REGULAR PAPER

Designing compact features for remote stroke rehabilitation monitoring using wearable accelerometers

Xi Chen[1] · Yu Guan[2] · Jian Qing Shi[3] · Xiu-Li Du[4] · Janet Eyre[5]

Abstract

Stroke is known as a major global health problem, and for stroke survivors it is key to monitor the recovery levels. However, traditional stroke rehabilitation assessment methods (such as the popular clinical assessment) can be subjective and expensive, and it is also less convenient for patients to visit clinics in a high frequency. To address this issue, in this work based on wearable sensing and machine learning techniques, we develop an automated system that can predict the assessment score in an objective manner. With wrist-worn sensors, accelerometer data is collected from 59 stroke survivors in free-living environments for a duration of 8 weeks, and we map the week-wise accelerometer data (3 days per week) to the assessment score by developing signal processing and predictive model pipeline. To achieve this, we propose two types of new features, which can encode the rehabilitation information from both paralysed and non-paralysed sides while suppressing the high-level noises such as irrelevant daily activities. Based on the proposed features, we further develop the longitudinal mixed-effects model with Gaussian process prior (LMGP), which can model the random effects caused by different subjects and time slots (during the 8 weeks). Comprehensive experiments are conducted to evaluate our system on both acute and chronic patients, and the promising results suggest its effectiveness.

Keywords Wrist-worn accelerometer sensor · Stroke rehabilitation · CAHAI score · Regression model

✉ Yu Guan
yu.guan@warwick.ac.uk

Xi Chen
chenxi@hainanbank.com.cn

Jian Qing Shi
shijq@sustech.edu.cn

Xiu-Li Du
duxiuli@njnu.edu.cn

Janet Eyre
janet.eyre@ncl.ac.uk

1 Hainan Rural Credit Union, Hainan, China

2 Department of Computer Science, University of Warwick, Coventry, UK

3 Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China

4 School of Mathematical Sciences, Nanjing Normal University, Nanjing, China

5 Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

1 Introduction

It is widely known that stroke is a worldwide health problem causing disability and death (Donnan et al. 2008), and it occurs when a blood clot cuts off oxygen supply to a region of the brain. Hemiparesis is a very common symptom of post-stroke that is the fractional or intact paralysis of one side of the body, i.e., the opposite side to where the blood clot occurs, and it results in difficulties in performing activities, e.g., with reduced arm movement. Patients can recover some of their capabilities with intense therapeutic input, so it is important to assess their recovery levels in time. There are many approaches to assess patients' recovery levels including brain imaging (Wintermark et al. 2005), questionnaire-based (Pietro et al. 2007), and lab-based clinical assessment (Barreca et al. 2005).

The brain imaging technique, is deemed as one of the most reliable approach, which can provide the information of brain hemodynamics (Wintermark et al. 2005). However, this approach requires special equipment and is very expensive in cost. Questionnaire-based approaches investigate the functional ability during a period using questionnaires, and

it can be categorised into two types: patient-completed and caregiver-completed (Pietro et al. 2007). Although it is much cheaper than brain imaging approaches, it may contain high-level of bias. For instance, patients may not remember their daily activities (i.e.,recall bias); the caregivers may not be able to observe the patient all the time. These biases make questionnaire-based approaches less precise. Lab-based clinical assessment approaches (Barreca et al. 2005), on the other hand, provide an alternative solution. The patients' upper limb functionality will be assessed by clinicians, e.g., by observing patients' capabilities of finishing certain pre-defined activities (Barreca et al. 2005). Compared with braining imaging or questionnaire-based approaches, the cost of lab-based clinical assessment approaches is reasonable with high accuracy. However, this assessment is normally taken in clinics/hospitals, which is not convenient for the patients, making continuous monitoring less feasible.

In this work, we aim to build an automated stroke rehabilitation assessment system using wearable sensing and machine learning techniques. Different from the aforementioned approaches, our system can measure the patients objectively and continuously in free-living environments. We collect accelerometer data using wrist-worn accelerometer sensors, and design compact features that can capture rehabilitation-related movements, before mapping these features to clinical assessment scores (i.e., the model training process). The trained model can be used to infer recovery-level for other unknown patients. In free living environments, there are different types of movements which may be related to different frequencies. For example, activities such as running or jumping may correspond to high-frequency signals, while sedentary or eating may be low-frequency signals. In this study, instead of recognising the daily activities explicitly, which is hard to achieve given limited annotation (e.g., without frame/sample-wise annotation), we transform the raw accelerometer data to the frequency domain, where we design features that can encode the rehabilitation-related movements. Specifically, wavelet transform (Walden et al. 2000) is used, and the wavelet coefficients can represent the particular frequency information at certain decomposition scales. In Preece et al. (2009), provide some commonly used wavelet features extracted from accelerometer data. However, to capture stroke rehabilitation-related activities, some domain knowledge should be taken into account to design better features. After stroke, patients have difficulties in moving one side (i.e., paralysed side) due to the brain injury, and data from paralysed side tends to describe more about the upper limb functional ability, than the non-paralysed side (i.e., normal side). However, such signals can be significantly affected by personal behaviours or irrelevant daily activities, and such noises should be suppressed before developing the predictive models. Various wavelet features were studied, and we propose two new types of

daily-activity-invariant features that can encode information from both paralysed/non-paralysed sides, before developing predictive models for stroke rehabilitation assessment. Specifically, in this work our contributions can be summarised as follows:

- *Stroke-rehab-driven features*: We propose two new types of compact wavelet-based features that can encode information from both paralysed and non-paralysed sides to represent upper limb functional abilities for stroke rehabilitation assessment. It can significantly suppress the influences of personal behaviours or irrelevant daily activities for data collected in the noisy free-living environment.
- *Automated assessment system*: Based on the proposed stroke-rehab-driven features, we developed the automated system by using the longitudinal mixed-effects model with Gaussian process prior (LMGP). Various predictive models were studied, and we find LMGP can model the random effects caused by the heterogeneity nature among subjects in a 8-week longitudinal study.
- *Comprehensive evaluation*: Comprehensive experiments are designed to study the effectiveness of our system. We comprehensively studied the feature subset on modelling the mixed-effects of LMGP. Compared with other approaches, the results suggest the effectiveness of the proposed system on both acute and chronic patients.

2 Background and related work

As described in Sect. 1, lab-based clinical assessment is one of the most effective stroke rehabilitation assessment methods. In this section, we introduce the lab-based approach named Chedoke Arm and Hand Activity Inventory (CAHAI) scoring (Barreca et al. 2006), based on which our automated system can be developed. Some sensing and machine learning techniques for automated health assessment are also reviewed in this section.

2.1 Chedoke arm and hand activity inventory (CAHAI)

CAHAI scoring is a clinical assessment method for stroke rehabilitation, and it is a fully validated measure Barreca et al. (2006) of upper limb functional ability with 9 tasks which are scored by using a 7-point quantitative scale. In the assessment, the patient will be asked to perform 9 tasks, including opening a jar of coffee, drawing a line with a ruler, calling 911, etc. and the clinician will score these behaviours based on patient's performance at a scale from 1 (total assist weak) to 7 (complete independence i.e., timely, safely) (Barreca et al. 2006). A task example "call 911" is shown in

Fig. 1. Thus the minimum and maximum summation scores are 7 and 63 respectively. A CAHAI score form can be found in Fig. 12 in Appendix 2.

2.2 Automated behaviour assessment using wearables

Recently, wearable sensing and machine learning (ML) techniques are comprehensively studied for automated health assessment. Compared with the traditional assessment approaches (e.g., via self-reporting, clinical assessment, etc.) which are normally subjective and expensive, the automated systems may provide an objective, low-cost alternative, which can also be used for continuous monitoring/assessment. Some automated systems are developed to assess the behaviours of diseases such as Parkinson's disease (zia ur et al. 2019; Hammerla et al. 2015), autism (Ploetz et al. 2012), depression (Little et al. 2020); or to monitor the health status such as sleep (Zhai et al. 2020; Supratak et al. 2017), fatigue (Bai et al. 2020; Ibrahim et al. 2020), or recover-level from surgery (Ratcliffe et al. 2020; Gurchiek et al. 2019), etc.

After collecting behaviour or physiological signals (e.g., accelerometers, ECG, audio, etc.), assessment/monitoring models can be developed. For application with high interpretability requirement, feature engineering can be a crucial step. For example, with gait parameters extracted from IMU sensors (such as stride, velocity, etc.), one can build simple ML models (e.g., random forest) for Parkinson's disease classification (zia ur et al. 2019) or fatigue score regression (Ibrahim et al. 2020). Compared with the redundant IMU data, gait parameters are more compact and interpretable, making it suitable for clinical applications. However, designing interpretable/clinically-relevant features can be a

Behaviour Observation for CAHAI Scoring

1. Open jar of coffee
2. Call 911
3. Draw a line with a ruler
4. Pour a glass of water
5. Wring out wash cloth
6. Do up five buttons
7. Dry back with towel
8. Put toothpaste on toothbrush
9. Cut medium resistance putty

Fig. 1 The clinical behaviour assessment for CAHAI scoring Barreca et al. (2006)

time-consuming process, which may also require domain knowledge (Zhai et al. 2020; Ibrahim et al. 2020; zia ur et al. 2019; Ratcliffe et al. 2020; Gurchiek et al. 2019).

On the other hand, when interpretability is less required, deep learning can be an alternative approach, which can be directly applied to the raw signal (Supratak et al. 2017) or engineered features (Hammerla et al. 2015; Zhai et al. 2020; Bai et al. 2020; Little et al. 2020 for (high-level) representation learning and classification/regression tasks. However, it normally requires adequate data annotation for better model generalisation.

2.3 Sensing techniques for automated stroke rehabilitation monitoring

With the rapid development of the sensing/ML techniques, researchers also start to apply various sensors for stroke rehabilitation monitoring. In Dolatabadi et al. (2017), Kinect sensor is used in a home-like environments to detect the key joints such that stroke patients' behaviour can be assessed. In Ganesh et al. (2018), a wireless surface Electromyography (sEMG) device is used to monitor the muscle recruitment of the post-stroke patients to see the effect of orthotic intervention. In clinical environments, five wearable sensors are placed on the trunk, upper and forearm of the two upper limbs to measure the reaching behaviours of the stroke survivors (Jung et al. 2018). To monitor motor functions of stroke patients during rehabilitation sessions at clinics, an ecosystem including a jack and a cube for hand grasping monitoring, as well as a smart watch for arm dynamic monitoring was designed (Bobin et al. 2019). These techniques can objectively assess/measure the behaviours of the stroke patients, yet they are either limited to clinical environments (Bobin et al. 2019; Jung et al. 2018; Ganesh et al. 2018 or constrained environments [e.g., in front of a camera Dolatabadi et al. (2017)].

Most recently, wrist-worn sensors are used for stroke rehabilitation monitoring for patients in free-living environment (Halloran et al. 2019; Tang et al. 2020). In each trial, 3-day accelerometer data are collected from both wrists (with a trial-wise annotation, i.e., CAHAI score), and for both works (Halloran et al. 2019; Tang et al. 2020) data analysis is performed using the sliding window approach. To reduce the data redundancy of the raw data, PCA features are extracted from each window (Halloran et al. 2019; Tang et al. 2020). Moreover, due to the lack of window-wise annotation, in Halloran et al. (2019) pseudo label is assigned to each window such that a random forest regressor can be trained, while in Tang et al. (2020) Gaussian Mixture Models (GMM) clustering approach is employed to learn the holistic trial-wise representation, before developing the regression model. Both methods (Halloran et al. 2019; Tang et al. 2020) suffer from the lack of annotation. In Halloran

et al. (2019), pseudo labeling is introduced, yet the trained model is affected by the introduced label noise. In Tang et al. (2020), the application of GMM clustering (on the sliding windows) makes it computationally expensive to large data, and the trained model does not generalise well to unseen subjects.

In our work, by analysing the nature of the paralysed/non-paralysed sides, we design stroke-rehab-driven features which can directly encode the long accelerometer sequence (e.g., a trial with 3-day accelerometer data) into a very compact representation. The features are expected to emphasis the stroke-related behaviours while suppressing the irrelevant activities. Based on the proposed features, a predictive model that is adaptive to different subjects/time-slots can be developed using LMGP (Shi et al. 2012) for CAHAI score prediction.

3 Methodology

In this section, we introduce our method from data collection, data pre-processing, feature design to predictive models. Our aim is to develop an automated model which can map the free-living 3-day accelerometer data into the CAHAI score. With the trained model, we can automatically infer the CAHAI score in an objective and continuous manner. To achieve this, we first reduce the data redundancy via preprocessing and design compact and discriminant features. Given the proposed features, a longitudinal mixed-effects model with Gaussian Process prior (LMGP) is used (Shi et al. 2012), which can further reduce the impact of large variability (caused by different subjects and time slots) for higher prediction results.

3.1 Data acquisition

3.1.1 Participants

Data is collected as part of a bigger research study which aims to use a bespoke, professionally-written video game as a therapeutic tool for stroke rehabilitation (Shi et al. 2013). Ethical approval is obtained from the National Research Ethics Committee and all work undertaken is in accordance with the Declaration of Helsinki. Written, informed consent from all the subjects is obtained. A cohort of 59 stroke survivors, without significant cognitive or visual impairment, are recruited for the study. Patients were divided into two groups, i.e.,

- *Group 1*: the acute patient group, consisting of 26 participants who enrolled into the study within 6 months after stroke;

- *Group 2*: the chronic patient group, including 33 participants who were 6 months or more post onset of stroke.

The distributions of acute/chronic condition, gender, dominant/non-dominant hand, paralysed/non-paralysed side with respect to age are shown in Fig. 2.

These 59 patients visit the clinic for the CAHAI scoring every week (a random day in weekdays) for a duration of 8 weeks. In the 8 weeks, they are asked to wear two wrist-worn sensors for 3 full days (including night time) a week. They are also advised to remove the device during shower or swimming. Since some patients need time to get familiar to this data collection procedure, for better data quality we do not use the first week's accelerometer data. The first week's CAHAI scores are used as medical history information.

3.1.2 Data collection

In contrast to other afore-mentioned sensing techniques (Jung et al. 2018; Bobin et al. 2019; Ganesh et al. 2018; Dolatabadi et al. 2017), in this study we collect the accelerometer data from wrist-worn sensors in free-living environments. The sensor used for this study, i.e., AX3 (Axivity Ltd 2020), is a triaxial accelerometer logger that is designed for physical activity/behaviour monitoring, and it has been widely used in the medical community [e.g., for the UK Biobank physical activity study Doherty et al. 2017]. The wrist bands are also designed such that the users can comfortably wear it without affecting their behaviours. The data is collected at 100Hz sampling rate, which can well preserve the daily activities of human being (Bouten et al. 1997). Different from human activity recognition which requires sample-wise or frame-wise annotation (Guan and Ploetz 2017; Ploetz and Guan 2018), the data collection in this study is relatively straight-forward. The patients put on both wrist-worn sensors 3 full days a week, before visiting clinicians for CAHAI scoring (i.e., week-wise annotation). In other words, we aim to use accelerometer data captured in free-living environments to represent the stroke survivors' upper limb activities to measure the degree of paresis (Stig JÃzgensen et al. 1999) (i.e., CAHAI score).

One problem with most commercial sensors is that only summary data (e.g., step count from fitbit), instead of raw data, are available. The algorithms of producing summary data are normally non-open source, and may vary from vendor to vendor—making the data collection and analysis device-dependent, and thus less practical in terms of generalisation and scalability. The AX3 device used in this study, on the other hand, outputs the raw acceleration information in x, y, z directions. It is simple and transparent, making the collected data re-usable, which is crucial for research communities.
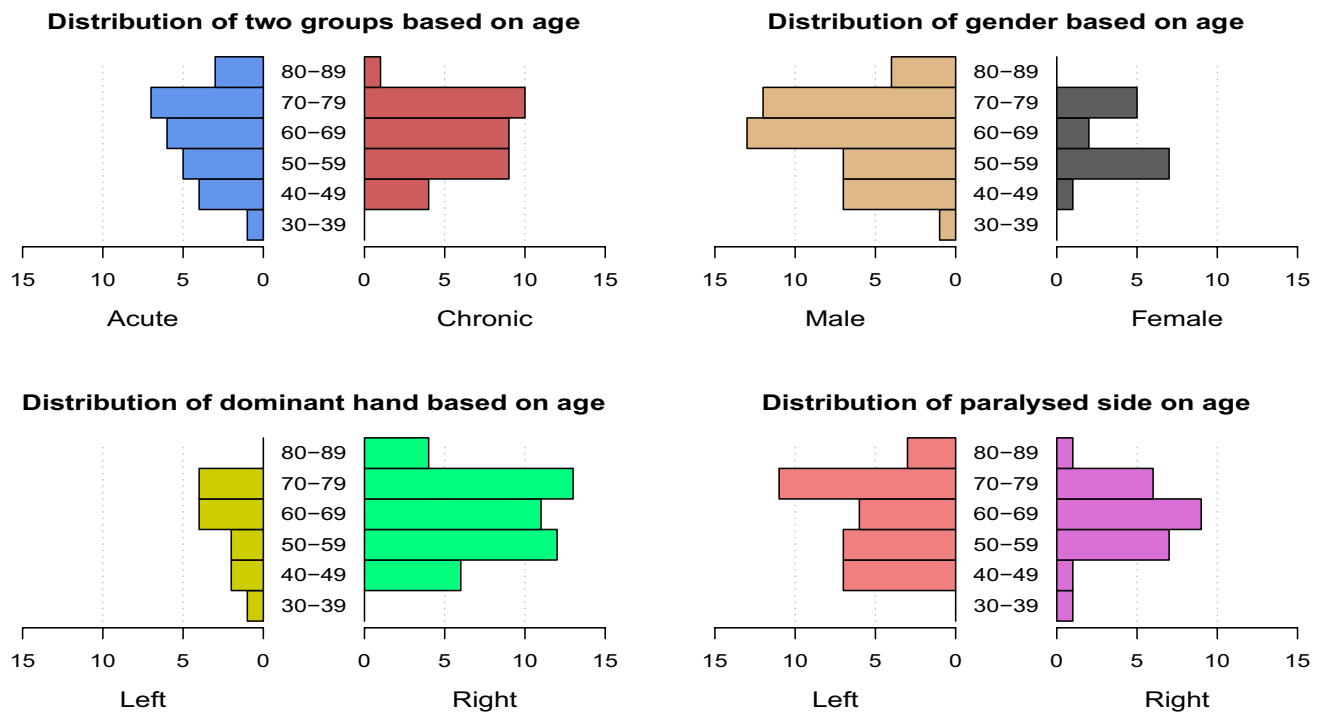
Fig. 2 Demographic information of the collected dataset (with 59 subjects): the distributions of acute/chronic condition, gender, dominant/non-dominant hand, paralysed/non-paralysed side with respect to age

3.2 Data pre-processing

For accelerometer data, signal vector magnitude (VM) (Karantonis et al. 2006) is a popular representation, which is simply the magnitude of the triaxial acceleration data defined as $a(t) = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)}$, where $a_x(t)$, $a_y(t)$, $a_z(t)$ are the acceleration along the x, y, z axes at timestamp t. The gravity effect can be removed by $VM(t) = |a(t) - 1|$. Because its simplicity and effectiveness, VM has been widely used in health monitoring tasks, such as fall detection (Karantonis et al. 2006), physical activity monitoring (Doherty et al. 2017), perinatal stroke assessment (Gao et al. 2019), etc. To further reduce the data volume, we used second-wise VM, i.e., the mean VM over each second (including 100 samples per second) will be used as new representation. Some second-wise VM examples (from two patients) can be found in Fig. 3.

3.3 The proposed stroke-rehab-driven features

3.3.1 Challenges

We aim to build a model that can map the 3-day time-series data to the CAHAI score. Different from other

wearable-based behaviour analysis tasks (e.g., Ploetz et al. 2012; Guan and Ploetz 2017), the annotation here is inadequate. Even if we used the second-wise VM data, each trial still included roughly 3 days × 24h/day × 3600s/h = 259200 samples (a.k.a. timestamps) with one annotation (i.e., CAHAI score). In contrast to the popular deep learning based human activity recognition approaches, which can be trained when with rich annotations (in frame-wise or sample-wise level), the lack of annotation makes it hard to learn effective representation directly (using machine/deep learning) from the raw data. Moreover, since the data is collected in free-living environments, and the 3 full days (per week) can be taken in weekdays or weekends, which may increase the intra-subject variability significantly, making it hard to model. To address the afore-mentioned issues, domain knowledge driven feature engineering may play a major role in extracting compact and discriminant signatures.

3.3.2 Wavelet features

For time-series analysis, wavelet analysis is a powerful tool to represent various aspects of non-stationary signals such as trends, discontinuities, and repeated patterns (Ayachi et al. 2016; Walden et al. 2000; Preece et al. 2009), which is especially useful in signal compression or noise reduction. Given its properties, wavelet features have been widely used in accelerometer-based daily living
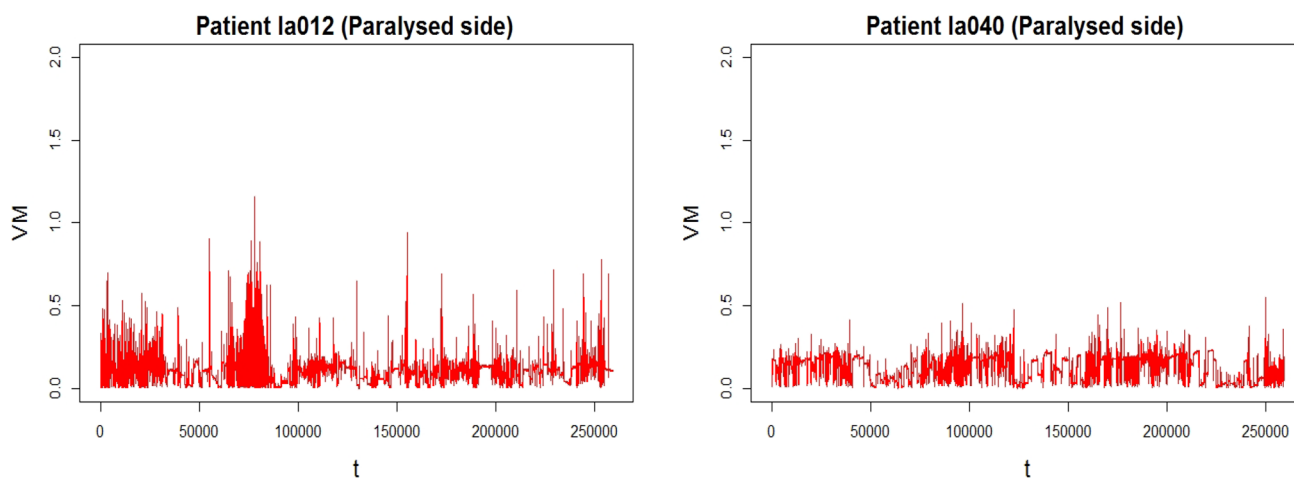
Fig. 3 The signal vector magnitude (VM) data collected from two patients (on the paralysed side); Patient la012 has a CAHAI score of 55, while Patient la040 has a CAHAI score of 26

activity analytics (Ayachi et al. 2016). In this work, we use discrete wavelet transform (DWT) and discrete wavelet packet transform (DWPT) as feature extractors, based on which new features were designed to preserve the stroke rehabilitation-related information. More details of DWT and DWPT can be found at Appendix 3.

After applying the DWT and DWPT, VM signals can be transformed to the wavelet coefficients at different decomposition scales. In this work, DWT coefficients at scales $\{2, 3, 4, 5, 6, 7\}$ and DWPT at scales $\{1.1, 1.2, 1.3, 1.4\}$ are employed, and the corresponding normalised Sum of Absolute value of the coefficients at different Decomposition scales (referred to as SAD features) are used as new representation. Specifically, SAD includes DWPT features defined as

$$
\begin{aligned}
SAD_{1.1} &= \frac{\|\mathbf{W}_{3.4}\|_1}{N/2^3} = 2^3 \frac{\|\mathbf{W}_{3.4}\|_1}{N}, \\
SAD_{1.2} &= \frac{\|\mathbf{W}_{3.5}\|_1}{N/2^3} = 2^3 \frac{\|\mathbf{W}_{3.5}\|_1}{N}, \\
SAD_{1.3} &= \frac{\|\mathbf{W}_{3.6}\|_1}{N/2^3} = 2^3 \frac{\|\mathbf{W}_{3.6}\|_1}{N}, \\
SAD_{1.4} &= \frac{\|\mathbf{W}_{3.7}\|_1}{N/2^3} = 2^3 \frac{\|\mathbf{W}_{3.7}\|_1}{N},
\end{aligned}
\tag{1}
$$

and DWT features defined as

$$
SAD_j = \frac{\|\mathbf{W}_j\|_1}{N/2^j} = 2^j \frac{\|\mathbf{W}_j\|_1}{N}, \qquad j = 2, 3, 4, 5, 6, 7,
\tag{2}
$$

where $\mathbf{W}$ presents the wavelet coefficients and N presents the length of the VM data. More technical details of DWT, DWPT, as well as the scale selection can be found in Appendix 4.

Through wavelet transformation, the long sequence (e.g., VM data in Fig. 3) can be transformed into compact SAD representation (i.e., 10-dimensional feature vector, with entries listed in Eqs. 1 and 2). In Fig. 4, we visualise compact SAD features corresponding to the paralysed sides of two patients (i.e., patients la012 and la040 from Fig. 3). We notice in the SAD feature space, it is not easy to distinguish the paralysed sides from these two different patients (in terms of CAHAI), indicating the necessity of developing more advanced stroke-related features (e.g., by also considering the non-paralysed side).
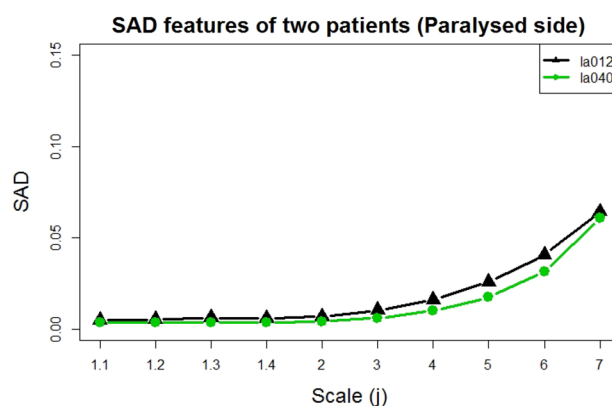

Fig. 4 10-dimensional SAD features extracted from the paralysed side of two patients (with different CAHAI scores); They exhibit similar patterns, indicating the necessity of developing more informative stroke-related features

3.3.3 Proposed features

Based on the compact SAD representation, we aim to further design effective features for reliable CAHAI score regression. In Figs. 3 and 4, we visualise the behaviour patterns in different feature spaces. Specifically, we plot the paralysed side of patient la012 (with CAHAI score 55), and la040 (with CAHAI 26) using VM representation (Fig. 3) and SAD representation (Fig. 4). From both figures, we can see the limitations of both representations. Although VM can demonstrate distinct patterns from both patients, it may be also related to the large intra-class variability (e.g., personalised behaviour patterns). Moreover, the redundancy as well as the high-dimensionality make it hard for modelling. On the other hand, SAD has low dimensionality, yet both patients exhibited high-level of similarity, indicating that SAD of the paralysed side alone is not enough for distinguishing patients with different recovery levels.

Given the observations, we further visualise SAD features from both paralysed/non-paralysed sides for both patients in Fig. 5. We can see patient la012 (with high recovery level) uses both hands (almost) equally while patient la040 (with low recovery level) tends to use the non-paralysed side more. These observations motivate us to design new features using both sides, instead of the paralysed side alone. In this work, we propose two types of features that combine both Paralysed side and Non-Paralysed side, namely 1) PNP^1 that encodes the ratio information with entries defined as:

$$PNP_k^1 = \frac{SAD_k^p}{SAD_k^{np}} \tag{3}$$

as well as its variant 2) **PNP^2** with entries defined as:

$$PNP_k^2 = \frac{SAD_k^{np} - SAD_k^p}{SAD_k^{np} + SAD_k^p}, \tag{4}$$

where k represents the scales defined in SAD features (as shown in Eqs. 1 and 2); p and np refer to the paralysed side and non-paralysed side respectively. We also visualise patient la012 and patient la040 using the new proposed features PNP^1 and PNP^2 in Fig. 6, from which we can see the proposed features can well distinguish these two patients, in contrast to SAD (Fig. 4). Although the proposed PNP features empirically exhibit the desired properties (i.e., compact and informative) for two patients, it should be pointed out that larger scale experiments should be conducted to evaluate the generalisation capability, which will be provided in the experimental section.

We summarise the procedure of generating PNP features as follows:

1. Given 3-day raw accelerometer data, calculating the signal vector magnitude (VM) with the gravity effect removed;
2. calculating the second-wise VM (mean VM value for each second) as the new representation;
3. calculating DWPT features at scales {1.1, 1.2, 1.3, 1.4} and DWT features at scales {2, 3, 4, 5, 6, 7}
4. given the DWPT and DWT features, calculating the 10-dimensional SAD features via Eqs. (1) and (2).
5. given SAD features, calculating the two proposed PNP^1 and PNP^2 features, via Eqs. (3) and (4).

We list 4 types of features, i.e., the original wavelet features extracted from paralysed (SAD^p) and non-paralysed sides (SAD^{np}) separately, as well as the two new proposed
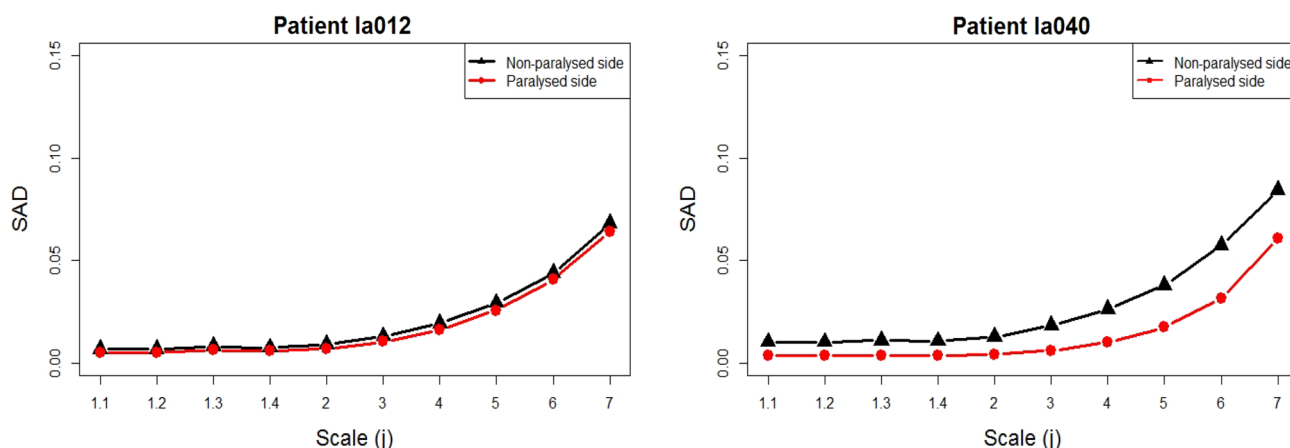

Fig. 5 SAD representation with both paralysed/non-paralysed sides from two different patients (la012 with CAHAI score 55, and la040 CAHAI score 26). SAD features from the non-paralysed side may contain discriminant information for stroke-rehab modelling
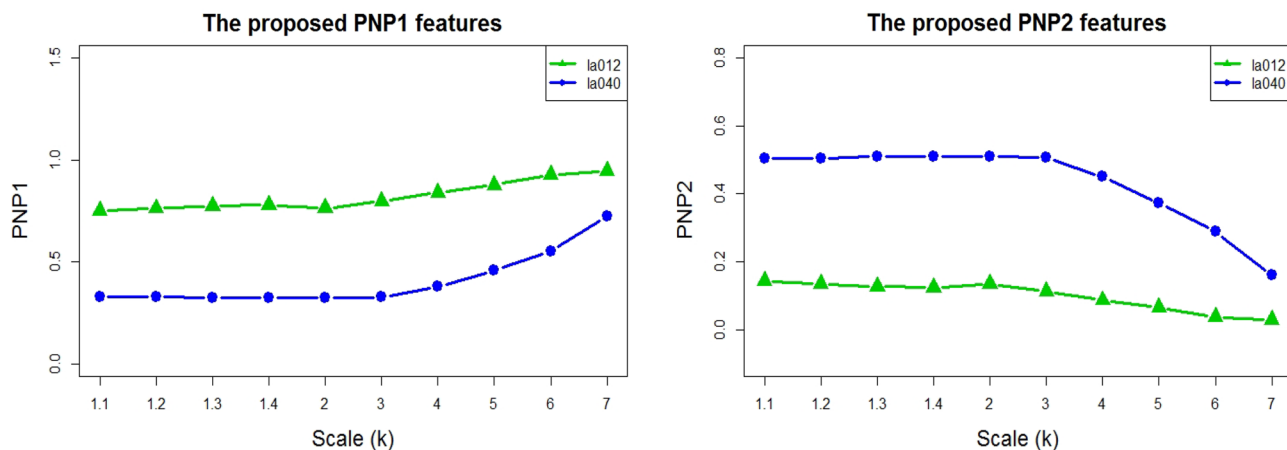
Fig. 6 Two proposed PNP representations for two patients(la012, and la040), which can provide discriminant information in distinguishing the patients with different recovery levels (clinical CAHAI score)

Table 1 The proposed rehab-driven features

Feature type	Feature entries for each type
SAD^p	$SAD^p_{1.1}, SAD^p_{1.2}, SAD^p_{1.3}, SAD^p_{1.4}, SAD^p_2, SAD^p_3, ..., SAD^p_7$
SAD^{np}	$SAD^{np}_{1.1}, SAD^{np}_{1.2}, SAD^{np}_{1.3}, SAD^{np}_{1.4}, SAD^{np}_2, SAD^{np}_3, ..., SAD^{np}_7$
PNP^1	$PNP^1_{1.1}, PNP^1_{1.2}, PNP^1_{1.3}, PNP^1_{1.4}, PNP^1_2, PNP^1_3, ..., PNP^1_7$
PNP^2	$PNP^2_{1.1}, PNP^2_{1.2}, PNP^2_{1.3}, PNP^2_{1.4}, PNP^2_2, PNP^2_3, ..., PNP^2_7$

features (PNP^1 and PNP^2). Based on 10 scales, we can form 40-dimensional feature vector, as shown in Table 1. However, there exist certain level of noises and redundancy (especially on SAD^p, and SAD^{np}), so it is crucial to develop feature selection mechanism or powerful prediction models for higher performance.

3.4 Predictive models

Based on the proposed representation, we aim to develop predictive models that can map features to the CAHAI score. Although we reduce the data redundancy significantly, there still exist data noises, which may encode irrelevant information. It is crucial to develop robust mechanism to select the most relevant features, and here we use a popular feature selection linear model (LASSO). To model the nonlinear random effects in the longitudinal study, we also propose to use the longitudinal mixed-effects model with Gaussian Process prior (LMGP).

It is worth noting that our model will also take advantage of the medical history information (i.e., CAHAI score during the first visit) to predict CAHAI scores for the rest 7 weeks

(i.e., week 2– week 8). From the perspective of practical application, CAHAI score from the initial week (referred to as *ini*) may be used as an important normalisation factor for different individuals.

3.4.1 The linear fixed-effects model

Since there may exist some redundant or irrelevant features for the prediction task, first we propose to use LASSO (Least Absolute Shrinkage and Selection Operator) for feature selection.

Given the 41-dimensional input variables (40 wavelet features listed in Table 1 and one CAHAI score from the initial week), first we standardise the data using z-norm, and each feature entry x_k will be normalised as $x_k^{new} = (x_k - \bar{x})/s_k$, where $\bar{x}$ and s_k are the mean and standard deviation of the kth feature. Based on the aforementioned model, namely LASSO, useful features can be selected, based on which prediction model can be developed. For simplicity, we first use linear model to predict the target CAHAI score y_i:

$$y_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{ij}, \ \epsilon_{ij} \sim N(0, \sigma^2), \tag{5}$$

where i stands for the ith trial/visit (during week 2–week 8) and j represents the jth patients; $\boldsymbol{x}_{ij}$ represents the selected feature vector; $\boldsymbol{\beta}$ are the model parameter vector to be estimated, and ϵ_{ij} is the random noise term.

3.4.2 Longitudinal mixed-effects model with Gaussian process prior (LMGP)

It is simple to use linear model for CAHAI score prediction. However, it ignores the heterogeneity nature among subjects in this longitudinal study. To model the heterogeneity, we propose to use a nonlinear mixed-effects model

(Shi et al. [2012]), which consists of the fixed-effects part and random-effects part. Specifically, the random-effects part contributes mainly on modelling the heterogeneity, making the the prediction process subject/time-adaptive for longitudinal studies. The longitudinal mixed-effects model with Gaussian Process prior (LMGP) is defined as follows:

$$y_{i,j} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + g(\boldsymbol{\phi}_{ij}) + \epsilon_{ij}, \ \epsilon_{ij} \sim N(0, \sigma^2), \tag{6}$$

where i,j stand for the ith patient at the jth visit (from week 2 to week 8); ϵ_{ij} refers to as independent random error and σ^2 is its variance. In Eq. ([6]), $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ is the fixed-effects part and $g(\boldsymbol{\phi}_{ij})$ represents the nonlinear random-effects part, and the latter can be modelled using a non-parametric Bayesian approach with a GP prior (Shi et al. [2012]).

It is worth noting that in LMGP the fixed-effects part $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ explains a linear relationship between input features and CAHAI, while the random-effects part $g(\boldsymbol{\phi}_{ij})$ is used to explain the variability caused by differences among individuals or time slots during different weeks. By considering both parts, LMGP provides a solution of personalised modelling for this longitudinal data analysis. In LMGP, it is important to select input features to model both parts, and we refer them to as fixed-effects features and random-effects features, respectively. The effect of the fixed-effects features will be studied in the experimental evaluation section.

For LMGP training, we first ignore the random-effects part, and only optimise the parameters $\hat{\boldsymbol{\beta}}$ of the fixed-effects part (via ordinary least squares OLS); With estimated parameters $\hat{\boldsymbol{\beta}}$, the residual $r_{ij} = y_{ij} - \boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} = g(\boldsymbol{\phi}_{ij}) + \epsilon_{ij}$ can be calculated, from which we can model the random-effects

$$g(\boldsymbol{\phi}_{i,j}) \sim GP(0, K(\cdot, \cdot; \boldsymbol{\theta})).$$

In this paper we choose $K(\cdot, \cdot; \boldsymbol{\theta})$ as the following three different kernels (linear, squared exponential and rational quadratic), and here we take the squared exponential as an example. The squared exponential (covariance) kernel function is defined as : $K(\boldsymbol{\phi}, \boldsymbol{\phi}'; \boldsymbol{\theta}) = v_0 \exp\left\{-d(\boldsymbol{\phi}, \boldsymbol{\phi}')/2\right\}$ where $d(\boldsymbol{\phi}, \boldsymbol{\phi}') = \sum_{q=1}^{Q} w_q \left(\phi_{i,j,q} - \phi'_{i,j,q}\right)^2$ is an extended distance between $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$. It involves the hyper-parameters $\boldsymbol{\theta} = (v_0, w_1, \dots, w_Q)$. In Bayesian approach, we may choose the value of those parameters based on prior knowledge. It is however a difficult task due to the large dimension of $\boldsymbol{\theta}$. We used an empirical Bayesian method.

The training procedure include two steps. (I) Estimate $\boldsymbol{\beta}$ and σ in Eq. ([5]); (II) Estimate the values of the hyper-parameters $\boldsymbol{\theta}$ by an empirical Bayesian method, i.e. maximise the marginal likelihood from $\boldsymbol{r}_i \sim N(\boldsymbol{0}, \boldsymbol{C}_i + \sigma^2\boldsymbol{I})$ for $i = 1, \dots, n$, where $\boldsymbol{C}_i \in \mathbb{R}^{J \times J}$ is the covariance matrix of $g(\cdot)$, and its

element is defined by $K(\boldsymbol{\phi}_{i,j}, \boldsymbol{\phi}_{i,j'}; \boldsymbol{\theta})$. To obtain a more accurate results, an iterative method may be used. Except the initial step, the error item in ([5]) used in step I is replaced by

$$\boldsymbol{\epsilon}_i = (\epsilon_1, \dots, \epsilon_J) \sim N(\boldsymbol{0}, \boldsymbol{C}_i + \sigma^2\boldsymbol{I}))$$

where all the parameters are evaluated by using the values obtained in the previous iteration.

The calculation of the prediction is relatively easy. The posterior distribution of $g(\boldsymbol{\phi}_i)$ is a multivariate normal with mean $\mathbf{C}(\mathbf{C} + \sigma^2\mathbf{I})^{-1}\boldsymbol{r}_i$ and the variance $\sigma^2\mathbf{C}(\mathbf{C} + \sigma^2\mathbf{I})^{-1}$.

The fitted value can therefore be calculated by the sum of $\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}$ and the above posterior mean. The variance can be calculated accordingly. The detailed description can be found in Shi et al. ([2011]).

4 Experimental evaluation

In this section, several experiments are designed to evaluate the proposed features as well as the prediction systems. The patients are splitted into two groups according to the disease nature, i.e., the acute patient group (26 subjects) and the chronic patient group (33 subjects). Experiments are conducted on both group separately.

Specifically for each group, leave one subject out cross validation(LOSO-CV) is applied. That is, for a certain group (acute or chronic) with n subjects, in each iteration 1 subject was used as test set while the rest $n - 1$ subjects were used for training. This procedure is repeated n times to test all the n subjects and average prediction performance (i.e., the mean predicted CAHAI) will be reported.

Since CAHAI score prediction is a typical regression problem, we use the root mean square error (RMSE) as the evaluation metric, and lower mean RMSE values indicate better performance.

4.1 Evaluation of the proposed feature PNP

In this subsection, we evaluate the effectiveness of the proposed PNP features. One most straight-forward approach is to calculate the correlation coefficients against the target CAHAI scores. In Table [2] we report the corresponding correlation coefficients (PNP_k^1, and PNP_k^2 in 10 scales) for acute/chronic patients group. The correlation coefficients of the original wavelet features (with paralysed side SAD_k^p, and non-paralysed side SAD_k^{np} in 10 scales) against CAHAI score are also reported for comparison. From Table [2], we can see:

- PNP features generally have higher correlation coefficients (than SAD) against the CAHAI scores.

Table 2 Correlation coefficients of the wavelet features and CAHAI score

Scale (k)	SAD_k^p	Acute SAD_k^{np}	Patients PNP_k^1	PNP_k^2	SAD_k^p	Chronic SAD_k^{np}	Patients PNP_k^1	PNP_k^2
k = 1.1	−0.41	0.32	0.68	−0.70	0.22	0.49	0.56	−0.56
k = 1.2	−0.42	0.33	0.69	−0.71	0.24	0.50	0.57	−0.56
k = 1.3	−0.43	0.32	0.70	−0.72	0.23	0.51	0.58	−0.57
k = 1.4	−0.42	0.33	0.69	−0.71	0.24	0.51	0.57	−0.57
k = 2	−0.42	0.31	0.69	−0.71	0.23	0.50	0.56	−0.55
k = 3	−0.42	0.27	0.67	−0.68	0.25	0.50	0.53	−0.52
k = 4	−0.43	0.20	0.60	−0.63	0.26	0.50	0.48	−0.47
k = 5	−0.42	0.10	0.49	−0.52	0.27	0.50	0.43	−0.42
k = 6	−0.37	−0.01	0.35	−0.38	0.27	0.48	0.35	−0.34
k = 7	−0.30	−0.10	0.19	−0.20	0.28	0.45	0.25	−0.24

- for PNP features, from Scale $k = 1.1$ to $k = 5$ there are higher correlations against the CAHAI scores.
- for chronic patients, SAD features (on the non-paralysed side) exhibit comparable correlation scores with PNP features.

These observations indicate the necessities of selecting useful features on building the prediction system. Although PNP demonstrates more powerful prediction capacity, in some cases, SAD (e.g., extracted from the non-paralysed side) may also provide important information for a certain population (e.g., chronic patients).

For better understanding the relationship between these features, we also report the cross-correlation between each feature pairs. Noting we also include the medical history feature, i.e., the initial week-1 CAHAI score. From Fig. 7, and we have the following observations:

- For both patient groups, the PNP features are highly correlated. PNP features within the same type (PNP^1 or PNP^2) tend to be positively correlated, while PNP features from different types tend to be negatively correlated.
- For acute patients, SAD features for each side (paralysed side SAD^p or non-paralysed side SAD^{np}) are highly (positively) correlated, yet the SAD features from different sides are less correlated. For chronic patients, however, SAD features from both sides are highly (positively) correlated.
- In general, PNP features, SAD features and the medical history information *ini* are less correlated, indicating them as potentially complementary information to be fused.

Based on the above findings, it is clear that within each feature types, there may exist high-level of feature redundancy, and it is necessary to select the most relevant

feature subsets. For acute and chronic patient groups, the optimal feature subset may vary due to the different movement patterns (e.g., on paralysed/non-paralysed sides). Although the proposed PNP features can alleviate this problem to some extent, it is beneficial to combine the less correlated features (i.e., PNP, SAD, and *ini*).

4.2 Evaluation of the predictive models

4.2.1 Feature selection

Based on the feature correlation analysis in Sect. 4.1, it is important we select the most relevant features from various sources (i.e., PNP, SAD, and *ini*). Different from the correlation-based approach which can select each feature independently (by the correlation coefficient), LASSO can select the features by solving a linear optimisation problem with sparsity constraint, and it takes the relationship of the features into consideration. Based on LASSO we select the most important features for both acute/chronic patients, as shown in Table 3.

It is also worth mentioning that the wavelet-based features can bring certain levels of interpretability. SAD_j represents the point energy in the signal at the decomposition level j based on the energy preserving condition (see Appendix 4 for more details). Specifically, it relates to the degree of energy among the different activity levels (in different frequency domain based on the decomposition scale j). The activities such as jumping or lifting an object may correspond to high-frequency signal, while sedentary or eating may be low-frequency signal. Based on these, we can interpret the key features in Table 3. For example, for acute patients key features (which is high-related to stroke-rehab modelling) correspond to asymmetric activities in low/medium-frequency level (i.e., with PNP_3^2, PNP_6^1), non-paralysed-based activities in low/medium-frequency level (i.e.,
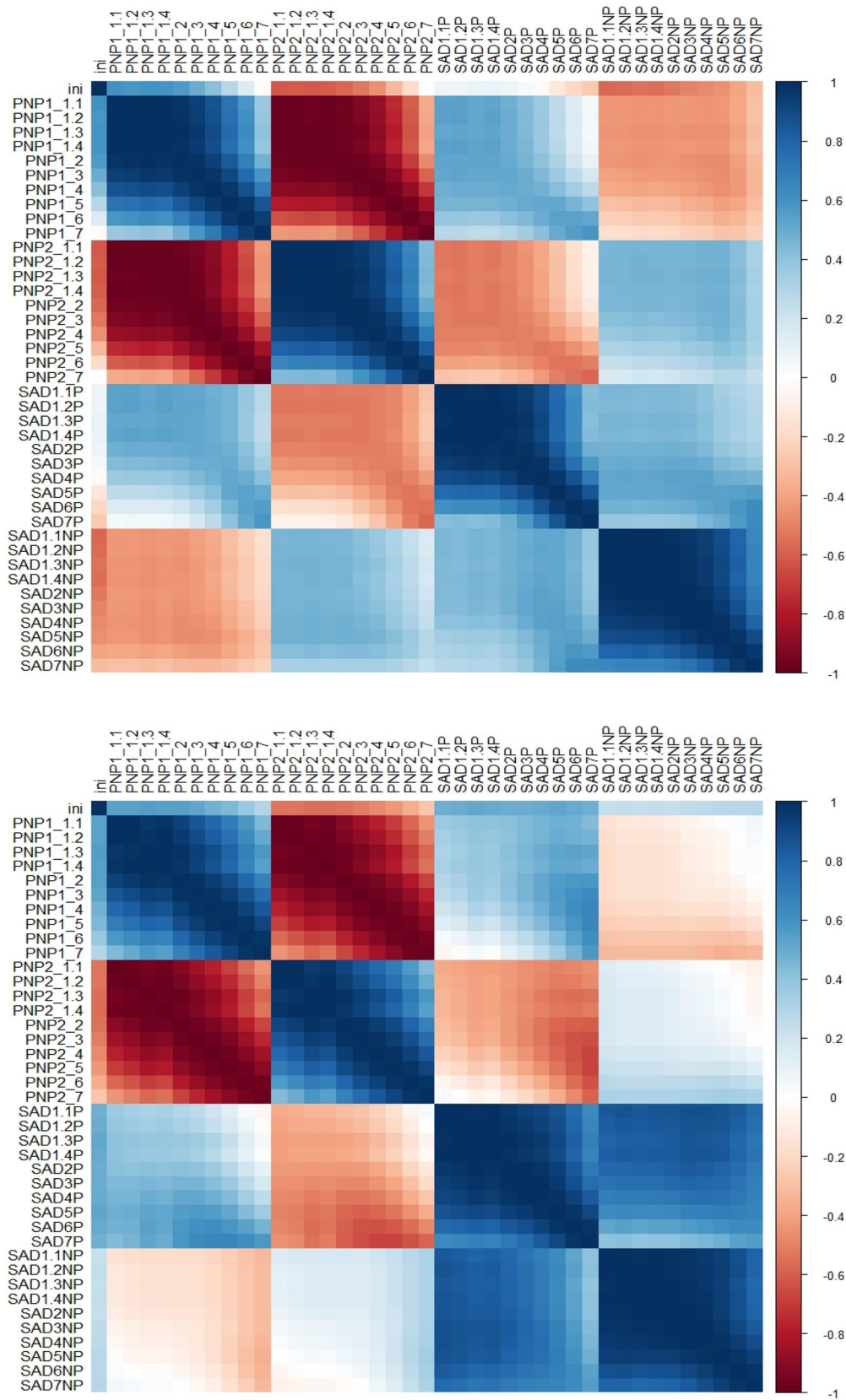
Fig. 7 Cross-correlation of the candidate features for two patient groups (top: acute patients; bottom: chronic patients). In general, PNP features, SAD features and the medical history information *ini* are less correlated, compared with within-feature correlation (e.g., within PNP features)

Table 3 Selected features using LASSO

Acute patients	Chronic patients
PNP_3^2, PNP_6^1, SAD_2^{np}, $SAD_{1.2}^p$ SAD_6^{np}, ini	$PNP_{1.4}^1$, SAD_4^p, SAD_2^{np}, $PNP_{1.3}^2$ PNP_4^1, $PNP_{1.1}^2$, ini, PNP_6^1 $SAD_{1.4}^{np}$, SAD_6^{np}
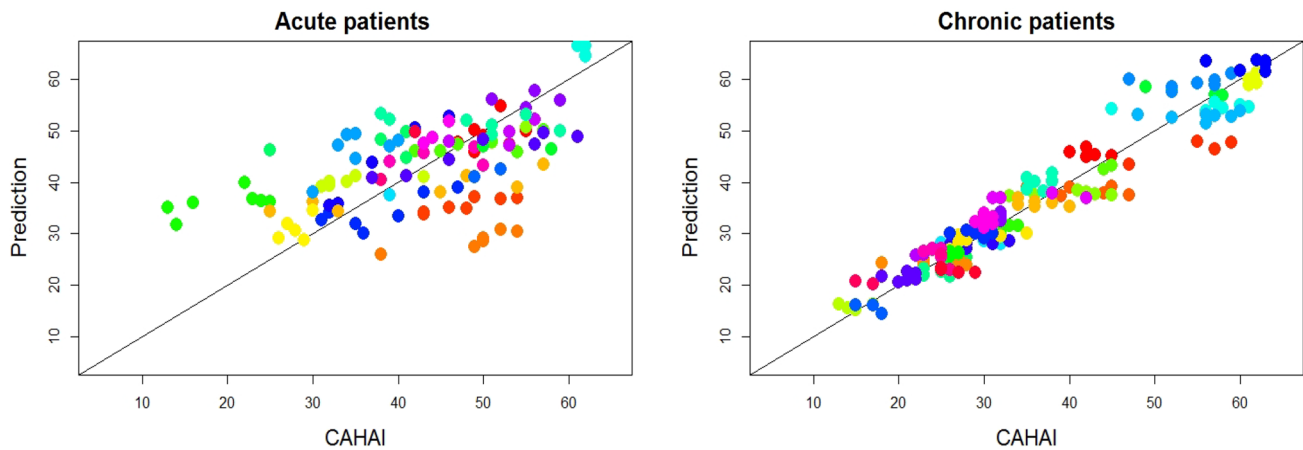
Fig. 8 Linear model prediction vs clinical CAHAI; Left: Acute patients (RMSE 7.24); Right: Chronic patients (RMSE 3.29). Each point corresponds to a trial (i.e., data collected from 3 days), and different colours represent different subjects
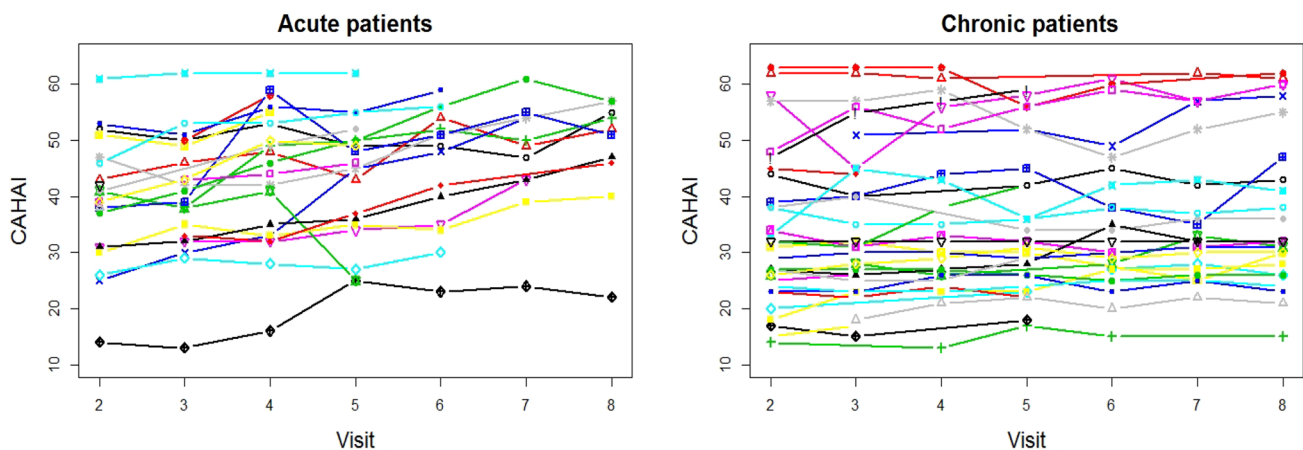
Fig. 9 Clinical assessed CAHAI distribution with respect to visit; stroke rehabilitation levels may be stable for chronic patient while may vary substantially for acute patients

with SAD_2^{np}, SAD_6^{np}), and paralysed-side based activities in high-frequency level (i.e., with $SAD_{1.2}^p$).

4.2.2 Performance of linear fixed-effects model

Based on the selected features, we perform leave-one-patient-out cross validation on these two patient groups respectively using the linear fixed-effects model. As shown in Fig. 8, the prediction results of the chronic patients (with mean RMSE 3.29) tend to be much better than the ones of the acute group (with mean RMSE 7.24). One of the main reasons might be the nature of the patient group. In Fig. 9, we plot the clinical CAHAI distribution (i.e., the ground truth CAHAI) from week 2 to week 8, and we can see the clinical CAHAI scores are very stable for chronic patients. On the other hand, for acute patients who suffered from stroke in the past 6 months, their health statuses were less stable and affected significantly by various factors, and in this case the simple linear fixed-effected model yields less promising results.
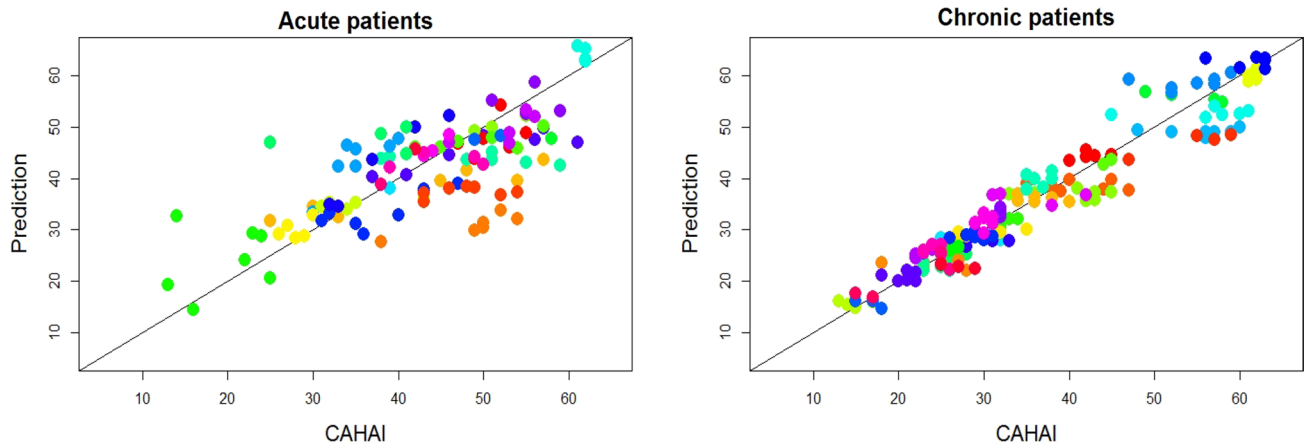
Fig. 10 LMGP prediction vs clinical CAHAI; Left: Acute patients (RMSE 5.75); Right: Chronic patients (RMSE 3.12). Each point corresponds to a trial (i.e., data collected from 3 days), and different colours represent different subjects

4.2.3 Performance of Longitudinal mixed-effects model with Gaussian process prior (LMGP)

We also develop LMGP for both patient groups. We have applied different covariance kernels in LMGP models and found the one with powered exponential kernel achieves the

best results. The following discussion will therefore focus on the model with this kernel. More results of using other kernels can be found in Appendix 5.

Here, we use the selected features (from Table 3) as the fixed-effects features and random-effects features. Similar to the linear fixed-effects model, we evaluate the performance
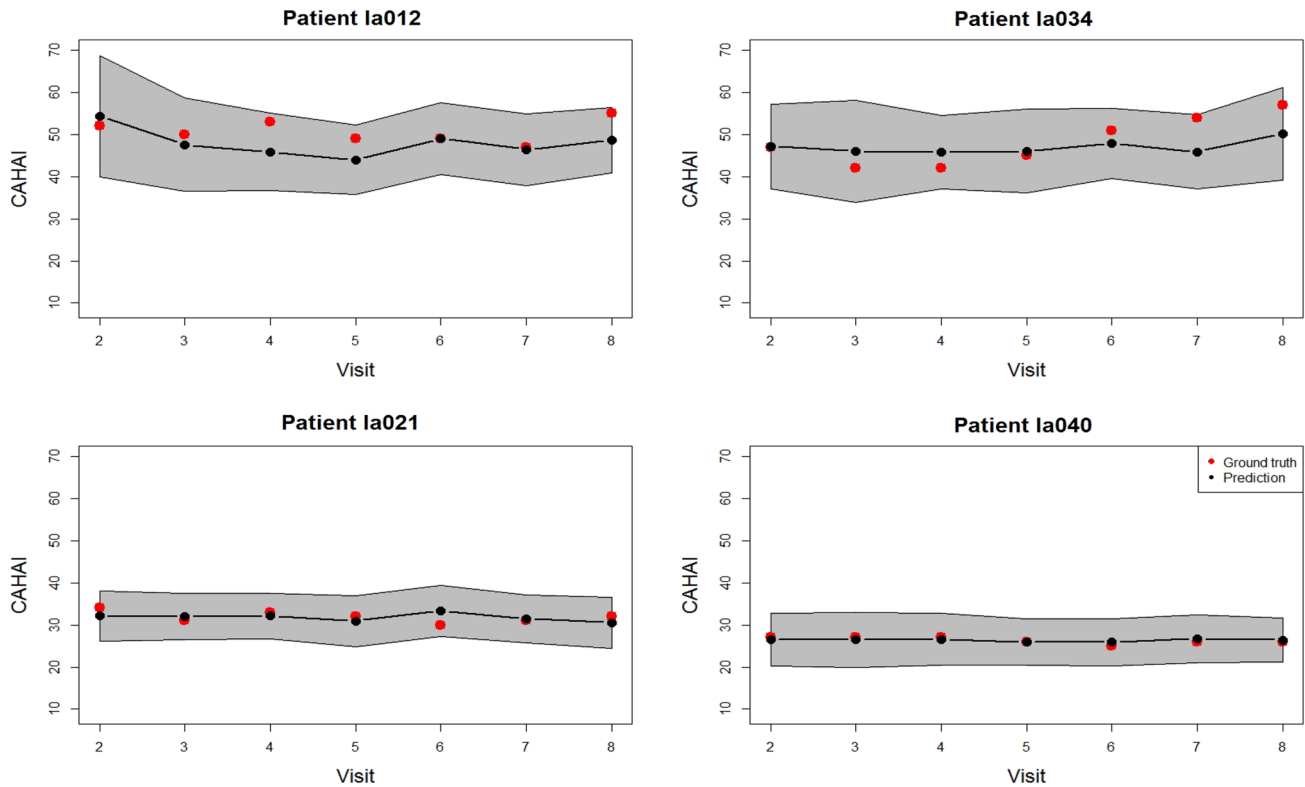

Fig. 11 Continues monitoring using LMGP for four patients (top: two chronic patients; bottom: two acute patients); dark points are the trial-wise/week-wise (i.e., each trial including data collected from 3 days

per week) prediction and red points are the corresponding ground truth CAHAI scores

based on leave-one-patient-out cross validation, and the mean RMSE values are reported in Fig. 10, from which can see LMGP can further reduce the errors when compared with the fixed-effects linear model, with mean RMSE 5.75 for acute patients and 3.12 for chronic patients, respectively.

Based on LMGP, we also perform "continuous monitoring"—with week-wise predicted CAHAI score—on four patients (two for each patient group) from week 2 to week 8, and the results are reported (with mean and 95% confidence interval) in Fig. 11, which is extremely helpful when uncertainty measurement is required.

4.2.4 On the fixed-effects part of LMGP

LMGP includes two key parts, i.e., the linear fixed-effects and the non-linear random-effects part, and it is important to choose the key features for modelling. Since the fixed-effects part measures the main (linear) relationship between the input features and the predicted CAHAI, we study the corresponding feature subsets. For random-effects part, we use the full LASSO features (as shown in Table 3).

To select the most important feature subset for the fixed-effects part modelling, we rank the features (from Table 3) based on two criteria: LASSO coefficients, and correlation coefficients (between features and CAHAI, as described in

Sect. 4.1). Table 4 demonstrates ranked features, and here only the top 50% features (i.e., top three features for acute patients and top five features for chronic patients) are used to model the fixed-effects part, and the settings as well as the results are reported in Table 5.

It is interesting to observe the performance may vary when different feature subsets are applied. Specifically, with the top feature subsets, modelling the LMGP's fixed-effects part can further reduce the errors to some extent for acute patients, in contrast to chronic patients with increased errors. The top 5 features selected via the LASSO criterion yields the worst performance for chronic patients, and one possible explanation could be the lack of feature *ini*—the initial health condition—a major attribute for chronic patient modelling (see Fig. 9).

4.2.5 Model comparison

Based on our proposed (41-dimensional) stroke-rehab-driven features, we compare LMGP with a number of classical predictive models, such as neural network (NN), support vector regression (SVR) and random forest regression(RF) for acute/chronic patient groups. It is worth noting that we cannot use the popular deep learning structures such as convolutional neural network(CNN) or recurrent neural

Table 4 Feature importance ranking (based on two criteria) for acute/chronic patients

-	Acute patients	Chronic patients
LASSO coefficients (absolute value)	PNP_3^2, PNP_6^1, SAD_2^{np}, $SAD_{1.2}^p$ SAD_6^{np}, *ini*	$PNP_{1.4}^1$, SAD_4^p, SAD_2^{np}, $PNP_{1.3}^2$ PNP_4^1, $PNP_{1.1}^2$, *ini*, PNP_6^1 $SAD_{1.4}^{np}$, SAD_6^{np}
Correlation coefficients (absolute value)	PNP_3^2, *ini*, $SAD_{1.2}^p$, PNP_6^1 SAD_2^{np}, SAD_6^{np}	*ini*, $PNP_{1.4}^1$, $PNP_{1.3}^2$, $PNP_{1.1}^2$ $SAD_{1.4}^{np}$, SAD_2^{np}, PNP_4^1, SAD_6^{np} PNP_6^1, SAD_4^p

Table 5 LMGP's fixed-effects part modelling results (RMSE) based on different feature subsets

Acute patients	Fixed-effects features	Random-effects features	RMSE
	Full six features in Table 3	full 6 features in Table 3	5.75
	Top three features (Corr criterion in Table 4): PNP_3^2, *ini*, $SAD_{1.2}^p$	Full six features in Table 3	5.37
	Top three features (LASSO criterion in Table 4): PNP_3^2, PNP_6^1, SAD_2^{np}	Full six features in Table 3	5.51

Chronic patients	Fixed-effects features	Random-effects features	RMSE
	Full ten features in Table 3	Full 10 features in Table 3	3.12
	Top 5 features (Corr criterion in Table 4): *ini*, $PNP_{1.4}^1$, $PNP_{1.3}^2$ $PNP_{1.1}^2$, $SAD_{1.4}^{np}$	Full 10 features in Table 3	3.20
	Top 5 features (LASSO criterion in Table 4): $PNP_{1.4}^1$, SAD_4^p, SAD_2^{np} $PNP_{1.3}^2$, PNP_4^1	Full 10 features in Table 3	5.12

Table 6 Predictive model comparison based on the proposed stroke-rehab-driven features (in LOSO-CV setting)

Predictive models	RMSE (acute)	RMSE (chronic)
Neural network	10.50	4.93
Support vector regression (linear)	7.47	3.25
Support vector regression (rbf)	9.67	4.92
Random forest regression	8.19	3.93
Linear fixed-effects model	7.24	3.29
LMGP	5.75	3.12

Table 7 Method comparison (in LOSO-CV setting)

Methods	RMSE (acute)	RMSE (chronic)
Tang et al. (2020)	15.98	12.76
Halloran et al. (2019)	10.12	12.14
Ours	5.75	3.12

network(RNN) on the time-series signal, due to the lack of frame-wise or sample-wise annotation. Yet with the stroke-rehab-driven features and trial-wise annotation, simple neural networks such as multi-layer perceptron (MLP) can be applied, and here we use a 3-layer MLP.

LOSO-CV is applied with the mean RMSE values reported in Table 6, from which we observe linear models (linear SVR and linear fixed-effects model) yield better results than non-linear methods (NN, SVR with rbf, and RF). One of the explanation is the over-fitting effect, where the trained non-linear models do not generalise well to the unseen patients/environments in this longitudinal study setting. RF is normally known as a classifier with high generalisation capability, yet it may suffer from the low-dimensionality of the selected features (6 features for acute patients and 10 features for chronic patients). Given the simplicity of the linear models and the designed low-dimensional features, linear models tend to suffer less from the over-fitting effect, with reasonable results in these challenging environments. Compared with linear models, our LMGP can further model the longitudinal mixed-effects (i.e., with linear fixed-effect part and non-linear random-effects part), making the system adaptive to different subjects/time-slots, with the lowest errors.

We also compare our approach with other automated CAHAI score regression methods (Tang et al. 2020; Halloran et al. 2019) in the existing literature. Different from our approach, Tang et al. (2020) and Halloran et al. (2019) are pure data-driven approaches. To address the lack of annotation problem, Tang et al. use GMM clustering (on the sliding windows) Tang et al. (2020) to learn latent features that can be aggregated for trial-wise representation, while Halloran et al. (2019) employ pseudo labelling strategy for trial-wise

representation. However, both data-driven features cannot suppress the substantial noises in the original accelerator signal, and such noises (e.g., irrelevant daily activities) significantly affect the performance of both approaches. In contrast, by taking advantage of the domain knowledge, our proposed stroke-rehab-driven representation is compact yet informative, and from Tables 6 and 7 we can see it tends to have lower errors than (Tang et al. 2020; Halloran et al. 2019 irrespective of the predictive models for both patient groups.

5 Conclusions

In this work, we develop an automated stroke rehabilitation assessment system using wearable sensing and machine learning techniques. We collect accelerometer data using wrist-worn sensors, based on which we build models for CAHAI score prediction, which can provide objective and continuous rehabilitation assessment. To map the long time-series (i.e., 3-day accelerometer data) to the CAHAI score, we propose a pipeline which can perform from data cleaning, feature design, to predictive model development. Specifically, we propose two compact features which can well capture the rehabilitation characteristics while suppressing the irrelevant daily activities, which is crucial on analysing the data collected in free-living environments. We further use LMGP, which can make the model adaptive to different subjects and different time slots (across different weeks). Comprehensive experiments are conducted on both acute/chronic patients, and very promising results are achieved, especially on the chronic patient group. We also study different feature subsets on modelling the fixed-effects part in LMGP, and experiments suggest the errors can be further reduced for the challenging acute patient population.

Due to irrelevant daily activities and strong heterogeneity among subjects, it is very challenging for researchers in mathematics, computing sciences and other areas to deal with free-living data. It is also crucial to develop models which have good mathematical properties and have physical explanation particularly in medical research. Hopefully, the ideas of the new features and the models discussed in this paper can provide some hints on addressing similar problems in health research.

Appendix

List of abbreviations/notations

- VM: Signal vector magnitude
- DWT: Discrete wavelet transform
- DWPT: Discrete wavelet packet transform
- LMGP: Longitudinal mixed-effects Gaussian process prior

- SAD: Normalised sum of absolute value of the wavelet coefficients at different decomposition scales
- PNP: Wavelet features that combine both paralysed side and non-paralysed side

The CAHAI score form

See Fig. 12.

Discrete wavelet transform and discrete wavelet packet transform

The DWT procedure includes two parts: decomposition and reconstruction. Decomposition part will be the main focus in this project. We now consider more details of the DWT using matrix algebra:

$$\mathbf{W} = \mathcal{W}\mathbf{X}, \tag{7}$$

where W is the output of matrix of DWT coefficients in different scales. W is the orthonormal matrix containing different orthonormal wavelet bases [more details can be checked in Daubechies (2006) and Walden et al. (2000)] and it satisfies $\mathcal{W}^T\mathcal{W} = \mathbf{I}_N$. $\mathbf{X}$ is the raw signal. The signal $\mathbf{X}$ with length $N = 2^J$, the $N \times N$ orthonormal matrix W can be separated into J + 1 submatrices, each of which can produce a partitioning of the vector $\mathbf{W}$ of DWT coefficients in each scale j, j = 1,2,..., J. To be more specific, Eq. (7) can be rewritten as follows:

$$\mathcal{W}\mathbf{X} = \begin{bmatrix} \mathcal{W}_1 \\ \mathcal{W}_2 \\ \vdots \\ \mathcal{W}_J \\ \mathcal{V}_J \end{bmatrix} \mathbf{X} = \begin{bmatrix} \mathcal{W}_1\mathbf{X} \\ \mathcal{W}_2\mathbf{X} \\ \vdots \\ \mathcal{W}_J\mathbf{X} \\ \mathcal{V}_J\mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_J \\ \mathbf{V}_J \end{bmatrix} = \mathbf{W}, \tag{8}$$

where $\mathbf{W}_j$ is a column vector of length $N/2^j$ representing the differences in adjacent weighted averages from scale 1 to scale J, $\mathbf{V}_J$ is the last column contained in $\mathbf{W}$ which has the same length with $\mathbf{W}_J$. $\mathbf{W}_j$ is defined as detailed coefficients at scale j. $\mathbf{V}_J$ contains the approximated coefficients at the J-th level. $\mathcal{W}_j$ has dimension $N/2^j \times N$, where j = 1,2,...,J and $\mathbf{V}_J$ has the same dimension with $\mathbf{W}_J$. Note that the rows of design orthonormal matrix $\mathcal{W}$ depend on the decomposition level j-th. In other words, the value of J depends on the DWT decomposition scale of the raw signal. The maximum decomposition level j equals J since our signal $\mathbf{X}$ has length $N = 2^J$.

We now further consider wavelet packet transform DWPT. The DWPT is the expansion of the discrete wavelet transformation. In DWT, each scale is calculated by passing only the previous wavelet approximated coefficients through discrete-time low and high pass quadrature mirror filters. However, in the DWPT, both the detailed and approximation coefficients are decomposed to create the full binary tree. More details can be found in Walden et al. (2000).

Commonly used wavelet features

In the discrete wavelet transform (DWT), $\mathbf{W}_j$ represents DWT coefficients in the j-th decomposition scale. DWT can be written as $\mathbf{W} = \mathcal{W}\mathbf{X}$, where $\mathbf{W}$ is a column vector with length 2^j and $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_J, \mathbf{V}_J]^T$, $\mathcal{W}$ is the orthonormal matrix which satisfies $\mathcal{W}^T\mathcal{W} = \mathbf{I}_n$ and contains different filters. Due to the orthonormality of DWT, which means that $\mathbf{X} = \mathcal{W}^T\mathbf{W}$ and $\|\mathbf{X}\|^2 = \|\mathbf{W}\|^2$, $\left\|\mathbf{W}_j\right\|^2$ shows energy in the DWT coefficients with decomposition level j. Now the energy preserving condition can be written as:

$$\|\mathbf{X}\|^2 = \|\mathbf{W}\|^2 = \sum_{j=1}^{J} \left\|\mathbf{W}_j\right\|^2 + \|\mathbf{V}_J\|^2, \tag{9}$$

where $\mathbf{X}$ is our VM data (the signal vector magnitude of accelerometer data; see Sect. 3.2) with length N, $j = 1, 2, ..., J$ is the discrete wavelet transform decomposition level. $\mathbf{W}_j$ denotes the detailed coefficient in scale j, and is a vector of length $N/2^j$ representing the differences in adjacent weighted averages from scale 1 to scale J. $\mathbf{V}_J$ denotes the approximated coefficients in the Jth level and has the same length as $\mathbf{W}_J$. Based on the decomposition, each $\left\|\mathbf{W}_j\right\|^2$ represents a special part of the energy in our VM data which relates to the certain frequency domain (Preece et al. 2009; Walden et al. 2000). Then the sample variance from Walden et al. (2000) can be decomposed as:

$$\widehat{\sigma}_{\mathbf{X}}^2 = \frac{1}{N}\|\mathbf{W}\|^2 - \overline{X} = \sum_{j=1}^{J} \frac{\left\|\mathbf{W}_j\right\|^2}{N}. \tag{10}$$

The term $\frac{\|\mathbf{w}_j\|^2}{N}$ represents the sample variance (corresponding to j at different scales of DWT decomposition) in our VM data $\mathbf{X}$.

There are many wavelet features (e.g., Preece et al. 2009) for the classification of dynamic activities from accelerometer data using DWT. On this basis, we extract the features from the energy preserving condition and sample variance mentioned previously.

We aim to look for the features which imply the recovery level among the stroke patients (see Sect. 3.3). Now, we define the features in the j-th level discrete wavelet transform and discrete wavelet packet transform:

$$\mathbf{SSD}_j = \frac{\left\|\mathbf{W}_j\right\|^2}{N/2^j} = 2^j \frac{\left\|\mathbf{W}_j\right\|^2}{N}.$$

Fig. 12 The CAHAI score form Barreca et al. (2006)

$$SAD_j = \frac{\|\mathbf{W}_j\|_1}{N/2^j} = 2^j \frac{\|\mathbf{W}_j\|_1}{N}.$$

After we check the correlation between the important wavelet feature PNP (Sect. 3.3) and CAHAI score, the branch of features PNP using SAD based perform better than those using SSD based in Table 8. Hence we consider the commonly used feature SAD_j in this paper.

In our analysis, we assume the discrete wavelet decomposition level $J = 7$ which is the same level as in Sekine et al. (1998) and contains enough low-frequency component as the stroke patients' movement. The frequency domain with seven scales is shown in Table 9

So far, we have decomposed the VM data $\mathbf{X}$ to get $\mathbf{W}_1$, $\mathbf{W}_2$, ... , $\mathbf{W}_7$ using DWT. Since the frequency domain at scale 1 is so wide (0.50–1 hz), it is better to divide it into smaller one, then using DWPT in Appendix 3, we can further decompose $\mathbf{W}_1$ into $\mathbf{W}_{3.4}$, $\mathbf{W}_{3.5}$, $\mathbf{W}_{3.6}$ and $\mathbf{W}_{3.7}$ which are the results of the 3rd stage of DWPT, each coefficient vector with length $N/2^3$ has the same dimension as the coefficients in the third level of DWT decomposition, that is

$$\|\mathbf{X}\|^2 = \|\mathbf{W}\|^2 = \|\mathbf{W}_{3.4}\|^2 + \|\mathbf{W}_{3.5}\|^2 + \|\mathbf{W}_{3.6}\|^2 + \|\mathbf{W}_{3.7}\|^2 + \sum_{j=2}^{J} \|\mathbf{W}_j\|^2 + \|\mathbf{V}_J\|^2.$$

For the detailed coefficients $\mathbf{W}_j$ at decomposition level j, $\|\mathbf{W}_j\|^2$ presents its energy and the raw data with length N. Hence the physical explanation of SSD_j is that it stands for the point energy at the decomposition level j. Moreover, from the Eq. (10), $\frac{\|\mathbf{W}_j\|^2}{N}$ represents the sample variance at the decomposition level j, SSD_j also has properties of both the energy preserving condition and the sample variance in wavelet analysis with constant 2^j.

Comparing with SSD_j (sum of square value of DWT coefficients at scale j (with normalisation)), we define other features call SAD_j, which is sum of absolute value of DWT coefficients at scale j (with normalisation):

Table 9 The frequency domain from scale 1 to scale 7 by using DWT

	Scale 7	Scale 6	Scale 5
Frequency	0.0078–0.0156 hz	0.0156–0.0312 hz	0.0312–0.0625 hz
	Scale 4	Scale 3	Scale 2
Frequency	0.0625–0.125 hz	0.125–0.25 hz	0.25–0.50 hz
	Scale 1		
Frequency	0.50–1 hz		

Table 8 The correlation between SAD and SSD based wavelet features and CAHAI score for acute and chronic patients

Scale (k)	PNP_k^1 (SSD)	Acute PNP_k^2 (SSD)	Patients PNP_k^1 (SAD)	PNP_k^2 (SAD)	PNP_k^1 (SSD)	Chronic PNP_k^2 (SSD)	Patients PNP_k^1 (SAD)	PNP_k^1 (SAD)
k = 1.1	0.60	− 0.65	0.68	− 0.70	0.45	− 0.45	0.56	− 0.56
k = 1.2	0.60	− 0.66	0.69	− 0.71	0.46	− 0.45	0.57	− 0.56
k = 1.3	0.63	− 0.69	0.70	− 0.72	0.49	− 0.48	0.58	− 0.57
k = 1.4	0.62	− 0.68	0.69	− 0.71	0.47	− 0.47	0.57	− 0.57
k = 2	0.65	− 0.69	0.69	− 0.71	0.45	− 0.45	0.56	− 0.55
k = 3	0.63	− 0.67	0.67	− 0.68	0.39	− 0.38	0.53	− 0.52
k = 4	0.59	− 0.63	0.60	− 0.63	0.31	− 0.30	0.48	− 0.47
k = 5	0.46	− 0.50	0.49	− 0.52	0.29	− 0.27	0.43	− 0.42
k = 6	0.32	− 0.38	0.35	− 0.38	0.20	− 0.16	0.35	− 0.34
k = 7	0.16	− 0.19	0.19	− 0.20	0.13	− 0.10	0.25	− 0.24

Table 10 The frequency domain from scale 1.1 to scale 7 by using DWPT and DWT

	Scale 1.1	Scale 1.2	Scale 1.3
Frequency	0.5–0.625 hz	0.625–0.75 hz	0.75–0.875 hz
	Scale 1.4	Scale 2	Scale 3
Frequency	0.875–1 hz	0.25–0.50 hz	0.125–0.25 hz
	Scale 4	Scale 5	Scale 6
Frequency	0.0625–0.125 hz	0.0312–0.0625 hz	0.0156–0.0312 hz
	Scale 7		
Frequency	0.0078–0.0156 hz		

Now we have coefficients at 10 decomposition scales by using DWT and DWPT: $\mathbf{W}_{3.4}$, $\mathbf{W}_{3.5}$, $\mathbf{W}_{3.6}$, $\mathbf{W}_{3.7}$, $\mathbf{W}_2$, $\mathbf{W}_3$, $\mathbf{W}_4$, $\mathbf{W}_5$, $\mathbf{W}_6$ and $\mathbf{W}_7$. Based on these detailed coefficients, we define the commonly used wavelet features again:

$$Scale\ 1.1: SAD_{1.1} = \frac{\|\mathbf{W}_{3.4}\|_1}{N/2^3} = 2^3\frac{\|\mathbf{W}_{3.4}\|_1}{N},$$

$$Scale\ 1.2: SAD_{1.2} = \frac{\|\mathbf{W}_{3.5}\|_1}{N/2^3} = 2^3\frac{\|\mathbf{W}_{3.5}\|_1}{N},$$

$$Scale\ 1.3: SAD_{1.3} = \frac{\|\mathbf{W}_{3.6}\|_1}{N/2^3} = 2^3\frac{\|\mathbf{W}_{3.6}\|_1}{N},$$

$$Scale\ 1.4: SAD_{1.4} = \frac{\|\mathbf{W}_{3.7}\|_1}{N/2^3} = 2^3\frac{\|\mathbf{W}_{3.7}\|_1}{N},$$

$$Scale\ j: SAD_{j} = \frac{\|\mathbf{W}_{j}\|_1}{N/2^j} = 2^j\frac{\|\mathbf{W}_{j}\|_1}{N}, \qquad j = 2, 3, 4, 5, 6, 7.$$

There are 10 features which provide reliable and valid information (corresponding to more frequency domains) from different frequency domains. The frequency domain of these features, among 10 scales, is listed in Table 10.

Performance of LMGP through three different kernels

Three kernels are used in LMGP, and they are linear kernel, powered exponential kernel and rational quadratic kernel. We use the selected features (from Table 3) as the fixed-effects features and random-effects features, and the results are reported in Table 11.

Table 11 Performance of LMGP based on three kernels

Selected kernels in LMGP	RMSE (acute)	RMSE (chronic)
Linear kernel	5.89	3.13
Powered exponential kernel	5.75	3.12
Rational quadratic kernel	7.58	3.24

References

Barreca, S., Stratford, P., Masters, L., Lambert, C., Griffiths, J., McBay, C.: Validation of three shortened versions of the chedoke arm and hand activity inventory. Physiother. Can. **58**, 1–9 (2006)

Axivity Ltd. AX3, 3-Axis Logging Accelerometer. https://axivity.com/product/ax3. [Online; accessed Jul 2020]

Ayachi, F.S., Nguyen, H.P., Lavigne-Pelletier, C., Goubault, E., Boissy, P., Duval, C.: Wavelet-based algorithm for auto-detection of daily living activities of older adults captured by multiple inertial measurement units (imus). Physiol. Meas. **37**(3), 442–461 (2016)

Bai, Y., Guan, Y., Ng, W.-F.: Fatigue assessment using ecg and actigraphy sensors. In: Proceedings of the 24rd International Symposium on Wearable Computers, ISWC, New York, NY, USA. Association for Computing Machinery (2020)

Barreca, S.R., Stratford, P.W., Lambert, C.L., Masters, L.M., Streiner, D.L.: Test-retest reliability, validity, and sensitivity of the Chedoke arm and hand activity inventory: a new measure of upper-limb function for survivors of stroke. Arch. Phys. Med. Rehabil. **86**(8), 1616–1622 (2005)

Bobin, M., Bimbard, F., Boukallel, M., Anastassova, M., Ammi, M.: Spectrum: Smart ecosystem for stroke patient's upper limbs monitoring. 13, 02 (2019)

Bouten, C., Koekkoek, K., Verduin, M., Kodde, R., Janssen, J.: A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. IEEE Trans. Bio-med. Eng. **44**, 136–47 (1997)

Daubechies, I.: Orthonormal bases of compactly supported wavelets. Commun. Pure. Appl. Math, pp 909–996, (2006)

Doherty, A., Jackson, D., Hammerla, N., Ploetz, T., Olivier, P., Granat, M.H., White, T., van Hees, V.T., Trenell, M.I., Owen, C.G., Preece, S.J., Gillions, R., Sheard, S., Peakman, T., Brage, S., Wareham, N.J.: Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. PLoS One **12**(2), 1–14 (2017)

Dolatabadi, E., Zhi, Y., Ye, B., Coahran, M., Lupinacci, G., Mihailidis, A., Wang, R., Taati, B.: The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy. pages 375–381, 05 (2017)

Donnan, G., Fisher, M., Macleod, M., Davis, S.: Stroke. Lancet **371**(2), 1612–1623 (2008)

Ferrari, P., Friedenreich, C., Matthews, C.: The role of measurement error in estimating levels of physical activity. Am. J. Epidemiol. **166**, 832–40 (2007)

Ganesh, A. C., Renganathan, B. S., Rajakumaran, C., Preejith, S. P., Shubham, K., Jayaraj, J., Mohanasankar, S.: Post-stroke rehabilitation monitoring using wireless surface electromyography: A case study. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp 1–6, (2018)

Gao, T., Long, Y., Guan, Y., Basu, A., Baggaley, J., Ploetz, J.: Towards reliable, automated general movement assessment for perinatal stroke screening in infants using wearable accelerometers. In: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 3(1), March (2019)

Guan, Y., Ploetz, T.: Ensembles of deep lstm learners for activity recognition using wearables. In: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 1(2), (2017)

Gurchiek, R., Choquette, R., Beynnon, B., Slauterbeck, J., Tourville, T., Toth, M., McGinnis, R.: Open-source remote gait analysis: a post-surgery patient monitoring application. Sci. Rep. 9, 17966 (2019)

Halloran, S., Tang, L., Guan, Y., Shi, J. Q., Eyre, J.: Remote monitoring of stroke patients' rehabilitation using wearable accelerometers. In: Proceedings of the 23rd International Symposium on Wearable Computers, ISWC 19, New York, NY, USA. Association for Computing Machinery, pp 72–77 (2019)

Hammerla, N.Y., Fisher, J.M., Andras, P., Rochester, L., Walker, R., Ploetz, T.: Pd disease state assessment in naturalistic environments using deep learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI. AAAI Press, pp 1742–1748 (2015)

Ibrahim, A.A., Küderle, A., Gaßner, H., Klucken, J., Eskofier, B.M., Kluge, F.: Inertial sensor-based gait parameters reflect patient-reported fatigue in multiple sclerosis. J. Neuroeng. Rehabil. 17(1), 165 (2020)

Jorgensen, H.S., Nakayama, H., Raaschou, H., Olsen, T.S.: Stroke: Neurologic and functional recovery the Copenhagen stroke study. Phys. Med. Rehabil. Clin. N. Am. 10(4), 887–906 (1999)

Jung, H., Park, J., Jeong, J., Ryu, T., Kim, Y., Lee, S. I.: A wearable monitoring system for at-home stroke rehabilitation exercises: A preliminary study. In: 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pp 13–16, (2018)

Karantonis, D.M., Narayanan, M.R., Mathie, M., Lovell, N.H., Celler, B.G.: Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans. Inf. Technol. Biomed. 10(1), 156–167 (2006)

Little, B., Alshabrawy, O., Stow, D., Ferrier, I., McNaney, R., Jackson, D., Ladha, K., Ladha, C., Ploetz, T., Bacardit, J., Olivier, P., Gallagher, P., O'Brien, J.: Deep learning-based automated speech detection as a marker of social functioning in late-life depression. Psychol. Med. 51, 1–10 (2020)

Ploetz, T., Hammerla, N.Y., Rozga, A., Reavis, A., Call, N., Abowd, G.D.: Automatic assessment of problem behavior in individuals with developmental disabilities. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp, New York, NY, USA, Association for Computing Machinery, pp 391–400 (2012)

Ploetz, T., Guan, Y.: Deep learning for human activity recognition in mobile computing. Computer 51(5), 50–59 (2018)

Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D.: A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Trans. Biomed. Eng. 56(3), 871–879 (2009)

Ratcliffe, A.M., Zhai, B., Guan, Y., Jackson, D., Sneyd, J.R.: Patient-centred measurement of recovery from day-case surgery using wrist worn accelerometers: a pilot and feasibility study. Anaesthesia 76, 785–797 (2020)

Rehman, R.Z.U., Din, S., Guan, Y., Yarnall, A., Shi, J., Rochester, L.: Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: a comprehensive machine learning approach. Sci. Rep. 9, 12 (2019)

Sekine, M., Tamura, T., Ogawa, M., Togawa, T., Fukui, Y.: Classification of acceleration waveform in a continuous walking record. In: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), vol 3, pp 1523–1526 (1998)

Shi, J., Choi, T.: Gaussian process regression analysis for functional data. Chapman and Hall/CRC, London (2011)

Shi, J.Q.Y. Cheng, J. Serradilla, G. Morgan, C. Lambden, G. Ford, C. Price, H. Rodgers, T. Cassidy, L. Rochester, Eyre, J.A.: Evaluating functional ability of upper limbs after stroke using video game data. In: Imamura, K., Usui, S., Shirao, T., Kasamatsu, T., Schwabe, L., Zhong, N. (eds) International Conference on Brain and Health Informatics, volume 8211 of Lecture Notes in Artificial Intelligence. Springer, pp 181–192 (2013)

Supratak, A., Dong, H., Wu, C., Guo, Y.: Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. 25, 1998–2008 (2017)

Shi, J.Q., Wang, B., Will, E.J., West, R.M.: Mixed-effects Gaussian process functional regression models with application to dose response curve prediction. Stat. Med. 31(26), 3165–3177 (2012)

Tang, L., Halloran, S., Shi, J. Q., Guan, Y., Cao, C., Eyre, J.: Evaluating upper limb function after stroke using the free-living accelerometer data. Stat. Methods Med. Res. (2020)

Walden, A.T., Percival, D.B.: Wavelet methods for time series analysis. Cambridge series in statistical and probabilistic mathematics, 1st edn. Cambridge University Press, Cambridge (2000)

Wintermark, M., Sesay, M., Barbier, E., Borbély, K., Dillon, W.P., Eastwood, J.D., Glenn, T.C., Grandin, C.B., Pedraza, S., Soustiel, J.F., Nariai, T., Zaharchuk, G., Caillé, J.M., Dousset, V., Yonas, H.: Comparative overview of brain perfusion imaging techniques. J. Neuroradiol. 32(5), 294–314 (2005)

Zhai, B., Perez-Pozuelo, I., Clifton, E. A. D., Palotti, J., Guan, Y.: Making sense of sleep: multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. In: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 4(2), (2020)

Dr. Xi Chen received the Ph.D. degree from the Department of Mathematics, Statistics & Physics, Newcastle University 2021. He is currently a staff in Hainan Rural Credit Union, China. His research interests include functional data analysis and wavelet analysis for complex data. Besides, he is also interested in biomedical statistics and its various applications, such as stroke rehabilitation under the wavelet-domain in a functional data analysis framework.

Dr. Yu Guan is an Associate Professor in the Department of Computer Science, the University of Warwick. His research interestes are machine learning, wearable computing, activity recognition, computer vision, etc. He is an Associate Editor of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT).

Dr. Jian Qing Shi is a Professor of Statistics in the Department of Statistics and Data science, and also the director of the Center for Biostatistics, Southern University of Science and Technology (SUSTech), China. Before he joined in Sustech in 2020, he was working as a Reader in the School of Mathematics, Statistics & Physics, assistant director of the Cloud Computing for Big Data CDT in Newcastle University and Turing Fellow in the Alan Turing Institute UK. His research interests include functional data analysis, Bayesian nonparametric analysis for big data, missing data, meta-analysis and applications in medicine and system control. He is the principal investigator of several big research projects in both UK and China. He has been visiting research fellow in Newton Institute in Cambridge UK and SAMSI in USA. He and his team have achieved many important and original research works, including two recent international best paper awards.

Dr. Xiuli Du received the Ph.D. degree from the Department of Measuring and Testing Technologies and Instruments, College of Civil Engineering, Southeast University, in 2010. She is currently an Associate Professor in School of Mathematical Sciences, Nanjing Normal University. Her research interests are Time Series Analysis, Non(Semi)-parametric Statistics, Longitudinal Data Analysis, Image Processing and Functional Data Analysis.

Dr. Janet Eyre is a Professor of Paediatric Neuroscience at the institute of Neuroscience in the Newcastle University UK. Her research interests include Corticospinal system, plasticity, rehabilitation and video games. Professor Eyre has received several awards for her outstanding work, such as The NHS Innovations North Bright Ideas in Health Award, in 2009, the CELS Business for Life Awards— Partnership with the NHS, in 2010, The UnLtd and Higher Education Funding Council for England Entrepreneur Award. In 2011, she received tje Medical Future's Best Innovation to Improve Patient Care – Cardiovascular Innovation Award and The Medical Futures National Health Service Innovation of the Year Award, in 2012.