\boldsymbol{E}$. Since there is a closed form solution for ordinary least squares, the training time is extremely fast compared to algorithms that use iterative techniques such as gradient descent.

$$Y = g(\boldsymbol{W^{out}E}) \tag{3}$$

ESNs offer several advantages over traditional Recurrent Neural Networks (RNNs), mostly related to their simpler design and implementation. For example, most of the design choices when developing an ESN are concerned with the reservoir itself, including its size and the spectral radius parameter. Rather than having to consider multiple layers of functional elements, there is just a single set of active nodes, with the input and outputs providing interfaces to and from the reservoir. The second main difference with more established recurrent networks such as the Elman or Jordan models is that only the weights of the output nodes are modified during training, thereby providing a much simpler training procedure than those used for standard RNN techniques such as backpropagation through time.

However, ESNs have not seen as much uptake as could be expected given their simplicity and promising results on benchmark data sets. This may be a result of concerns about their theoretical understanding; the reservoir itself is commonly viewed as a black box, with little understanding on how the dynamics present in the input signal are being modelled. In addition, there is little guidance available on how to best tune the network parameters for a particular application, with optimal design often arising as a result of trial-and-error [9, 10]. Ozturk et al [11] cite this as a major limitation of ESN applicability, along with the argument that the primary hyper-parameter in ESN design—spectral radius—is not well correlated with network goodness-of-fit. Nevertheless, while ESNs have not been largely used in machine learning applications, they have found some value in neuroscience as models of the brain [12].

2.3. Applications

Despite the limitations associated with ESN construction described above, the field has demonstrated strong modelling capabilities when applied to problems with a temporal element. In particular they have demonstrated a strong ability to forecast chaotic time series, with considerable success on the well-known Mackey-Glass benchmark equation [13, 9]. In addition, they

have been successfully applied to real-world applications, including: classifying speech [14, 15, 16, 17], predicting stock market prices [18], modelling grammar in language tasks [19], and controlling robots [20]. However, they have rarely been used for classification problems in general, or in the specific field of medical analysis.

One exception is a study by Verplancke et al [21], predicting the likelihood that a patient would require dialysis after being admitted to the ICU given various biomarkers, including diuresis and creatinine levels. This application employed ESNs in a feature based time series classification approach, by reducing the dimensionality of the input sequence into a single scalar value, from which a classification can be made. An alternative classification method developed by Chen et al [22] uses ESNs as kernel functions in an SVM framework by fitting a reservoir to each input time series and comparing the distance between the models themselves. While this approach was shown to be competitive and more computationally efficient than the traditional Dynamic Time Warping (DTW) kernel function, it still necessitates the fitting of the linear readout weights for each time series in the data set, in addition to the overhead cost introduced by the distance calculations. Employing ESNs as a kernel under an SVM framework has also been used for forecasting, for example Shi and Han [23] adapted the Support Vector Regression technique to incorporate an ESN kernel for use in predicting chaotic time series. In their work on classifying speech patterns, Skowronski and Harris [14] use multiple reservoir readout filters to target different parts of the input time series, however, the configuration of Verplancke et al [21] required the fitting of a single set of readout weights and so remains more efficient.

In summary, reservoir computing provides significant advantages for modelling temporal data due to the use of a large reservoir providing the capability to model complex signals, while retaining a rapid training method. ESNs have been employed in several domains owing to these properties, but have been less well researched for classification problems and in medical situations in particular. In this paper, we will apply ESNs to the problem of classifying PD patients from movement data, considering multiple ways in which ESNs can be configured for classification.

Table 1: Details of the test subjects observed at each centre

Centre	Identifier	Control Subjects	PD Patients
LGI (first study)	*LGI1*	41	49
LGI (second study)	*LGI2*	29	58

3. Data Collection

3.1. Test Subjects

Two separate studies were run at Leeds General Infirmary (LGI) in Leeds, United Kingdom, to record both patients and control subjects while performing various physical movement tests. Local ethical approval was granted for both studies, with details of the centres and the number of recordings at each being found in Table 1. One of the tasks being recorded—finger tapping—is the source of the data used in this study. For ethical reasons, patients were observed while medicated, thereby increasing the difficulty of the task owing to the dampening effect of medication on motor disorder symptoms. Control subjects were age matched to the patient cohort and were screened for neurological disorders.

3.2. Finger Tapping Protocol

Finger tapping is a simple repetitive motion, whereby a person repeatedly taps their index finger against their thumb for a set duration of time. It is included in the Unified Parkinson's Disease Rating Scale (UPDRS)—the gold standard scale for staging PD—as item 3.4 [24]. Despite its simplicity, finger tapping provides an insight into a patient's condition by highlighting the cardinal symptom of PD known as bradykinesia, which is exhibited by slowness of movement, hesitations, and a reduced range of motion. The periodic nature of the task also allows for the observation of the sequence effect—a decrement in amplitude and speed of repetitive movements over a short period of time, more severe than standard physical fatigue.

Test subjects were recorded performing a subtly modified version of the finger tapping protocol laid out in the UPDRS, which contains the following instructions (emphasis added) [24]:

> "Instruct the patient to tap the index finger on the thumb 10 times as quickly **and** as big as possible."

The task was performed twice, once with the dominant hand before being repeated with the non-dominant hand. The task duration was set at thirty seconds rather than the ten tap limit enforced by the UPDRS, thereby allowing more time for bradykinesia symptoms to be exhibited. Test administrators were instructed to mark the patient's level of movement disorder out of *Normal*, *Mild*, *Moderate*, and *Severe*, with the UPDRS item for finger tapping providing the following guidance on how to form an assessment:

> "Rate each side separately, evaluating speed, amplitude, hesitations, halts and decrementing amplitude."

By digitizing the movement data, each of these aspects can be assessed objectively and in greater detail than the four-level criteria provided by the UPDRS.

3.3. Equipment

The finger tapping cycles were recorded by a pair of non-invasive electromagnetic sensors attached to the test subject's index finger and thumb as shown in Figure 2. These lightweight and small transducers record at a frequency of 60Hz with a resolution of 1.52mm, allowing for a detailed measurement of any movement disorders. To provide a measure of position, an electromagnetic signal is transmitted by a separate source unit placed approximately 50cm away, and received by the two sensors, which subsequently relay their relative positions in 3D space to an attached central processing unit. Movement is recorded in six degrees of freedom, however, only the positional data was used in this study.

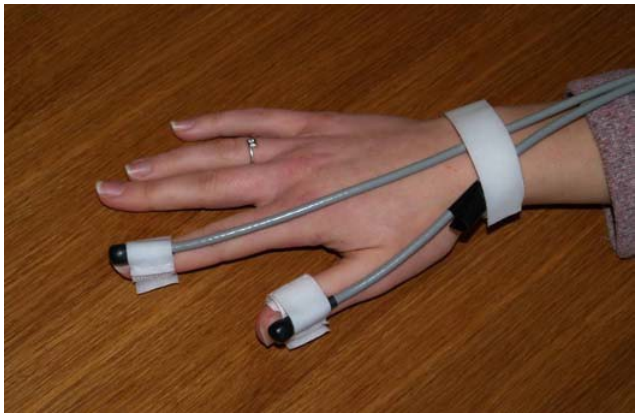

Figure 2: The position sensors used in the study were lightweight and unintrusive when performing movement tasks. Image was originally used in [25]

3.4. Data Processing

The raw data comprises $(\boldsymbol{X}_{ij}, \boldsymbol{Y}_{ij}, \boldsymbol{Z}_{ij})$ coordinate points for sensors $i \in \{1, 2\}$ and time steps $j = 1, 2, \ldots, N$, measuring the distance from the source unit. The Euclidean distance between the two sensors is calculated at each sample to obtain separation points $(\boldsymbol{x}_j^s, \boldsymbol{y}_j^s, \boldsymbol{z}_j^s)$, which are subsequently concatenated into a waveform. This separation signal is smoothed using a low-pass Butterworth filter with the cutoff frequency set at 5Hz, producing a clean sinusoidal separation trace as shown in Figure 3.

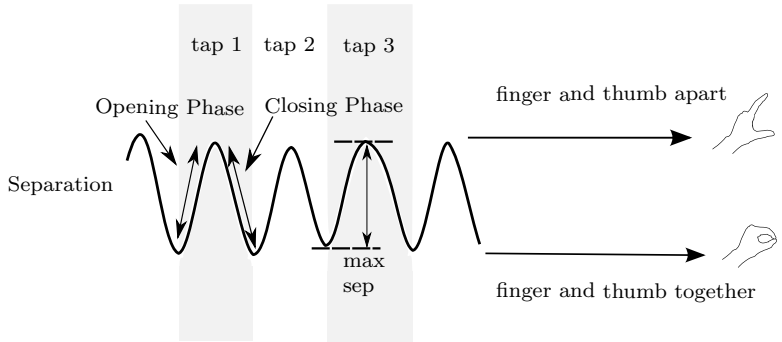

Figure 3: The finger tapping movement produces a sinusoidal wave when processed

From the separation signal, finger tapping velocity and acceleration are calculated as the first and second derivatives respectively. By locating local maxima and minima in the separation waveform, the data samples are

segmented into tapping cycles in which several identifying features can be calculated, as shown in Figure 4. An initial transient behaviour can be observed where the separation signal is not immediately identifiable as representing finger tapping behaviour, arising from the initial reaction to the cue to start the motion. To alleviate any discrepancies caused from misidentifying any taps during this period, the first tap cycle from each recording is discarded.

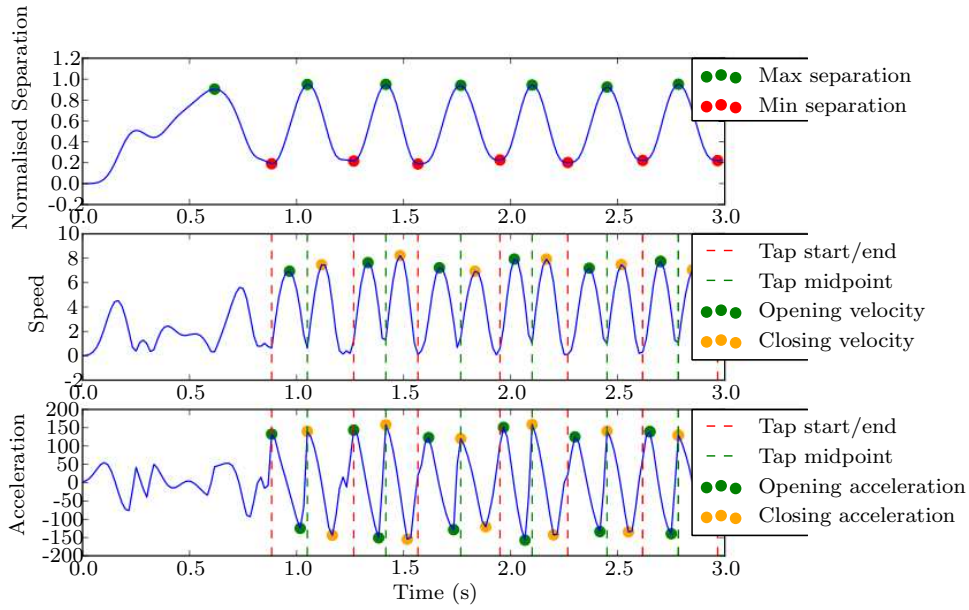

Figure 4: Identifying measures could be derived for each tap cycle from the separation trace

4. Methodology

The empirical investigation described in this paper has two principal aims: to determine the optimal configuration of ESN parameters for classification problems, and to establish whether directly modelling the separation waveform can prove as accurate as conventional models of summary features extracted with domain knowledge. This section details the experimental methodology used to explore these goals.

4.1. Configuring ESNs for Classification

There are multiple design considerations to be made when applying ESNs to classification problems, relating to both the network hyper-parameters and

the amount of data processing to perform. Potentially the most significant issue is how to produce a single output prediction for high-dimensional time series, impacting upon how $\boldsymbol{W^{out}}$ is fitted. The method used by forecasting applications of regressing on the desired target signal $\boldsymbol{D}$ by the harvested extended state matrix $\boldsymbol{E}$ (described in Section 2.1), is designed for fitting a network to a single time series with a known number of time steps n. However, when applying ESNs to classification problems there will be a number of training patterns S, each with a potentially varying number of time steps. In addition, there is no longer an appropriate length target signal $d(n)$ per pattern, instead there is a single class label. An alternative approach to classification that does not fit weights is that proposed by [26], whereby each class is represented by the principal components of its members' states at a given time-point. An input sequence is assigned a classification by identifying the similarity between its states and each candidate class's principle components. While this approach has shown promising preliminary results, in this work we will use the establish weight fitting method.

To be able to form ordinary least squares regression to fit $\boldsymbol{W^{out}}$, both $\boldsymbol{E}$ and $\boldsymbol{D}$ will need to have equal number of rows. However, $\boldsymbol{D}$ is now a vector of the class labels for each data pattern, having S rows. If the extended state matrix $\boldsymbol{E}$ was harvested from every time point from every pattern, it would have $\sum_{i=1}^{S} n_i$ entries (it cannot be assumed that every time series has equal samples). To be able to run regression, either $\boldsymbol{D}$ needs to be expanded to have $\sum_{i=1}^{S} n_i$ entries, or $\boldsymbol{E}$ needs to be collapsed to contain S rows. Regarding the first option, the target signal for each pattern could comprise the class label repeated for the number of time steps in that pattern; however, this will contain a large amount of redundant information. Instead, it would be preferable to aggregate the extended states into a single vector per training pattern.

Lukoševičius [27] recommends selecting the predicted class for each pattern as the mean of the outputs at each point in time. In practice this involves calculating $e(n)$ for each time step n, and deriving the mean values to produce a single vector of activation levels per time series. Concatenating these mean values results in a matrix $\boldsymbol{E}$ with S rows, as required to estimate values for $\boldsymbol{W^{out}}$. An alternative approach is to use the values of $e(n)$ present during the final time step to represent the overall prediction, owing to the recursive connections this will be a function of all the previous state values. This is the approach followed by Verplancke et al [21]. To better understand the characteristics of ESNs and their application to classifying movement

data, this study will investigate both of these methods.

The next issue to consider is how to process the data for input into the network. Since the movement task is periodic, it is possible to exploit this property as an inherent method of dimensionality reduction. Two approaches are taken in this work. The first simply passes the thirty second long separation waveform sample-by-sample into the reservoir via a single input neuron, with just the standard pre-processing detailed previously. The second approach is to segment the recording into its constituent tapping cycles and input these into the network one at a time, sampling the separation data from each tap at linearly separated offsets. The number of offsets to use in this tap sampling process was chosen to be twenty, calculated as sampling the tap cycle at ∼55Hz owing to the separation data having an average tap period of 0.36s, thereby not significantly down-sampling from the original recording rate of 60Hz.

The weight fitting method used in this work was ridge regression—with loss function displayed in Equation 4—whereby the value of the lambda penalty is determined by cross-validation. Using ridge regression helps to regularise the network to combat overfitting. Values for reservoir size, sparsity, and spectral radius were determined by prior investigation, with model accuracy proving relatively stable to the choice of these values. The values used in this experiment were 10% sparsity, a spectral radius of 0.4, and a reservoir size of 50 when inputting the data sample-by-sample and 20 when segmented. To reduce the effects of the initial transient behaviour of the reservoir, a number of initial values in the washout period were discarded, selected as fifty and five samples for the single sample case and the tap-by-tap input method respectively, as determined by visual inspection of the network activation levels.

$$\sum_{i=1}^{S}(\boldsymbol{D}_i - \boldsymbol{E}_i^T \boldsymbol{W}_i^{\boldsymbol{out}})^2 + \lambda \|\boldsymbol{W}_i^{\boldsymbol{out}}\|^2 \tag{4}$$

Table 2 summarises the aspects of movement disorder modelling under investigation, along with the labels used to refer to these factors for the remainder of the paper. The networks were trained on the positional data recorded from the two trials. Thirty repeats of ten-fold cross-validation were used to provide robust estimations of generalising ability, with the Area Under the Receiver Operating Characteristics Curve (AUC) used as the evaluation measure.

Table 2: Facets of ESN classification under investigation

Factor	Possible Values	
Data input type	Single sample (*single*)	Tap by tap (*tap*)
State aggregation method	Mean (*mean*)	Last state (*last*)

4.2. Comparison to Summary Features

The second goal of this study is to determine whether the ESNs are able to identify recordings from PD patients with a clinically relevant accuracy. This is assessed by comparing the networks to classifiers being input a set of seven summary features of the data, which have been captured with the assistance of an expert in neurodegenerative movement disorders. If the ESNs are able to reliably identify cases of PD without the need for domain knowledge it will have significant repercussions for the field of wearable medical technology.

As highlighted in Section 3, the thirty second recording is processed into a finger separation signal during the data collection phase. This waveform can be further segmented by the finger tapping cycles into a collection of short separation traces. In each finger tapping cycle, summary measures can be calculated, including the maximum separation, velocity, and acceleration. From these tap summary values, quantities summarising the entire recording can be determined by aggregating the tap scores in a meaningful way. The resulting seven values (summarised in Table 3) were chosen to represent multiple aspects of Parkinsonian movement disorders, including a decrement in amplitude and speed, hesitations in undergoing movements, and a decrement in amplitude of repetitive motions. Genetic Programming (GP) ensembles, which have demonstrated previous advantages in classification problems, are used as the classification model for these summary features [28, 29, 30].

An additional machine learning benchmark model is a GP approach named Temporal Expression Tree Classifiers (TETC), which have been previously applied to model movement data from the first study at LGI [31]. As with the ESNs, these models directly form a class prediction from the raw data. This technique trains symbolic regression models using a Genetic Algorithm (GA), with the separation waveform being passed into the model through a sliding input window. The advantage of this approach is that it is flexible enough to be applied to a wide range of time series data sets, only requiring the fitness function guiding the EA to be specified. The downside

Table 3: Bradykinesia features extracted from the raw data

Feature	Tap measure	Aggregation function
Tap frequency	NA	NA
Average tap separation	Peak separation	Mean
Average tap speed	Peak speed	Mean
Variability in amplitude	Peak separation	Coefficient of variation
Variability in speed	Peak speed	Coefficient of variation
Decrementing amplitude	Peak separation	Regression line gradient
Decrementing speed	Peak speed	Regression line gradient

is that it can be very slow to train, as it necessitates a long running population based search, which has to process a time series for each candidate individual.

5. Results

5.1. ESN configuration

As highlighted in Table 2, there were two factors related to ESN configuration for classification under investigation: whether to pass the separation waveform into the network sample-by-sample or segmented into tap cycles, and how to aggregate the extended states from each time step to form a single predicted output. Two possibilities were investigated for each of these choices: *mean* and *last* state aggregation functions and *single* and *taps* data processing method. The results of these experiments are displayed in Table 4 with the mean of the thirty cross-validated AUCs displayed along with 95% confidence intervals.

Table 4: Comparison of ESN configurations for classifying periodic time series

Network input	Stage aggregation	Mean AUC	CI
single	last	0.558	0.551 - 0.565
single	mean	0.575	0.567 - 0.583
taps	last	0.695	0.689 - 0.701
taps	mean	0.802	0.797 - 0.807

Immediately from the table, the impact of each of these factors can be seen. On both data sets, calculating the overall state of the network for

each time series as the *mean* of the states at each time step is shown to be significantly advantageous over using the final state value ($p < 0.001$ for *taps* input method and $p < 0.01$ for *single*). This finding is perhaps unexpected; the final reservoir state is a function of every previous value and could thereby be expected to provide a prediction taking into account the entire recording the same way the *mean* method does. In addition, under the *taps* network input method, the *mean* function does not take into account the ordering of the tap inputs directly, except that each state is a function of its previous value. A potential explanation for this finding is that the *last* method places greater emphasis on the final states, which, due to the increased task duration, could be indicating that both controls and patients are suffering from physical fatigue.

The second issue under investigation is whether to segment the data before passing it into the network. The results indicate that the networks achieve greater discriminatory ability by identifying trends within tap cycles, rather than inputting the signal sample-by-sample. Hypothesis testing showed this difference to be statistically significant at the 0.1% level for both state aggregation methods. This is a useful finding since less computational time is required to train networks due to the lowered number of time steps, with only a minimal increase in data processing time. Interestingly, the experimental setup which takes most advantage of the temporal nature of the data is the *single last* combination, however, this configuration produces the least accurate models overall, suggesting that the ESNs are not fully exploiting the recurrent connections to analyse the time dimension.

5.2. Comparison with bradykinesia features

Table 5: Comparison of ESN classifers, TETCs, and models of bradykinesia features

Model	Mean AUC	CI
bk	0.852	0.848 - 0.857
esn	0.802	0.797 - 0.807
tetc	0.791	0.786 - 0.795

To place the results of the ESN classifiers in context, Table 5 displays the mean AUC from the thirty repeats of ten-fold cross validation for a variety of models, including: the ESN classifier built using the optimal *mean* and *taps* configuration as identified above (labelled *esn*), the classifier modelling

summary features of bradykinesia (*bk*), and the GP expression tree approach previously used (*tetc*). Overall, the models formed of the summary bradykinesia features score most highly, suggesting that when domain knowledge is available it is preferable to use it. However, this can be time consuming, inflexible, and not always possible. On such occasions, both the GP and the ESN approach offer competitive models without requiring a large subjective pre-processing stage, aside from segmenting the periodic data into its constituent tap cycles.

Between the two biologically-inspired computational techniques, the ESN classifiers are slightly more accurate overall; their AUC of 0.802 on unseen data suggests clinically relevant diagnostic potential. To demonstrate this, an example ESN classifier was fitted on a training set during cross-validation, with its ROC curve on the unseen fold being plotted in Figure 5. On this particular validation fold, the model had an AUC of 0.84, and if one chose the operating point of the model as the highlighted point on the curve that equally maximises both the specificity and sensitivity, then it would have a sensitivity of 0.84 and a specificity of 0.75.

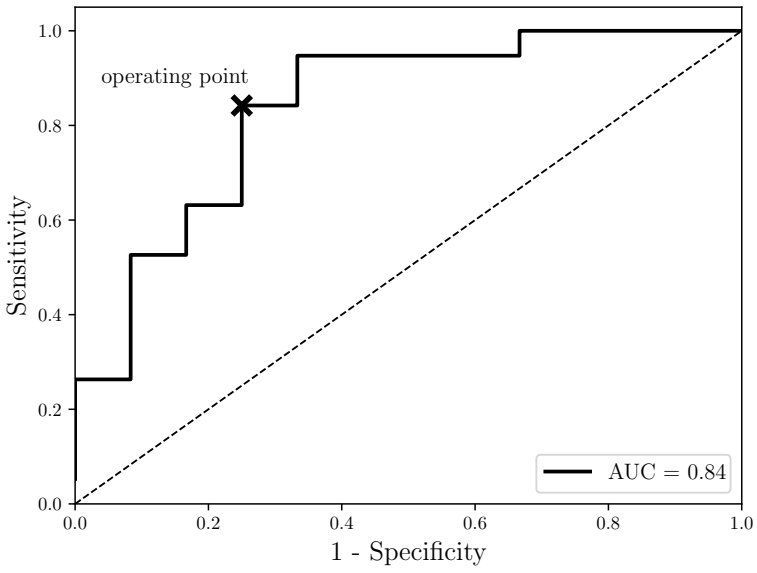

Figure 5: The ROC curve of an example ESN classifier as assessed on an independent validation set

Hypothesis testing (Wilcoxon signed-rank test with Holm multiple-correction adjustment) showed no significant difference between the ESN and TETC approaches, although both were significantly different from the models of bradykinesia features at the 0.1% level.

The primary difference between these two techniques is in terms of their suitability to applied situations, where training models with complex training methods (such as TETCs) can be computationally expensive, while ESNs are far quicker to train owing to the existence of a closed form solution for the output weights. A rapid training time provides significant benefit for model selection and evaluation, and allows the classifier to to be more flexible and to react more quickly to new data sources. This is a particularly attractive property in the application of wearable sensors discussed in this paper, where high-dimensional data sets can be computationally intensive to process and the sensor network can be expanded online. For example, if a new movement sensor is added to the system, an ESN model can rapidly incorporate this new stream as an additional network input. On the other hand, a classifier based on extracted features would necessitate waiting until a domain expert has intervened and potentially derived new summary measures using this additional data source.

6. Conclusions

The results of this empirical investigation have shown that the use of ESNs to model positional data is a viable technique for medical diagnosis of a movement disorder, with accuracies comparable to previous approaches that rely upon more complex training algorithms such as EAs. The primary advantage of ESNs is their rapid training time, which proves most advantageous when performing model selection over a variety of hyper-parameters. In addition, while they are not are as accurate as models formed using features extracted with expert supervision, the ESNs offer practical advantages with regards to their flexibility to adapt to new data sources, as well as having the scope to identify complex patterns in the data that are not identified by the summary measures. Typically, ESNs have been implemented in traditional signal forecasting applications; this paper has demonstrated their ability to perform classification provided careful configuration. Based on the results of this study, we recommend forming the final extended state matrix used for fitting the output weights as the mean of the extended state matrix for each individual time series. For our periodic data, a simple means of dimensional-

ity reduction by segmenting the waveform into its cycles proved advantageous over passing the signal in one sample at a time. While segmenting the data is easily implementable for a periodic movement such as finger-tapping, this approach can be adapted to other recordings of wearable sensors by splitting the data either by a suitable time period or by some measure relevant to the process that is producing the data. As the use of wearable sensors becomes more widespread, it is likely that segmentation will become valuable as a dimensionality reduction technique to facilitate simpler analysis.

As technology is progressing with transducers becoming increasingly smaller and non-invasive, there is a rising demand for personalised healthcare provided by wearable devices, and therefore a need for accurate time series models. Ideally, expert domain knowledge would be used to develop models capable of providing this functionality; however, as wearable technology becomes more prevalent and the number of applications increase, it will require significant human time to provide such analysis. In this work we have demonstrated that for this application, ESNs are capable of providing accurate classification models, comparable to those requiring expert movement disorder knowledge and pre-processing time to extract descriptive summary measures.

Bibliography

[1] M. Lukoševičius, H. Jaeger, Reservoir computing approaches to recurrent neural network training, Computer Science Review 3 (3) (2009) 127–149.

[2] H. Jaeger, The "echo state" approach to analysing and training recurrent neural networks-with an erratum note, Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148 (2001) 34.

[3] J. Jankovic, Parkinsons disease: clinical features and diagnosis, Journal of Neurology, Neurosurgery & Psychiatry 79 (4) (2008) 368–376.

[4] S. C. Mukhopadhyay, Wearable sensors for human activity monitoring: A review, Sensors Journal, IEEE 15 (3) (2015) 1321–1330.

[5] M. A. Lones, J. E. Alty, J. Cosgrove, S. Jamieson, S. L. Smith, Going through directional changes: evolving human movement classifiers using

an event based encoding, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM, 2017, pp. 1365–1371.

[6] W. Maass, T. Natschläger, H. Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations, Neural computation 14 (11) (2002) 2531–2560.

[7] T. Yamazaki, S. Tanaka, The cerebellum as a liquid state machine, Neural Networks 20 (3) (2007) 290–297.

[8] I. B. Yildiz, H. Jaeger, S. J. Kiebel, Re-visiting the echo state property, Neural networks 35 (2012) 1–9.

[9] D. Prokhorov, Echo state networks: appeal and challenges, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Vol. 3, IEEE, 2005, pp. 1463–1466.

[10] A. Rodan, P. Tino, Minimum complexity echo state network, IEEE transactions on neural networks 22 (1) (2011) 131–144.

[11] M. C. Ozturk, D. Xu, J. C. Príncipe, Analysis and design of echo state networks, Neural Computation 19 (1) (2007) 111–138.

[12] O. Barak, D. Sussillo, R. Romo, M. Tsodyks, L. Abbott, From fixed points to chaos: three models of delayed discrimination, Progress in neurobiology 103 (2013) 214–222.

[13] H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science 304 (5667) (2004) 78–80.

[14] M. D. Skowronski, J. G. Harris, Minimum mean squared error time series classification using an echo state network prediction model, in: IEEE International Symposium on Circuits and Systems, 2006, IEEE, 2006.

[15] M. D. Skowronski, J. G. Harris, Automatic speech recognition using a predictive echo state network classifier, Neural networks 20 (3) (2007) 414–423.

[16] D. Verstraeten, B. Schrauwen, D. Stroobandt, J. Van Campenhout, Isolated word recognition with the liquid state machine: a case study, Information Processing Letters 95 (6) (2005) 521–528.

[17] W. Maass, T. Natschläger, H. Markram, A model for real-time computation in generic neural microcircuits, in: Proceedings of the 15th International Conference on Neural Information Processing Systems, 2002, pp. 229–236.

[18] X. Lin, Z. Yang, Y. Song, Short-term stock price prediction based on echo state networks, Expert systems with applications 36 (3) (2009) 7313–7317.

[19] M. H. Tong, A. D. Bickett, E. M. Christiansen, G. W. Cottrell, Learning grammatical structure with echo state networks, Neural Networks 20 (3) (2007) 424–432.

[20] M. Salmen, P. G. Ploger, Echo state networks used for motor control, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2005, IEEE, 2005, pp. 1953–1958.

[21] T. Verplancke, S. Van Looy, K. Steurbaut, D. Benoit, F. De Turck, G. De Moor, J. Decruyenaere, A novel time series analysis approach for prediction of dialysis in critically ill patients using echo-state networks, BMC medical informatics and decision making 10 (1) (2010) 4.

[22] H. Chen, F. Tang, P. Tino, X. Yao, Model-based kernel for efficient time series analysis, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 392–400.

[23] Z. Shi, M. Han, Support vector echo-state machine for chaotic time-series prediction, IEEE Transactions on Neural Networks 18 (2) (2007) 359–372.

[24] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results, Movement Disorders 23 (15) (2008) 2129–2170.

[25] J. A. Edgar, MEng Project Report: The application of evolutionary algorithms towards the Diagnosis of Parkinson's Disease, Tech. rep., Department of Electronics, University of York (June 2007).

[26] A. Prater, Spatiotemporal signal classification via principal components of reservoir states, Neural Networks 91 (2017) 66–75.

[27] M. Lukoševičius, A practical guide to applying echo state networks, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 659–686.

[28] S. E. Lacy, M. A. Lones, S. L. Smith, Forming classifier ensembles with multimodal evolutionary algorithms, in: Proc. of the 2015 IEEE Congress on Evolutionary Computation, IEEE, 2015, pp. 723–729.

[29] S. E. Lacy, M. A. Lones, S. L. Smith, A comparison of evolved linear and non-linear ensemble vote aggregators, in: Proc. of the 2015 IEEE Congress on Evolutionary Computation, IEEE, 2015, pp. 758–763.

[30] S. E. Lacy, M. A. Lones, S. L. Smith, J. E. Alty, D. Jamieson, K. L. Possin, N. Schuff, Characterisation of movement disorder in Parkinson's disease using evolutionary algorithms, in: Proceeding of the fifteenth annual conference companion on Genetic and evolutionary computation conference companion, ACM, 2013, pp. 1479–1486.

[31] M. Lones, S. L. Smith, J. E. Alty, S. E. Lacy, K. L. Possin, D. Jamieson, A. M. Tyrrell, et al., Evolving classifiers to recognize the movement characteristics of Parkinson's disease patients, Evolutionary Computation, IEEE Transactions on 18 (4) (2014) 559–576.