Quantifying Movement Behavior of Chronic Low Back Pain Patients in Virtual Reality

TOMMI GRÖHN, Department of Computer Science, Aalto University, Finland
SAMMELI LIIKKANEN, DRDP, Institute of Biomedicine, University of Turku, Finland, and Orion Pharma, Finland
TEPPO HUTTUNEN and MIKA MÄKINEN, EstiMates Oy, Finland
PASI LILJEBERG, Department of Information Technology, University of Turku, Finland
PEKKA MARTTINEN, Department of Computer Science, Aalto University, Finland

Chronic low back pain (CLBP) is a globally common musculoskeletal problem. Measuring the sensation of pain and the effect of a treatment has always been a challenge for healthcare. Here, we study how the movement data, collected while using a virtual reality (VR) program, could be used as an objective measurement in patients with CLBP. A specific data collection method based on VR was developed and used with CLBP patients and healthy volunteers. We demonstrate that the movement data in VR can be used to classify individuals in these two groups with a high accuracy by using logistic regression. The most discriminative features are the duration of the movements and the total variation of movement velocity. Furthermore, we show that hidden Markov models can divide movement data into meaningful segments, which creates possibilities for defining even more detailed features, with potential to improve accuracy, when larger datasets become available in the future.

CCS Concepts: • **Applied computing** → **Health informatics**; • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Chronic low back pain, digital biomarker, digital therapeutics, gesture recognition, hidden Markov models, movement data, time series segmentation, virtual reality

ACM Reference format:
Tommi Gröhn, Sammeli Liikkanen, Teppo Huttunen, Mika Mäkinen, Pasi Liljeberg, and Pekka Marttinen. 2023. Quantifying Movement Behavior of Chronic Low Back Pain Patients in Virtual Reality. *ACM Trans. Comput. Healthcare* 4, 2, Article 11 (April 2023), 24 pages.
https://doi.org/10.1145/3582487

1 INTRODUCTION

Measuring pain has always been a challenge. **Numerical rating scale (NRS)** and **Visual Analog Scale (VAS)** scoring systems and **electronic Patient Reported Outcome (ePRO)** tools, which are widely used, are highly subjective methods to measure pain. They also take the patient's mind repeatedly back to the disease, which

T. Gröhn and S. Liikkanen contributed equally to this research.

Authors' addresses: T. Gröhn and P. Marttinen, Department of Computer Science, Aalto University, Finland; email: tommi.i.grohn@aalto.fi, pekka.marttinen@aalto.fi. S. Liikkanen, DRDP, Institute of Biomedicine, University of Turku, Finland, and Orion Pharma, Finland; email: sammeli.liikkanen@orionpharma.com. T. Huttunen and M. Mäkinen, EstiMates Oy, Finland; email: teppo.huttunen@estimates.fi, mika.makinen@estimates.fi. P. Liljeberg, Department of Information Technology, University of Turku, Finland; email: pasi.liljeberg@utu.fi.

2637-8051/2023/04-ART11 $15.00
https://doi.org/10.1145/3582487

is the opposite of what the patient wants to achieve [16]. Imaging techniques like MEG [44], fMRI [25], and PET [45] are used at clinics, but they are expensive and not suitable for remote and frequent monitoring, [6, 36]. Various neurophysiological tests such as **quantitative sensory tests (QSTs)**, autonomic tests, **microneurography (MCNG)**, **laser-evoked potentials (LEPs),** and **contact heat-evoked potential stimulators (CHEPSs)** can also be used in the assessment of pain [8, 17, 19, 28, 34, 47]. Most of these, however, are invasive and require specific test systems and trained users to perform the assessment. They are also not suitable for remote assessment. Measuring pain objectively is much needed, and more objective digital monitoring methods are under development [20, 23, 37, 41].

The use of digital tools in collecting data from implantable and external sensors provides opportunities for digital biomarkers in monitoring disease progress and treatment response. At best, sensor-based digital biomarkers are developed by using both physicians and patients when annotating the sensor-derived data (e.g., the severity of motor symptoms). Different sensors, wearables, and mobile devices allow the follow-up of vital and relevant biosignals such as heart rate variability, movement data, activity, sleep data, and skin conductance, and then clinically valid digital signals can be analyzed and detected [2, 35, 48].

Measuring the pain itself is not the only important objective. Another option is to measure the consequences of chronic pain over time instead of the actual sensation. The patient's ability to manage in their everyday life is critical. For instance, cognitive impairment is something that occurs in neurological illnesses, such as chronic pain, **Parkinson's disease (PD),** and **essential tremor (ET)**. It is shown, however, that even an adult brain is able to improve from a state of limited functionality. Thus, the ability to measure cognitive performance is an important part of monitoring the performance of any pain, PD, or ET intervention. The traditional methods used in these diseases are cognitive tests, such as the **Montreal Cognitive Assessment (MoCA)**, which involve answering a series of questions and/or performing simple cognitive tasks; however, these are subjective, and are even suggested to be poor with regard to diagnosis accuracy [5, 30–32, 46]. In the case of kinesiophobia, which chronic pain sometimes can cause, the **Tampa Scale for Kinesiophobia (TSK)** is a common and globally accepted scale for assessment [11, 14, 33].

The development of different digital biomarkers and assessments in chronic pain is active. The relationship between **chronic low back pain (CLBP)** and general physical activity and gait has been studied [24, 38]. The ability to classify patients in sub-populations based on the activity data seems possible. Assessment of gait and balance has been exploited with chronic stroke patients, who are also suffering from cognitive impairment issues [21]. In PD and ET, the automatic analysis of more specific motor symptoms, alongside physical activity and gait analysis, has been studied. Both laboratory and consumer-grade accelerometers have enabled the assessment of tremor and the quantification of the treatment effects [4, 10, 15, 18, 27, 40]. Various digital cognitive tests and games, specifically designed and built for this purpose, can be used both as digital biomarkers and as rehabilitating therapy for patients suffering from these diseases [1, 7, 12, 22, 39].

These novel digital biomarkers could be used as clinically validated and complementary endpoints in

- diagnosing diseases at early stages,
- monitoring the progression of diseases, and
- measuring the effect of therapeutic interventions, both with traditional drug therapeutics and **digital therapeutics (DTx)**.

We developed a specifically designed and built virtual reality program (Painlab) to collect movement data from both chronic pain and healthy participants in a laboratory-grade setting; see Figure 1. This was done in order to research how the movement data collected by accelerometers, while executing a predefined pattern of physical tasks in a VR environment, varies between CLBP patients and healthy volunteers. The collected data from Painlab was used to classify CLBP patients and healthy volunteers. In the future, the aim is to exploit this classification approach in the analysis of movement data from real-world settings when using VR.

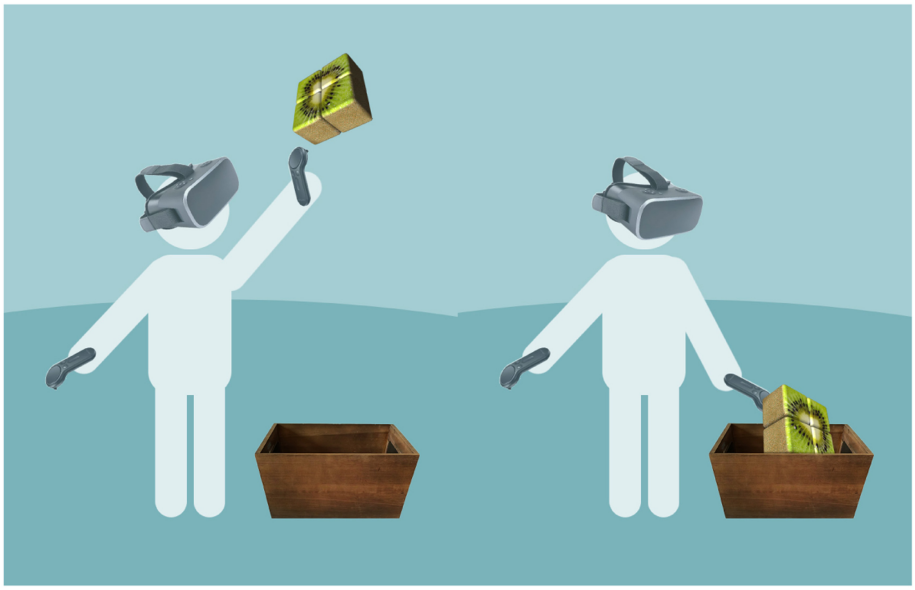

Fig. 1. An illustration of a subject doing an activity of up “picking an object" in the Painlab VR environment.

Here we report a methodology where we first identify the most useful features to differentiate the two cohorts, after which we use **hidden Markov models (HMMs)** to establish segments in the movement data. Finally, we compare different logistic regression models with the non-segmented and the segmented data.

2 DATA DESCRIPTION

Movement data was collected from CLBP patients ($N = 10$) and healthy volunteers ($N = 10$) by using a VR program that was specifically designed for our purposes. We chose patients with severe chronic pain in our study, because we wanted a dataset with as big differences in the movement data between the two populations as possible. Severe CLBP patients were also the potential target population for the DTx that was being developed alongside this study. The inclusion criteria for CLBP patients are:

- CLBP for at least 3 months (the most painful condition should be low back pain)
- Ambulatory
- Average pain intensity of ≥4 over the past week on a 0–10 NRS, either at rest or on back-extension movement
- Oswestry Disability Index of ≥26%
- Medium (34–41) or high TSK score (42–68)
- Can stoop without severe pain

Study participants were recruited by using social media advertisements or by contacting CLBP patients in previous similar trials. The study and the collection of the data were performed in the clinical research unit at Orion Corporation Orion Pharma, Espoo, Finland. After the participants arrived on site, they were screened and examined by the study site personnel. Next they were asked to try the VR solution for about 30 minutes, during which the study personnel observed the movements to ensure the safety of the participant. Once the safety of the participant was confirmed, they were ready for the data collection phase with the Painlab solution.

Basic demographic variables, the Keele STarT Back Screening Tool, Pain interference Short Form 3a, Pain interference Short Form 6b, Pain Score NRS, Tampa Scale for Kinesiophobia (TSK), **Oswestry Disability Index (ODI)**, and familiarity with digital devices assessments were collected from the participants in the beginning of the study visit. The TSK, **Game Experience Questionnaire (GEQ)**, and **Modified Game Experience Questionnaire (Modified GEQ)** assessments were collected from the participants 1 week after the study visit. Nine patients were female and one patient was male. Furthermore, four controls were female and six of them were male. The average age for the patients was $\mu = 55$ ($\sigma = 7.3$) and for the controls $\mu = 35$ ($\sigma = 6.3$). However, the age difference does not explain our results; see Appendix A. All the participants were White.

The Painlab solution was designed to enable a systematic collection of wearable data when using the VR environment in a laboratory-grade setting. The movements, designed together with physiotherapists and pain experts, were based on a VR treatment that reported significant efficacy in a randomized clinical trial with CLBP patients [7]. It was built using Unity v2019.3.7f1 and delivered on an off-the-shelf OCULUS Quest and Touch VR headset and accompanying Touch hand-held controllers. All the scenes in VR run at a refresh rate minimum of 70Hz. The data sampling rate of the movement data from the headset and the hand-held controllers was set at 30Hz.

All subjects were asked to perform the same sequence of specific movements by using the same movement pattern and instructions. There were *four patterns of movements* for both hands, and **each pattern consisted of three movements**; see Figure 3. The first four patterns are denoted with 1R, 2R, 3R, and 4R, where the letter R means the right hand. The same patterns were also performed with the left hand. Let's denote these patterns with 1L, 2L, 3L, and 4L to emphasize that the participants perform the same patterns of movements with both hands. Furthermore, we describe movements with notation 1R1, 1R2, 1R3, 2R1, . . ., where 1R2, as an example, describes the second movement of the first pattern performed by the right hand. The average time to perform all the patterns of movements was 182 seconds ($\sigma = 60.3$ seconds). The data collection of two control subjects did not succeed when they performed movements with their right hand.

All eight patterns of movements are visualized in Figures 3 and 4. Moreover, the x-axis represents movements from left to right, the y-axis vertical movements, and the z-axis movements from front to back. The first pattern of movements is plotted in Figure 2. It consists of three different movements:

- Grab an object in front of you and put it in a box before you (1R1, 1L1).
- Grab an object that is above on the right side of you and put it in a box before you (1R2, 1L2).
- Grab an object that is above on the left side of you and put it in a box before you (1R3, 1L3).

The remaining patterns of movements follow the similar procedure, where the subject grabs an object from a specific position and puts it in a box. Therefore, only the location of the objects is described in the second, third, and fourth patterns without a repetitive description of the box. Objects that are located low or high are clearly indicated. Otherwise, the object is vertically in the middle. The locations of the objects in the following movements are:

- Object on the right side of the subject (2R1, 2L1)
- Object below and behind the subject (2R2, 2L2)
- Object below on the right side of the subject (2R3, 2L3)
- Object behind the subject (3R1, 3L1)
- Object on the left side of the subject (3R2, 3L2)
- Object on the right side of the subject (3R3, 3L3)
- Object on the left side of the subject (4R1, 4L1)
- Object below on the left side of the subject (4R2, 4L2)
- Object above on the left side of the subject (4R3, 4L3)

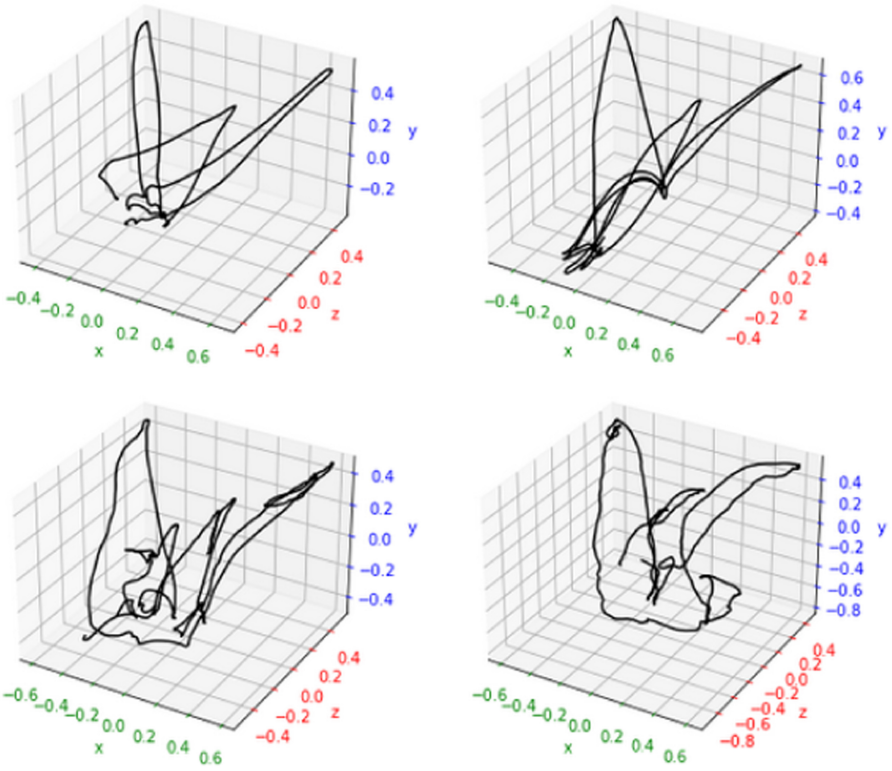

Fig. 2. The first pattern of movements 1L. The trajectory of the active left hand is plotted in a 3-dimensional space. The two plots above are trajectories from healthy individuals and the two plots below are from patients. Clearly, there is a pattern of three movements that occur in all these plots. However, the healthy individuals seem to have a more steady hand to create the pattern than the patients.

The movements used in our study were selected to mimic real-life tasks and movements that chronic pain patients might be avoiding in their everyday lives. Hence, the results are expected to generalize to other movements as long as these conditions hold, even if the exact tasks or the coordinate systems used to measure the movements were different.

Positional data of the active hand is processed further. First, velocity is calculated out of the positional data. Let's describe the 3-dimensional velocity of a subject i with

$$\mathbf{X}_i = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1T_i} \\ x_{21} & x_{22} & \dots & x_{2T_i} \\ x_{31} & x_{32} & \dots & x_{3T_i} \end{bmatrix},$$

where x_{jk} describes the velocity of the j-axis at the time point k. Furthermore, T_i describes the length of a pattern of movement, which is, for example, approximately 450, i.e., 30Hz · 15s, for the first pattern 1R in Figure 3.

Furthermore, by 3-dimensional *normalized velocity* is meant the matrix

$$\mathbf{X}'_i = \begin{bmatrix} x_{11}/c_1 & x_{12}/c_2 & \dots & x_{1T_i}/c_{T_i} \\ x_{21}/c_1 & x_{22}/c_2 & \dots & x_{2T_i}/c_{T_i} \\ x_{31}/c_1 & x_{32}/c_2 & \dots & x_{3T_i}/c_{T_i} \end{bmatrix},$$

where $c_t = \sqrt{(x_{1t})^2 + (x_{2t})^2 + (x_{3t})^2}$.

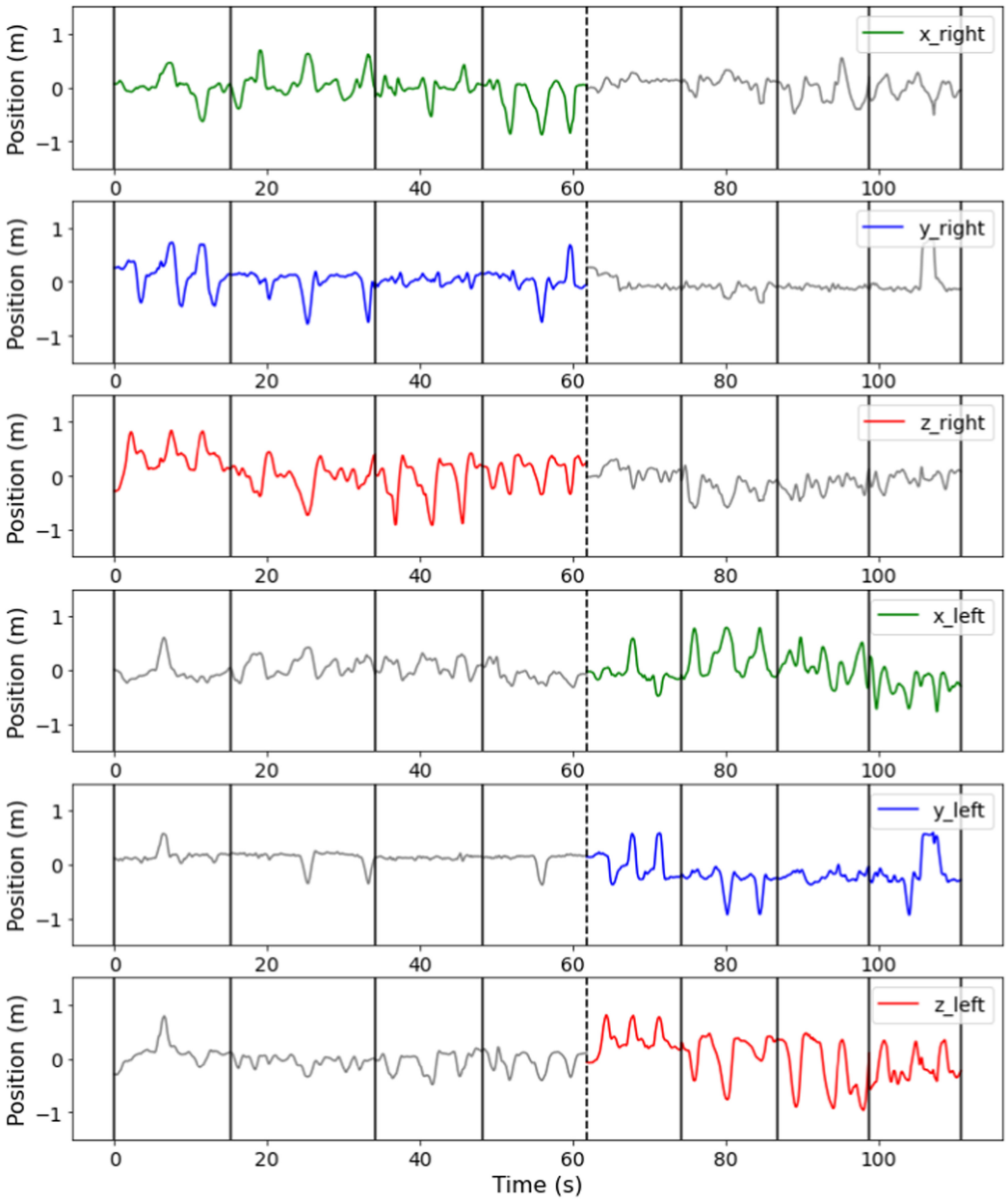

Fig. 3. An example of the whole time series data produced by a *healthy individual*. First, the subject performs four patterns of movements with their right hand. Then the subject performs the same patterns with their left hand. The data of the active hand is marked with the colors green, blue, and red. Moreover, the subject performed the first four patterns slower than the last four, as the subject learned at least a bit the game logic by playing. However, there is no clear pattern in the data of the passive hand.

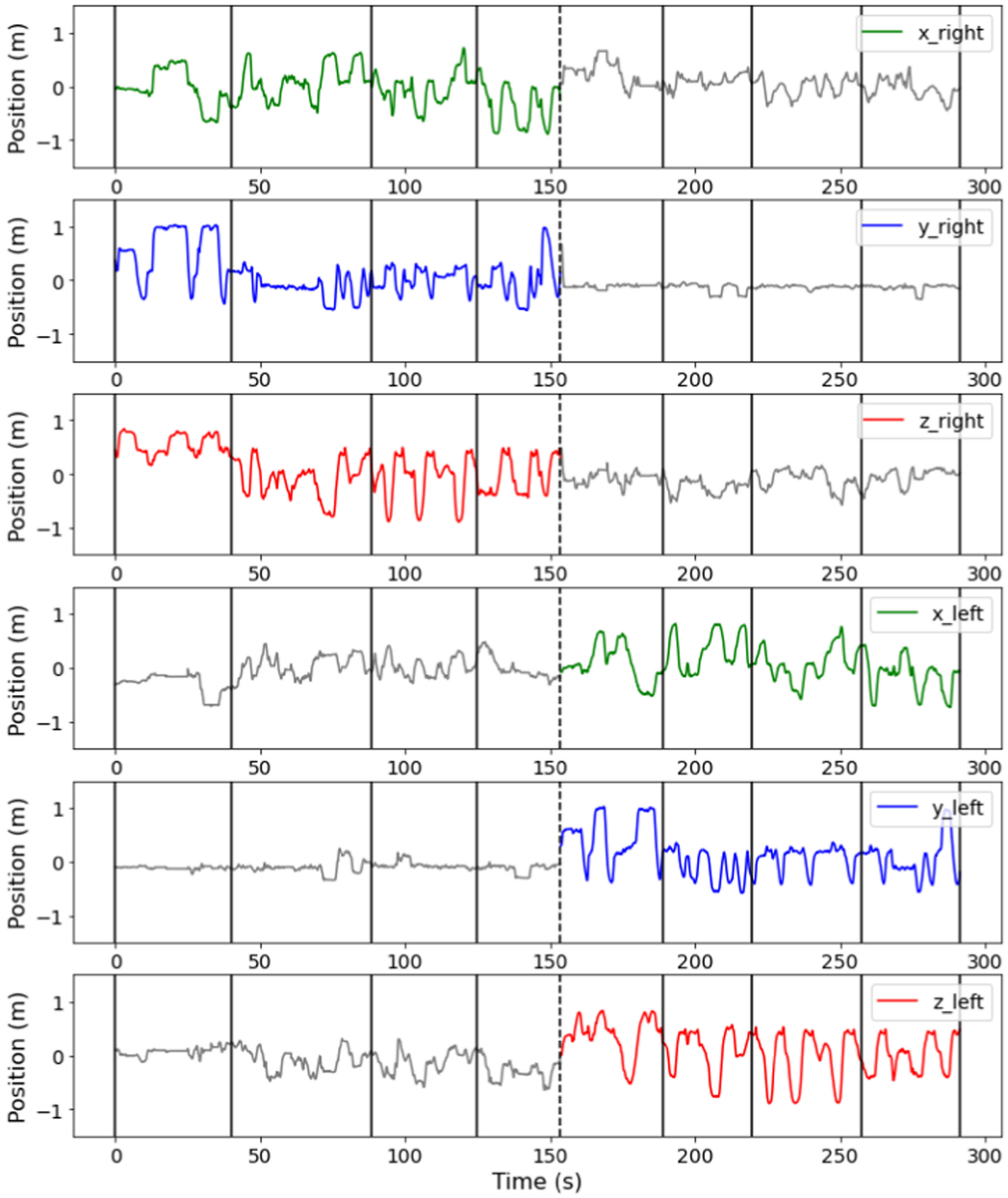

Fig. 4. An example of the whole time series data produced by a *chronic low back pain patient*. Similarly as before, the patient performs four patterns of movements using their right hand only and then the same patterns with their left hand. As the patterns are predefined, the data of the active hand has similar shapes compared to Figure 3. However, the patient performs the movements more slowly than the healthy individual. Note the different scale of the x-axis in Figures 3 and 4.

3 METHODS

Our research is a retrospective case control study. The movements in our data are similar to the ones used in a VR treatment with a good efficacy in a previous randomized clinical trial with CLBP patients [7], and therefore good candidates for CLBP treatments. Here, we introduce computational methods to quantify the progress of pain patients. The methods are divided into three stages:

(1) *Feature selection.* We select the most prominent features that differentiate the patients and the control group out of the movement data.
(2) *Time series segmentation using HMMs.* The patterns of movements are segmented with HMMs for further data analysis. An example of the segmentation can be seen in Figure 9, where a pattern of movements is segmented into 11 segments. Practically, these segments are pieces of movements, e.g., moving the hand to left.
(3) *Classifying patients and controls with logistic regression.* We train two logistic regression models. The first one uses non-segmented and the second one uses segmented data. The performances of these models are compared with each other.

The data is used differently in stages (1) and (2) compared to stage (3). We extract prominent features out of the dataset in stage (1) and learn the structure of the dataset in stage (2). In these stages, we only use data from the left hand as the data from right and left hands are not directly comparable. Indeed, participants perform the movements of the right hand first and therefore more slowly than with the left hand. Furthermore, we have enough different patterns (1L, 2L, 3L, 4L) to obtain the results we aim for in stages (1) and (2). However, in stage (3) we classify the participants, and to make the dataset as large as possible, we normalize the time series from both hands on the same scale and combine them, before dividing the resulting dataset into training and test sets for the classification stage. Additionally, we use HMMs in a more traditional manner in Appendix B than in stages (2) and (3).

3.1 Feature Selection

To estimate the usability of different features, distributions between the healthy individuals and the patients were compared with each other. The comparison was visualized with box-and-whisker plots. This section describes which four features we chose to investigate and why.

As observed in Figures 3 and 4, the patients seem to perform movements slower than the healthy individuals. Therefore, the used *time* to perform each pattern of movements is calculated first for all the subjects.

Intuitively, healthy individuals do not avoid moving their hand with high speed as much as the patients. Let $\mathbf{X}_i : \Omega \to \mathbb{R}^3$ be a random variable to measure 3-dimensional velocity of the subject i. Then the variance matrix of the random variable $\mathbf{X}_i$ is calculated with

$$\mathrm{Var}(\mathbf{X}_i) = \mathbb{E}[(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])^T] \in \mathbb{R}^{3\times 3}.$$

To summarize the multidimensional variance in a scalar value, *the total variation of velocity* is calculated: $\mathrm{tr}[\mathrm{Var}(\mathbf{X}_i)]$, which is expected to be larger for healthy individuals than for patients. Indeed, patients may avoid moving with high speed. Naturally, the estimate $\mathrm{tr}[\mathrm{Var}(\mathbf{X}_i)]$ treats all data points in the time series similarly, without taking into account any sequential structure.

Also, the mean and the variance are calculated for *the distance between the head and the active hand.* Intuitively, the control group should have a bigger mean than patients, as patients are afraid of large movements. For a similar reason, the control group should have a higher variance than patients as well. Namely, healthy individuals likely spend more time in extremes of a movement, either having their hand very far from or very close to their body.

All in all, four different features are investigated for classifying patients. They are the time, total variation of velocity, mean distance between the head and the active hand, and variance of the distance between the

active hand and head. We perform a two-sample t-test for each variable and calculate the corresponding p-value. Two-sample t-test compares two random variables with each other with a null hypothesis that the underlying population means are the same. This test is a good one for feature selection because ideal features differentiate the patients and the control group out of the movement data in two distinct groups.

3.2 Time Series Segmentation Using HMMs

The time series data consists of predefined patterns, as seen in Figures 3 and 4. The structure of these patterns is approximated with HMMs in this section. HMMs are a commonly used method in pattern recognition, as they automatically segment sequential data in parts [3, 9, 26, 29, 42, 43]. HMMs handle similar patterns in time series well even if their durations differ. We train HMMs using standard expectation-maximization and forward-backward algorithms. Therefore, HMMs do not need a lot of data to be trained.

Here, HMMs learn a sequence of hidden states $\mathbf{q}_i \in \mathbb{R}^{T_i}$ using 3-dimensional observations $\mathbf{X}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_{T_i}\} \in \mathbb{R}^{3\times T_i}$ for each subject i. Here, hidden states practically represent parts of a movement. For example, lifting the hand up could be one part of a movement, and the second part could be moving the hand to the left. As the players perform movements in a predefined order in the Painlab study, we use a left-to-right topology for hidden states. Therefore, the learned sequence of hidden states has a simple form, $\mathbf{q}_i = \{1, 1, \ldots, 2, 2, \ldots, \ldots, K, K\}$, where K is the total number of hidden states.

First, we visually investigate with *which data channel* HMMs segment the patterns of movements best. Three different data channels are given to HMMs: positional data, velocity, and normalized velocity. HMMs with $K = 11$ hidden states are visually compared with each other in Figures 7, 8, 9, 18, and 19. The visual comparison is a simple and powerful technique to observe whether the model consistently segments time series data or not. If a model segments time series data inconsistently between the subjects, the model is obviously not learning the structure of the predefined patterns.

Practically, K was chosen to equal the number 11 based on visual inspection of Figures 12, 13, and 14. The number of hidden states K could also have been, for example, 10 or 12, which is taken into account in the data analysis in Section 3.3. However, three hidden states, as an example, would have clearly been too few to capture the essential information about the pattern. Similarly, segmenting time series with 20 hidden states creates many segments that are extremely short for many individuals.

After plotting the segmented data, *posterior predictive checking* is presented for the HMM. Posterior predictive checking is a very illustrative way of showing what the model learns and what it does not. As posterior predictive checking is practically simulating replicated data under the fitted model and then comparing these to the observed data, we can then systematically compare discrepancies between real and simulated data [13]. Posterior predictive checking is presented only with HMMs that use normalized velocity because they segment the data the best, which can be seen in Figures 9, 18, and 19.

3.3 Classifying Patients and Controls with Logistic Regression

Two logistic regression models are presented in this section. The first one takes *non-segmented data* and the second one *segmented data* as features. In non-segmented data, the features presented in Section 3.1 are calculated using the whole time series. In segmented data, the features are calculated based on multiple segments created by an HMM, which we describe in detail in this section. The most promising features to differentiate the investigated groups are the *time* and the *total variation of velocity* based on Section 4.1. We calculate the average predictive likelihood and accuracy for both models. We use the binomial test for classification accuracy, to test that the results are significantly better than random, and calculate the corresponding p-value.

The two logistic regression models are compared with four different datasets. The first dataset equals patterns 1L and 1R, the second 2L and 2R, the third 3L and 3R, and the fourth 4L and 4R. To maximize the size of the test set, we use leave-one-out cross-validation, which is illustrated in Figure 5. We have $18 + 20 = 38$ data points for

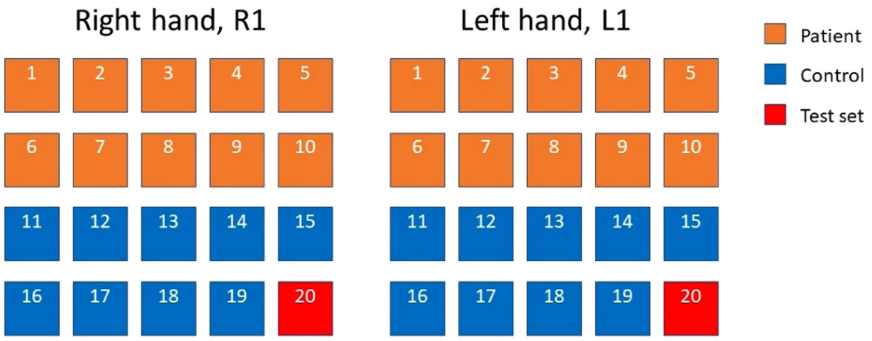

Fig. 5. Cross-validation setup. Data from both hands for 19 participants are chosen to train the classification model. The remaining participant's data from both hands is the whole test set in the current iteration. Naturally, all participants equal the test set iteratively one by one, regardless of whether the participant belongs to the control group or to the patients. Note that two classifications are made for the participant in the test set: one for the left hand and the other for the right hand.

each of these four datasets, as the data collection was not successful for two participants' right hand. Considering all the time intervals, we have $4 \cdot 38 = 152$ data points in total.

Again, the eight patterns are clearly visualized in Figures 3 and 4. Both these figures illustrate how the subjects perform the first four patterns of movements faster than the last four patterns. Therefore, the time and the total variation of velocity features are always normalized by the corresponding *mean*. After the normalization, the models can compare the data created by the left and right hands.

The segmented data in the other logistic regression is created with HMMs according to the results in Section 4.2. We have knowledge about the structure of each pattern, as we know the locations of all objects, which we described in Section 2. Therefore, we can identify the three movements of each pattern based on the segments created by HMMs. The time interval is divided into movements and noise between the movements. As an example, in Figure 9, the segments $\{2, 3, 6, 7, 10, 11\}$ would correspond to the movements and the remaining segments $\{1, 4, 5, 8, 9\}$ would correspond to the noise. The time and the total variation of velocity are calculated for both parts of these time intervals, which leads to four features instead of only two, which could increase the classification accuracy. Time series data is segmented to $K = 10, K = 11$, and $K = 12$ hidden states with an HMM to make sure the results between these choices of K are similar to another.

4 RESULTS

4.1 Feature Selection

Here, we investigated whether some features derived in Section 3.1 have a different distribution between the healthy individuals and the patients. Clearly, the features time and total variation of velocity separate the investigated groups well, as illustrated in Figures 6, 15, 16, and 17 (see Appendix C), but the mean and variance of the distance between the head and the active hand do not discriminate the two groups well (see p-values for the two-sample t-test in Tables 1 and 2). The features time and the total variation of velocity have p-values < 0.05. On the contrary, the mean and variance of the distance have p-values > 0.05 for all patterns.

Time is a good feature to describe differences between the groups, meaning the patients are slower than the healthy individuals. Indeed, the patients are stiff as they are avoiding feeling sudden pain. The total variation of velocity also separates the investigated groups well because the healthy individuals have a heavier tail in their velocity distribution than the patients. In other words, the healthy individuals are not as afraid of high speed or rapid changes in speed as the patients are. Neither the mean nor the variance of the distance between the head and the hand separates the groups well, which is a bit surprising because the patients should avoid large movements. Therefore, only time and the variance of velocity are used as features in upcoming models.

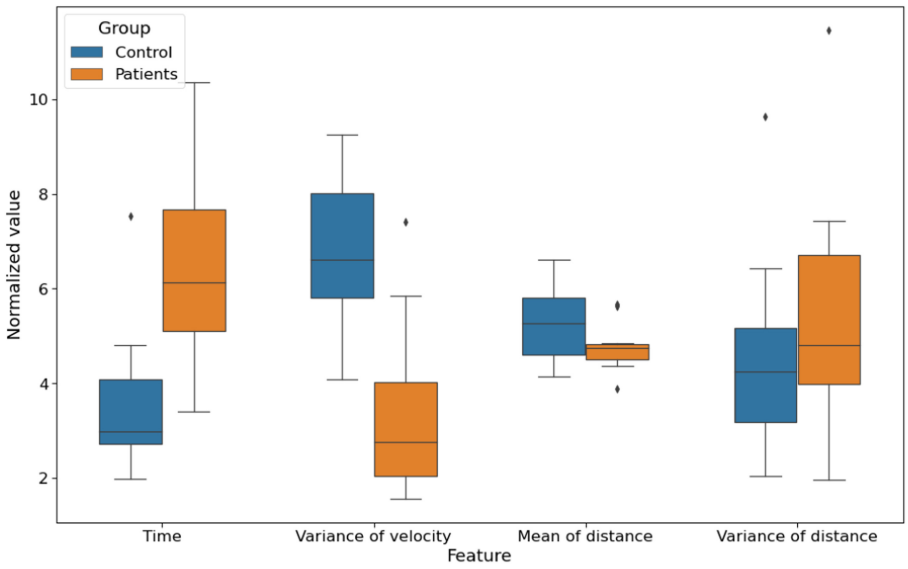

Fig. 6. Comparing distributions between healthy individuals and patients over different features *on the pattern 1L*. Clearly, time and variance of velocity have distributions between the groups that do not overlap with each other. The same is observed in Figures 15, 16, and 17. However, the mean and the variance of velocity have distributions with a clear overlap. Indeed, these features are not suitable to differentiate the investigated groups.

Table 1. Two-sample t-test for the Features Time and Velocity's Variance

	Time		Velocity's Variance	
	p-value	CI95%	p-value	CI95%
Pattern 1	0.0035	$[-20, -4.6]$	0.00096	$[0.07, 0.23]$
Pattern 2	0.001	$[-19, -5.8]$	0.000003	$[0.17, 0.32]$
Pattern 3	0.000117	$[-16, -6.2]$	0.000006	$[0.17, 0.35]$
Pattern 4	0.000009	$[-19, -9.1]$	0.000005	$[0.16, 0.31]$

Collected p-values and the confidence intervals of the difference in means on the patterns 1L, 2L, 3L, and 4L.

Table 2. Two-sample t-test for the Features Mean of the Distance and Variance of the Distance

	Mean, Distance		Variance, Distance	
	p-value	CI95%	p-value	CI95%
Pattern 1	0.16	$[-0.02, 0.13]$	0.52	$[-0.01, 0.00]$
Pattern 2	0.29	$[-0.04, 0.14]$	0.98	$[-0.01, 0.01]$
Pattern 3	0.063	$[-0.01, 0.18]$	0.86	$[-0.01, 0.01]$
Pattern 4	0.051	$[-0.01, 0.16]$	0.21	$[-0.00, 0.02]$

Collected p-values and the confidence intervals of the difference in means on the patterns 1L, 2L, 3L, and 4L.

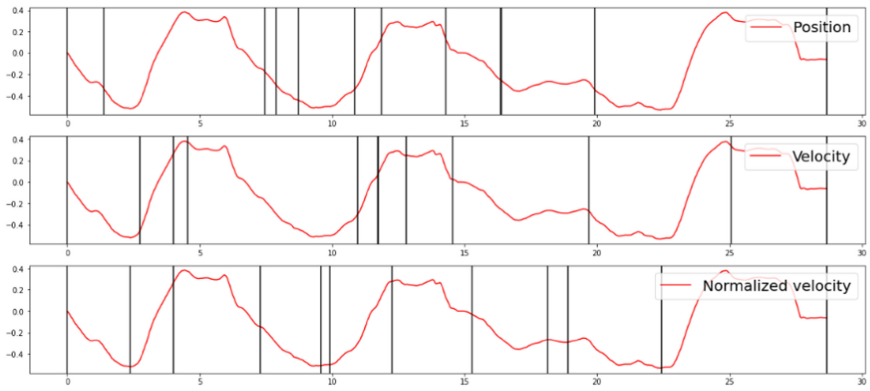

Fig. 7. All the red lines correspond to the same patterns of movements on the z-axis by the same patient. However, the time series data is segmented with three different hidden Markov models, which use different observations in their parameter estimation. Obviously, the hidden Markov model with *positional data* seems not to learn a lot of essential information about the time series, as the model leaves the first and last peak unsegmented. However, the conclusion about why the hidden Markov model with *normalized velocity* is a better model than the one trained with *velocity* requires comparison between this figure and Figure 8.

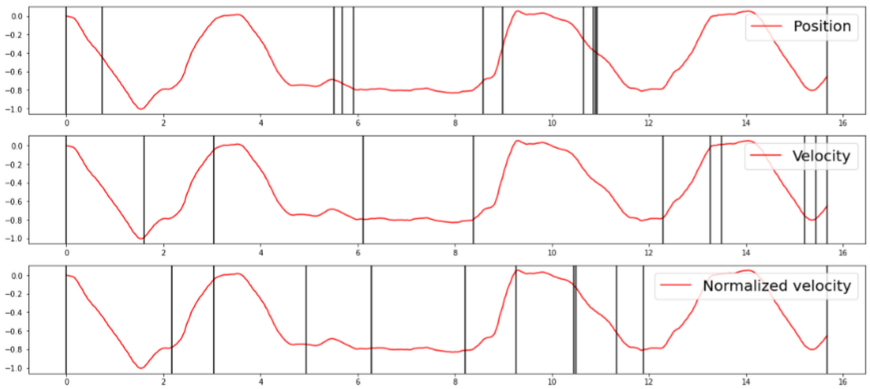

Fig. 8. This figure is similar to Figure 7 but now the time series of the same pattern is created by a healthy individual. The comparison of these two figures shows how the hidden Markov model with *velocity* segments the second red peak in multiple parts in Figure 7 but completely leaves the peak unsegmented in this figure. Indeed, **the hidden Markov model with *normalized velocity* is the only model that segments the time series of these two subjects consistently**. It may seem that the hidden Markov model with normalized velocity should have segmented the last peak even further, which is located between 12 and 16 seconds in this Figure 8. In reality, the x-axis and the y-axis are the most dominating dimensions at this time point.

4.2 Time Series Segmentation Using HMMs

Different HMMs are investigated and compared according to the methodology introduced in Section 3.2. Essentially, the goal of this section is to understand *how* different HMMs segment time series data and if some HMM segments time series the best.

The same pattern of movements is segmented by a patient with three different observations (positional, velocity, normalized velocity) in Figure 7. Only one dimension z is shown in the plots even if the models were fitted on 3-dimensional data. The same plotting is conducted for a healthy individual in Figure 8. The comparison of

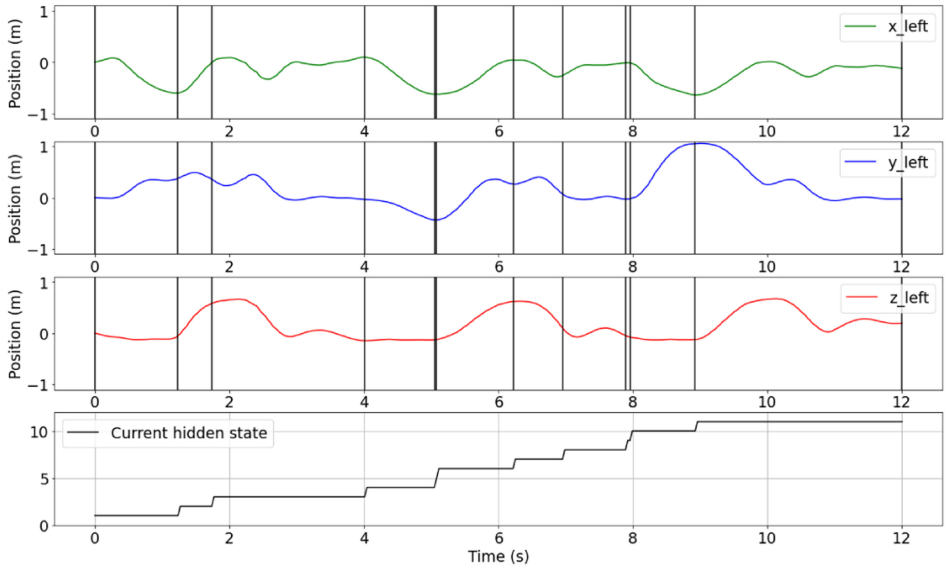

Fig. 9. Segmenting the last pattern of movements of a healthy individual with an HMM that uses *normalized velocity*. The segmentation is similar to the segmentation in Figures 18 and 19, which indicates that the hidden Markov model segments time series data reliably.

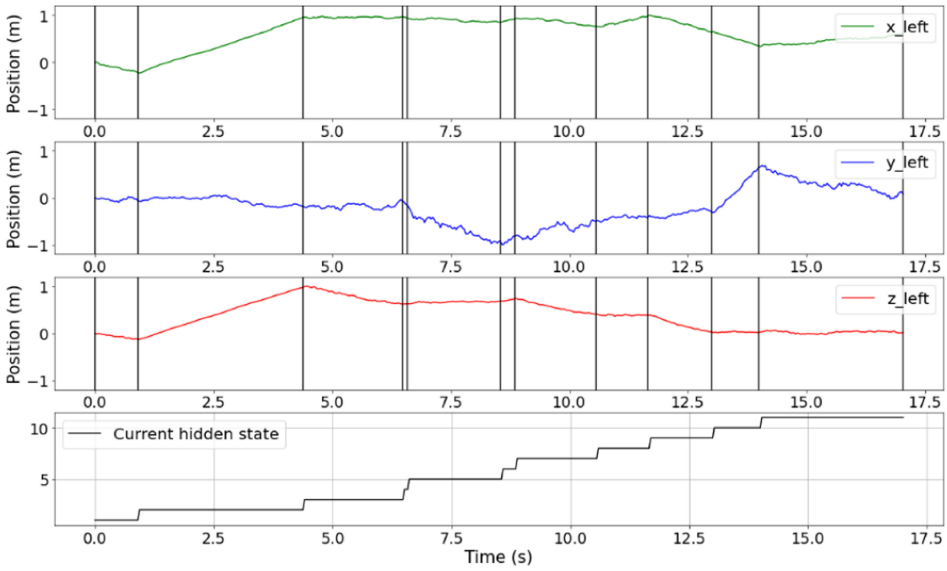

Fig. 10. Posterior predictive checking with an HMM that uses *normalized velocity*. The data has similar form to time series data in Figure 9. For example, the derivative of the x- and z-axis is positive in the second segment in both these figures.

these figures illustrates quite well how the HMM with *normalized velocity* is the only model out of these three that segments the time series of these two subjects consistently.

Of course, the conclusion that the HMM with *normalized velocity* segments time series reliably should not rely only on Figures 7 and 8. Therefore, the reader can compare Figures 9, 18, and 19 with each other, which have been segmented similarly; see Appendix D.

Table 3. Logistic Regression with Non-segmented and Segmented Data: Classification Accuracy and Average Predictive Likelihood on the Test Set

	Non-segmented Logistic Regression		Segmented Logistic Regression	
	Pred. Likelihood, Mean	Accuracy	Pred. Likelihood, Mean	Accuracy
Pattern 1	0.71 ± 0.063	0.84 ± 0.12	0.74 ± 0.072	0.89 ± 0.098
Pattern 2	0.74 ± 0.055	0.89 ± 0.098	0.79 ± 0.062	0.89 ± 0.098
Pattern 3	0.73 ± 0.053	0.87 ± 0.11	0.76 ± 0.065	0.84 ± 0.12
Pattern 4	0.71 ± 0.062	0.84 ± 0.12	0.76 ± 0.069	0.79 ± 0.13
All patterns	0.72 ± 0.028	0.86 ± 0.055	0.76 ± 0.032	0.86 ± 0.055

The uncertainty ± sign corresponds to 95 percent confidence intervals, which are calculated using the t-distribution.

Lastly, posterior predictive checking, i.e., simulated data, is presented for the fitted HMM in Figures 10, 20, and 21. We can now systematically compare discrepancies between the simulated data and the real data in Figure 9. Data generated with a HMM that uses normalized velocity is in many ways realistic. Big movements, e.g., moving the left hand to the left upper corner, are parts of both real-world and generated data. Naturally, the generated data approximates real-world data with straight lines. Therefore, the real-world data has more rounded shapes than the simulated time series.

4.3 Classifying Patients and Controls with Logistic Regression

Section 3.3 introduced two models: logistic regression with non-segmented and segmented data. The predicted results for logistic regression with segmented data are calculated with three different numbers of hidden states: 10, 11, and 12. We represent the results of segmented logistic regression with $K = 10$ hidden states in this section because the model has the median classification accuracy out of these three models. However, two other segmented logistic regression models give very similar results; see Table 5 in Appendix E.

The average predictive likelihoods and classification accuracies of both logistic regression models are written in Table 3. The average likelihood measures the average certainty that the model gives for its prediction of the participant's true label. It is higher for the segmented than for the non-segmented logistic regression. Furthermore, both models have very good and similar classification accuracies (segmented: $\frac{130}{152}$, non-segmented: $\frac{131}{152}$). We performed a binomial test for both models to test the significance of the classification accuracy. Binomial test for $n = 152, K = 130, p = 0.5$ creates a p-value $< 10^{-8}$, which is a very significant value considering the standard significance level of 0.05.

5 DISCUSSION

Our research led to many significant contributions and suggestions for future research. These conclusions are emphasized with a cursive font in the following paragraphs.

The patients and the control group were classified very reliably. The logistic regression with non-segmented and segmented data classified the data into patients and healthy controls with similar and high accuracies (0.86) (see Table 3), which highly significantly deviate from a random classifier (see the last paragraph in Section 4.3). However, the classification accuracy of the logistic regression with segmented data could have been even better if there had been more data in use. Then the segmentation of different patterns would have been more exact and also the coefficients would have been estimated better.

The time and the total variation of velocity were powerful features to differentiate the patients and the control group; see Figure 6 and Tables 1 and 2. However, some other features could also have been useful in the classification, which were not discovered in this research. For example, it would be interesting to ask subjects to follow specific objects with their hand and then measure the distance between the active hand and the object. Patients could be expected to have problems in following these objects.

Hidden Markov models can divide movement data into meaningful segments, which creates possibilities for defining even more detailed features, with potential to improve accuracy, when larger datasets become available in the future. The segmented data contains all the information of the pattern that the non-segmented data has but also information about the pattern's structure. Therefore, the segmented data has a lot of potential to improve the classification accuracy. The logistic regression with segmented features outperformed the logistic regression with non-segmented features already with our small dataset when the performance was measured in terms of average predictive likelihood; see Table 3. Furthermore, many machine learning algorithms with big datasets could find out interesting relationships between segmented features, which logistic regression does not.

To apply methods suggested in this article in a real treatment situation in a game-like environment, comparable tasks should be automatically labeled by the software. The label should be specific for every possible task in the system. Also, the start and the end point of a task should be well specified without a need for an interpretation in, e.g., situations where a subject shuts down the VR system within the task. To make the final product intriguing to play, one could let the subject choose the order of comparable tasks.

6 CONCLUSION

We were able to classify the CLBP patients and the healthy volunteers with high accuracy with the movement data collected by using a VR program. We were also able to divide the movement data into useful segments for analysis by using HMMs.

Next, the correlation between clinical improvements, by using globally accepted clinical scales as the reference data, and the changes in the movement data should be researched. Because the variation of movement velocity in the controlled tests seems a potential measurement for classification of CLBP patients, the measurement of arbitrary movement by using accelerometer data should be further researched in the future. Obviously, this requires using repetitive longitudinal measurements. Furthermore, large study populations are recommended to enable clinical validity in the future. The methods based on time series segmentation should be further developed in the future to explain the changes in the movement data in CLBP patients over the time, while their condition is slowly changing.

The findings in this research are a first step toward an objective quantification of chronic low back pain with virtual reality movement data. A reliable biomarker would not only help patients understand how their condition is improving over time but would also benefit the healthcare professionals who are trying to assess the severity of a patient's condition before and during the treatment.

APPENDICES

A CORRELATION BETWEEN AGE AND THE FEATURES USED IN MODELS

As the patients are clearly older than the control group on average, we investigate the correlations between the age and the features that are used is our models in Section 4.3. These variables' Pearson correlation coefficients are compared with another within the group in Figure 11, which shows that there are differences in how the age is connected with the features in the classification model. Furthermore, we conducted a two-way ANOVA with age and case-control indicators for each of the four patterns and two predictive features. The results presented in Table 4 show that in all eight cases the majority of variation was explained by the case-control status, as the sum of squares $SS(1)$ is constantly much bigger than $SS(2)$ and $SS(12)$. None of the age effects were statistically significant, and only one age-group interaction was significant ($p \approx 0.05$). Hence, we conclude that age does not explain our results.

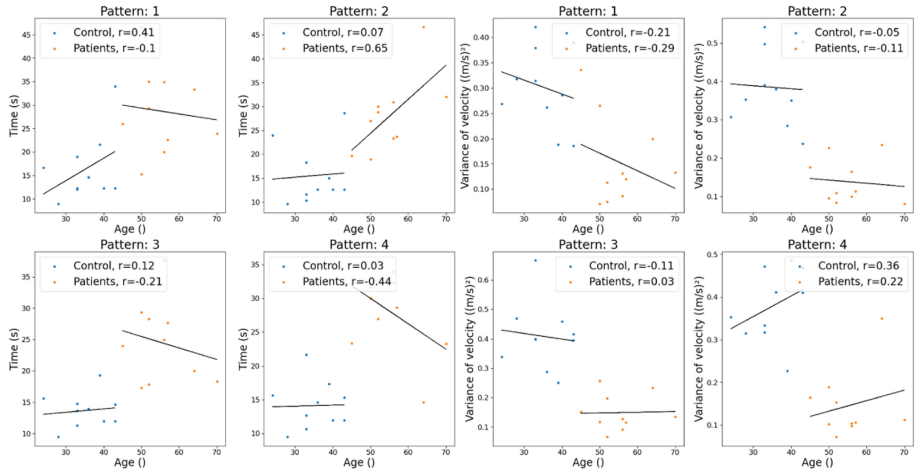

Fig. 11. Correlations between the age and features time and total variation of velocity. The features are calculated from the data created by the left hand. We see how the sign of the Pearson correlation coefficients varies between different patterns, which indicates that the effect of participant's age does not have a consistent effect on our predictions. Thus, even if the age groups are different in our dataset, the predictive features seem to be not consistently dependent on age.

Table 4. ANOVA

	Case Control		Age		Intersection	
	SS(1)	Pr(> F)	SS(2)	Pr(> F)	SS(12)	Pr(> F)
Pattern 1L, time	761	0.0027	41.0	0.42	209	0.081
Pattern 2L, time	787	0.00069	0.23	0.94	206	0.048
Pattern 3L, time	589	0.00018	589	0.32	12.5	0.49
Pattern 4L, time	968	0.000026	12.1	0.53	0.174	0.94
Pattern 1L, var. of vel.	0.110	0.0011	0.0128	0.20	0.00194	0.61
Pattern 2L, var. of vel.	0.303	0.000010	0.000684	0.77	0.000083	0.92
Pattern 3L, var. of vel.	0.335	0.000021	0.000215	0.88	0.000038	0.95
Pattern 4L, var. of vel.	0.274	0.000016	0.00202	0.61	0.000886	0.73

Features time and total variation of velocity are explained with age and case-control indicators.

B HMM: PREDICTIVE PERFORMANCE

B.1 Methods

Two HMMs are trained in this section: one model with the patients' and one with the control group's data. These models are compared with each other to find out which model predicts the likelihood of an unseen individual better. Indeed, this approach takes advantage of HMMs in a more standard way than how we used them previously.

Three indicators describe here the predictive performance of HMMs: the *comparative predictive performance*, the *control group's predictive performance*, and the *binary accuracy*. In these equations, patients are labeled with $1, \ldots, 10$ and healthy individuals with $11, \ldots, 20$. The predictive performance is validated with the leave-one-out cross-validation. This leave-one-out comparison is presented for different data channels (positional data, velocity, and normalized velocity), different number of hidden states (1–30), but only for the pattern of movements 1L. Only one pattern of movement is enough, as none of these models classified the subjects better than the logistic regression models.

The comparative predictive performance is a logical quantity to measure. At the end, a central goal of this research is to find differences between the investigated groups. The model that maximizes the difference between these two HMMs can be therefore considered a significant one. Intuitively, the comparative predictive performance compares the likelihoods of two different HMMs for each subject. Let's define the comparative predictive performance:

$$\sum_{i=1}^{10} \left\{ \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{pain}}^{-i}\right)\right) - \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{control}}^{-(i+10)}\right)\right) \right\} - \sum_{i=11}^{20} \left\{ \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{pain}}^{-(i \bmod 10)}\right)\right) - \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{control}}^{-i}\right)\right) \right\},$$

where $\mathbf{X}_i$ are the observations. Furthermore, ϕ_{pain}^{-i} are the learned parameter values of the HMM trained with patients' data but *without* the individual i. On the other hand, $\phi_{\text{control}}^{-(i+10)}$ are the trained parameter values with the control group's data but *without* the individual $(i + 10)$. Indeed, both parameters are trained then with 9 individuals. Naturally, the same logic holds for the case when the subject $i \in \{11, \ldots, 20\}$, i.e., the subject belongs to the control group.

The likelihood of the HMM that is trained only with the control group's data is also investigated. This control group's predictive performance describes how much adding a new state improves the likelihood of the time series. The control group's predictive performance is defined in the following manner:

$$\sum_{i=11}^{20} \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{control}}^{-i}\right)\right).$$

The binary accuracy is an interesting and a very concrete indicator of the predictive performance of HMM. However, the binary classification does not tell anything about the certainty of the model's prediction. Furthermore, the binary accuracy is a step function where each jump is $\frac{1}{20} = 0.05$ at the smallest, as the size of dataset is 20. Let's define the Binary accuracy:

$$\sum_{i=1}^{10} 1\left\{ \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{pain}}^{-i}\right)\right) - \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{control}}^{-(i+10)}\right)\right) > 0 \right\} + \sum_{i=11}^{20} 1\left\{ \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{control}}^{-i}\right)\right) - \log\left(p\left(\mathbf{X}_i \mid \phi_{\text{pain}}^{-(i \bmod 10)}\right)\right) > 0 \right\}.$$

B.2 Results

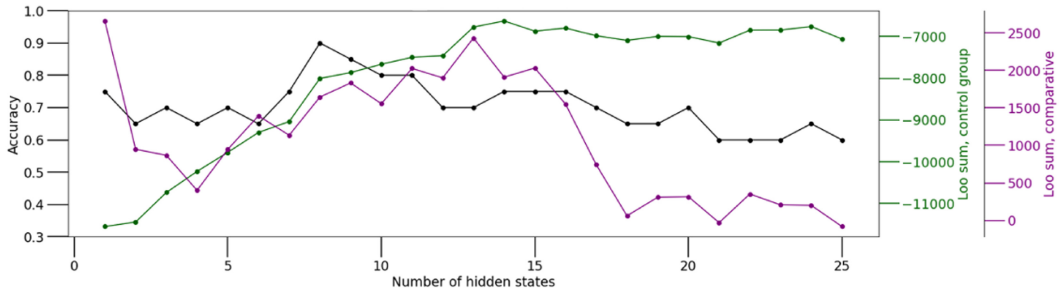

Fig. 12. The predictive performance as a function of hidden states when the observations are 3-dimensional *velocity* on pattern 1L. The classification accuracy is not bad but clearly worse than with the logistic regression models in Section 4.3. The control group's predictive performance plotted with green describes how much the likelihood gets better by adding a new hidden state. Practically somewhere around 10 to 15 hidden states this improvement is not very high. This last observation can also be made from the models in Figures 13 and 14.

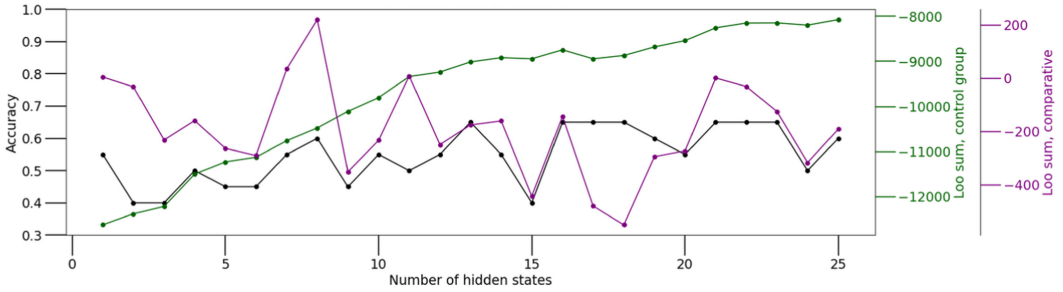

Fig. 13. The predictive performance as a function of hidden states when the observations are 3-dimensional *normalized velocity* on pattern 1L. Clearly, the classification accuracy is poor. Again, the control group's predictive performance does not increase very much after 10 hidden states.

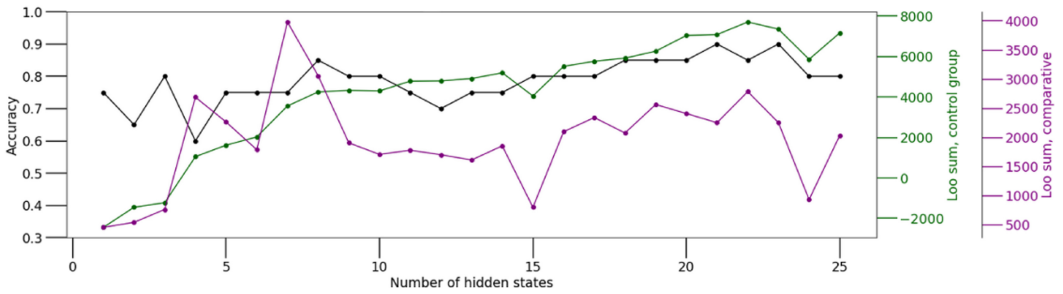

Fig. 14. The predictive performance as a function of hidden states when the observations are 3-dimensional *positional data* on pattern 1L. Clearly, the classification accuracy is not better compared with the logistic regression models in Section 4.3. Also in this figure, the control group's predictive performance stabilizes around 10 to 15 hidden states.

C FEATURE SELECTION

Figures 15, 16, and 17 all support the conclusion made in Figure 6.

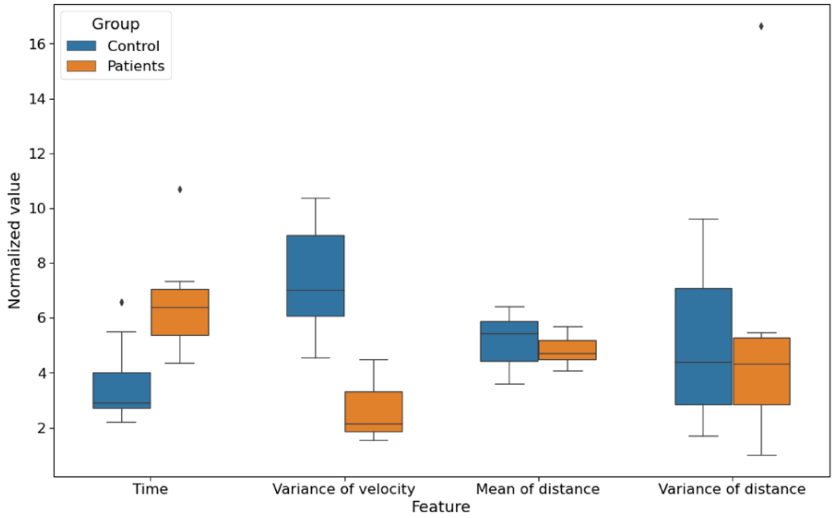

Fig. 15. Comparing distributions between the healthy individuals and the patients over different features *in the pattern 2L*.

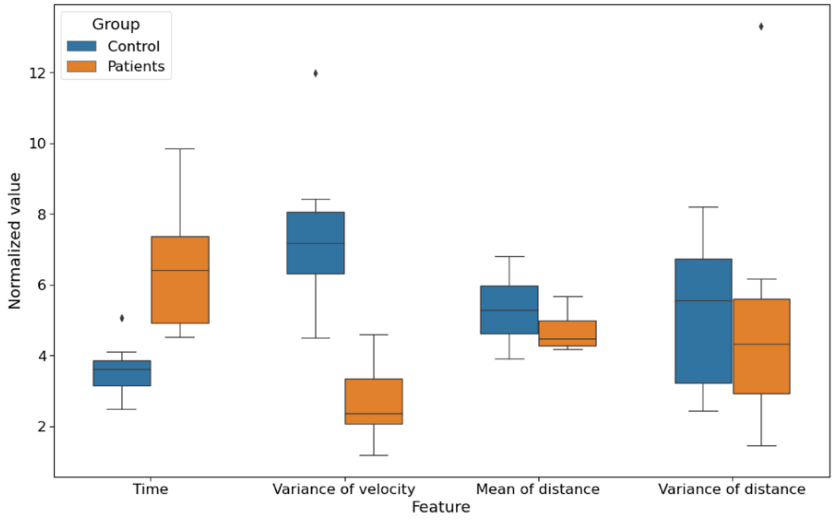

Fig. 16. Comparing distributions between the healthy individuals and the patients over different features *in the pattern 3L*.

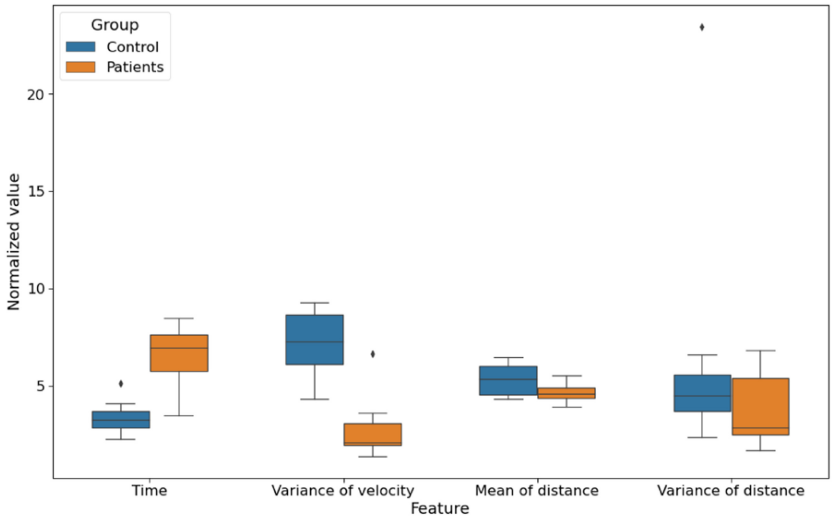

Fig. 17. Comparing distributions between the healthy individuals and the patients over different features *in the pattern 4L*.

D TIME SERIES SEGMENTATION USING HMMS

Hidden Markov models segment time series data in similar segments, which becomes clear when Figures 9, 18, and 19 are compared with each other.

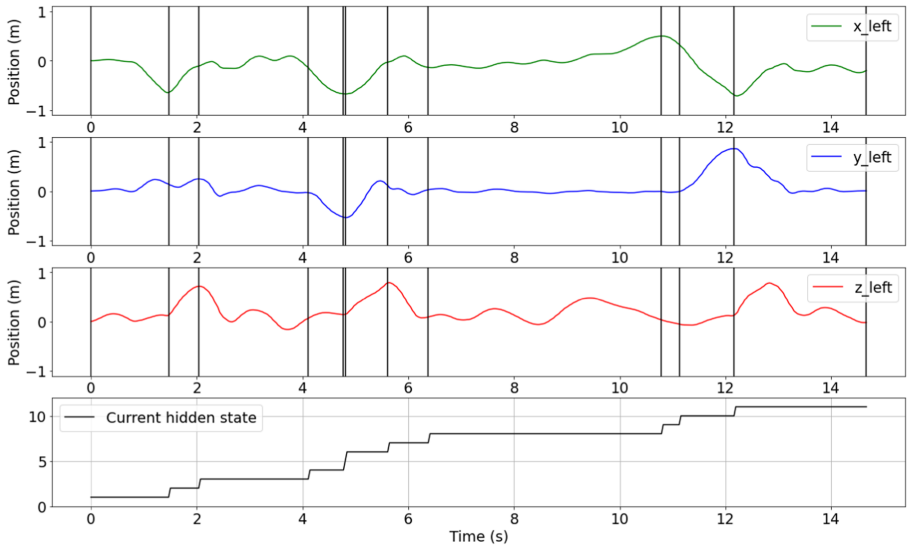

Fig. 18. Segmenting the last pattern of movements of a healthy individual with an HMM that uses *normalized velocity.*

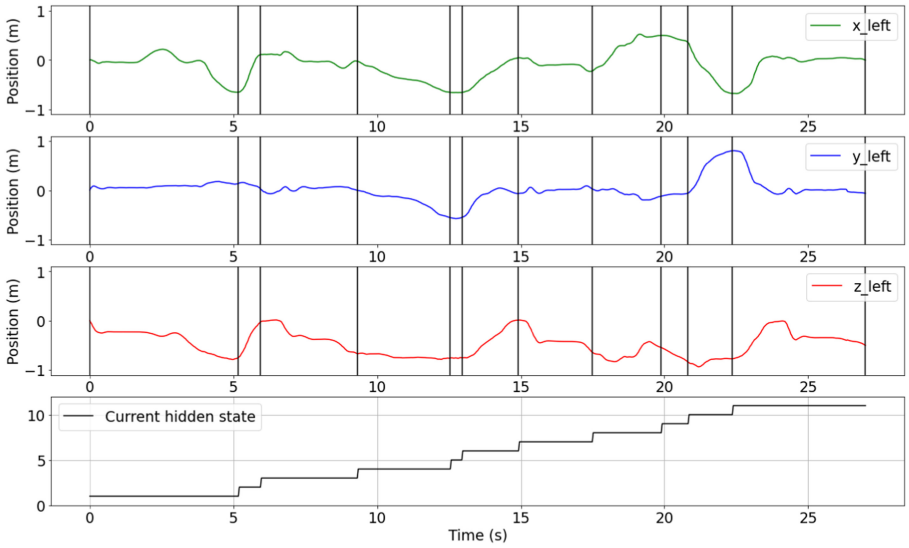

Fig. 19. Segmenting the last pattern of movements of a patient with an HMM that uses *normalized velocity.*

Furthermore, Figures 20 and 21 support the conclusion about the posterior predictive checking described in Figure 10.

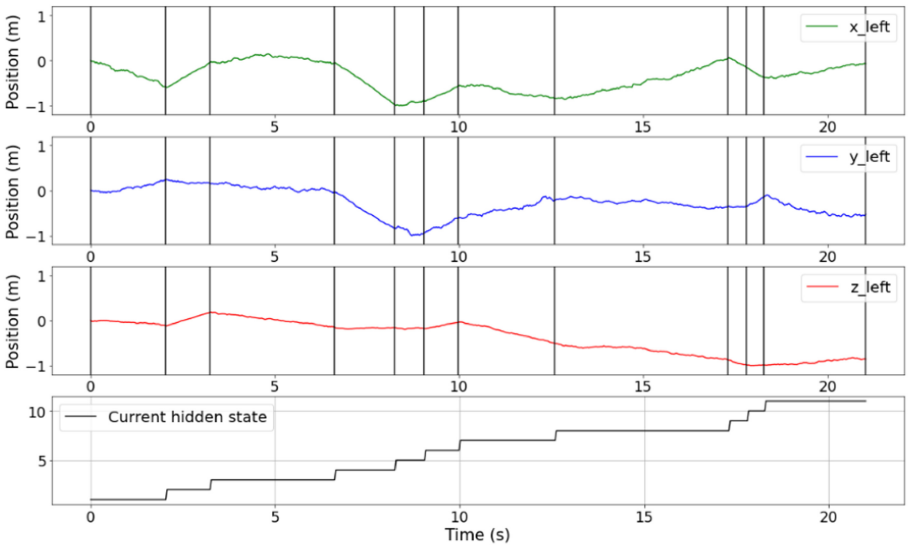

Fig. 20. Posterior predictive checking with a hidden Markov model that uses *normalized velocity.*

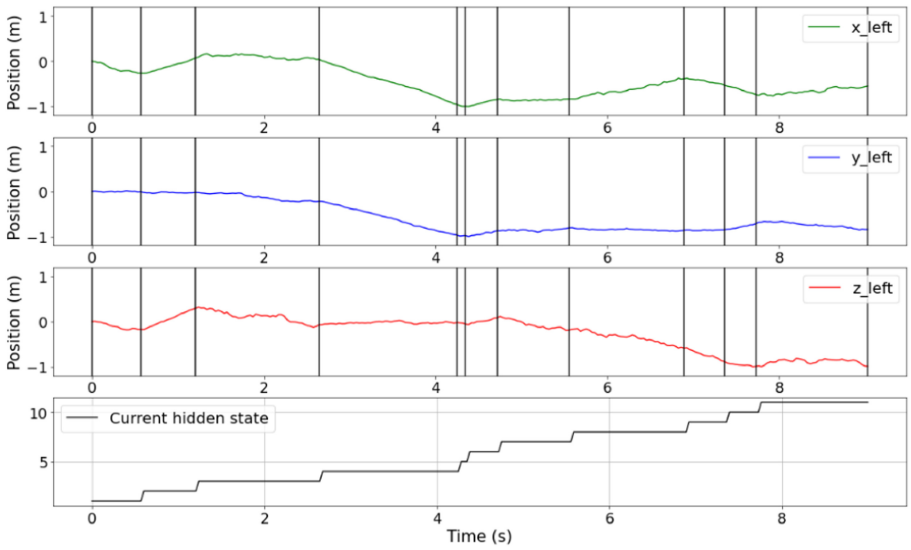

Fig. 21. Posterior predictive checking with a hidden Markov model that uses *normalized velocity*.

E LOGISTIC REGRESSION WITH SEGMENTED DATA, DETAILS

We also segment the time series with the number of hidden states $K = 11$ and $K = 12$ to make sure the results are similar to what we obtained in Table 3. Indeed, the accuracies and average predictive likelihoods are very similar with each other between all different segmented logistic regression models.

Table 5. Logistic Regression with Segmented Data: Classification Accuracy and Average Predictive Likelihood on the Test Set

	Segmented Logistic Regression, $K = 11$		Segmented Logistic Regression, $K = 12$	
	Pred. Likelihood, Mean	Accuracy	Pred. Likelihood, Mean	Accuracy
Pattern 1	0.74 ± 0.074	0.84 ± 0.12	0.73 ± 0.075	0.84 ± 0.12
Pattern 2	0.80 ± 0.057	0.87 ± 0.11	0.80 ± 0.061	0.92 ± 0.086
Pattern 3	0.76 ± 0.065	0.82 ± 0.12	0.75 ± 0.064	0.87 ± 0.11
Pattern 4	0.74 ± 0.073	0.82 ± 0.12	0.72 ± 0.072	0.82 ± 0.12
All patterns	0.75 ± 0.033	0.84 ± 0.059	0.75 ± 0.033	0.86 ± 0.055

The uncertainty ± sign corresponds to 95 percent confidence intervals, which are calculated using the t-distribution.

REFERENCES

[1] Shagun Adlakha, Deepak Chhabra, and Pratyoosh Shukla. 2020. Effectiveness of gamification for the rehabilitation of neurodegenerative disorders. *Chaos, Solitons & Fractals* 140 (Nov. 2020), 110192. https://doi.org/10.1016/j.chaos.2020.110192

[2] Seyed Amir Hossein Aqajari, Rui Cao, Emad Kasaeyan Naeini, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salanterä, Ariana Nelson, and Amir M. Rahmani. 2020. Pain assessment tool with electrodermal activity for post-operative patients: A method validation study. *JMIR mHealth and uHealth* 9 (Oct. 2020), e25258. https://doi.org/10.2196/25258

[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin.

[4] Kc. Brennan, Eva Jurewicz, Blair Ford, Seth Pullman, and Elan Louis. 2002. Is essential tremor predominantly a kinetic or a postural tremor? A clinical and electrophysiological study. *Movement Disorders: Official Journal of the Movement Disorder Society* 17 (April 2002), 313–316. https://doi.org/10.1002/(ISSN)1531-8257

[5] Sam Crowley, Angela Mickle, Margaret Wiggins, Josue Cardoso, Song Lai, Jared Tanner, Roland Staud, Roger Fillingim, Catherine Price, and Kimberly Sibille. 2022. Relationships between cognitive screening composite scores and pain intensity and pain disability in adults with/at risk for knee osteoarthritis. *Clinical Journal of Pain* 38 (May 2022), 470–475. https://doi.org/10.1097/AJP.0000000000001042

[6] K. Davis, E. Racine, and B. Collett. 2012. Neuroethical issues related to the use of brain imaging: Can we and should we use brain imaging as a biomarker to diagnose chronic pain? *Pain* 153 (Aug. 2012), 1555–1559. https://doi:10.1016/j.pain.2012.02.037

[7] Christopher Eccleston, Emma Fisher, Sammeli Liikkanen, Toni Sarapohja, Carina Stenfors, Satu Jääskeläinen, Andrew Rice, Leena Mattila, Taru Blom, and J. Bratty. 2022. A prospective, double-blind, pilot randomized controlled trial of an "embodied" virtual reality intervention for chronic low back pain in adults. *Pain* 163 (Sept. 2022), 1700–1715. https://doi.org/10.1097/j.pain.0000000000002617

[8] Dimitar Evdokimov, Johanna Frank, Alexander Klitsch, Stefan Unterecker, Bodo Warrings, Jordi Serra, Aikaterini Papagianni, Nadine Saffer, Caren Altenschildesche, Daniel Kampik, Rayaz Malik, Claudia Sommer, and Nurcan Üçeyler. 2019. Reduction of skin innervation is associated with a severe fibromyalgia phenotype. *Annals of Neurology* 86 (Aug. 2019), 504–516. https://doi.org/10.1002/ana.25565

[9] Gernot Fink. 2008. *Markov Models for Pattern Recognition: From Theory to Applications*. Springer-Verlag, London.

[10] Luay Fraiwan, Ruba Khnouf, and Abdel Mashagbeh. 2016. Parkinson's disease hand tremor detection system for mobile application. *Journal of Medical Engineering & Technology* 40 (March 2016), 1–8. https://doi.org/10.3109/03091902.2016.1148792

[11] Douglas French, Christopher France, François Vigneau, Julie French, and R. Evans. 2007. Fear of movement/(re)injury in chronic pain: A psychometric assessment of the original English version of the Tampa scale for kinesiophobia (TSK). *Pain* 127 (Feb. 2007), 42–51. https://doi.org/10.1016/j.pain.2006.07.016

[12] L. Garcia, B. Birckhead, P. Krishnamurthy, J. Sackman, I. Mackey, R. Louis, V. Salmasi, T. Maddox, and B. Darnall. 2021. An 8-week self-administered at-home behavioral skills-based virtual reality program for chronic low back pain: Double-blind, randomized, placebo-controlled trial conducted during COVID-19. *Journal of Medical Internet Research* 23 (Feb. 2021), e26292. https://doi.org/10.2196/26292

[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Taylor & Francis. https://books.google.fi/books?id=ZXL6AQAAQBAJ.

[14] Lydia Gómez, Alicia López-Martínez, and Gema Ruiz-Párraga. 2010. Psychometric properties of the spanish version of the Tampa Scale for Kinesiophobia (TSK). *The Journal of Pain: Official Journal of the American Pain Society* 12 (Oct. 2010), 425–435. https://doi.org/10.1016/j.jpain.2010.08.004

[15] Sanne Haas, R. Zoethout, Kristien Dyck, Marina De Smet, Laura Rosen, M. Murphy, K. Gottesdiener, R. Schoemaker, Adam Cohen, and Joop Van Gerven. 2011. The effects of TPA023, a GABAA 2,3 subtype-selective partial agonist, on essential tremor in comparison to alcohol. *Journal of Psychopharmacology (Oxford, England)* 26 (Sept. 2011), 282–291. https://doi.org/10.1177/0269881111415731

[16] Patrick A. Healy. 2019. The use of behavioral pain assessment tools and pain outcomes in nonverbal patients. *Honors Undergraduate Theses* 598 (2019).

[17] Henriette Janner, Catharina Klausenitz, Nancy Gürtler, Klaus Hahnenkamp, and Taras Usichenko. 2018. Effects of electrical transcutaneous vagus nerve stimulation on the perceived intensity of repetitive painful heat stimuli: A blinded placebo- and sham-controlled randomized crossover investigation. *Anesthesia & Analgesia* 126 (Jan. 2018), 1. https://doi.org/10.1213/ANE.0000000000002820

[18] Raed Joundi, John-Stuart Brittain, Ned Jenkinson, Alexander Green, and Tipu Aziz. 2011. Rapid tremor frequency assessment with the iPhone accelerometer. *Parkinsonism & Related Disorders* 17 (Feb. 2011), 288–290. https://doi.org/10.1016/j.parkreldis.2011.01.001

[19] S. K. Jääskeläinen. 2017. Quantitative sensory tests (QST) are promising tests for clinical relevance of anti-nociceptive effects of new analgesic treatments. *Scandinavian Journal of Pain* 17 (Oct. 2017), 144–145.

[20] Gráinne Kelly, Catherine Blake, Camillus Power, Declan O'Keeffe, and Brona Fullen. 2011. The association between chronic low back pain and sleep. *Clinical Journal of Pain* 27 (Feb. 2011), 169–181. https://doi.org/10.1097/AJP.0b013e3181f3bdd5

[21] Myoung Kim, Seong-Gil Kim, Young-Jun Shin, Eun-Hong Choi, and Yuwon Choe. 2017. The relationship between anterior pelvic tilt and gait, balance in patient with chronic stroke. *Journal of Physical Therapy Science* 30 (Jan. 2017), 27–30. https://doi.org/10.1589/jpts.30.27

[22] Scott Kollins, Denton DeLoss, Elena Canadas, Jacqueline Lutz, Robert Findling, Richard Keefe, Jeffery Epstein, Andrew Cutler, and Stephen Faraone. 2020. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): A randomised controlled trial. *Lancet Digital Health* 2 (Feb. 2020), e168–e178. https://doi.org/10.1016/S2589-7500(20)30017-0

[23] Heli Koskimäki, Henna Mönttinen, Pekka Siirtola, Hanna-Leena Huttunen, Raija Halonen, and Juha Röning. 2017. Early detection of migraine attacks based on wearable sensors: Experiences of data collection using Empatica E4. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp'17)*. Association for Computing Machinery, New York, NY, 506–511. https://doi.org/10.1145/3123024.3124434

[24] Brent Leininger, Craig Schulz, Zan Gao, Gert Bronfort, Roni Evans, Zachary Pope, Nan Zeng, and Mitchell Haas. 2017. The relationship between accelerometer-determined physical activity and clinical low back pain measures in adolescents with chronic or sub-acute recurrent low back pain. *Journal of Orthopaedic & Sports Physical Therapy* 47 (Sept. 2017), 1–20. https://doi.org/10.2519/jospt.2017.7345

[25] Tianjiao Li, Shuo Zhang, Eri Ikeda, and Hiroyuki Kobinata. 2022. Functional connectivity modulations during offset analgesia in chronic pain patients: An fMRI study. *Brain Imaging and Behavior* (March 2022), 1–9. https://doi.org/10.1007/s11682-022-00652-7

[26] Roderick J. A. Little and Donald B. Rubin. 1986. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

[27] Walter Maetzler, Jochen Klucken, and Malcolm Horne. 2016. A clinical view on the development of technology-based tools in managing Parkinson's disease: Technology-based clinical PD assessment. *Movement Disorders* 31 (June 2016), 1263–1271. https://doi.org/10.1002/mds.26673

[28] Ingemar S. J. Merkies, Catharina G. Faber, and Giuseppe Lauria. 2015. Advances in diagnostics and outcome measures in peripheral neuropathies. *Neuroscience Letters* 596 (June 2015), 3–13. https://doi.org/10.1016/j.neulet.2015.02.038.

[29] Sunanda Mitra and T. Acharya. 2007. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37 (2007), 311–324.
[30] Shotaro Moriyasu, Takahiro Shimizu, Makoto Honda, Yoshikazu Ugawa, and Ritsuko Hanajima. 2022. Motor cortical plasticity and its correlation with motor symptoms in Parkinson's disease. *eNeurologicalSci* 29 (Sept. 2022), 100422. https://doi.org/10.1016/j.ensci.2022.100422
[31] Muzeyyen Oz, Aynur Demirel, Yasemin Özel Aslıyüce, and Özlem Ülger. 2022. Education level is a strong determinant of cognitive function as measured by MoCA in people with chronic low back pain. *Musculoskeletal Science and Practice* 58 (Jan. 2022), 102503. https://doi.org/10.1016/j.msksp.2022.102503
[32] Muzeyyen Oz, Yasemin Özel Aslıyüce, Aynur Demirel, Hatice Çetin, and Özlem Ülger. 2021. Determination of cognitive status and influencing variables in patients with chronic neck pain: A cross-sectional study. *Applied Neuropsychology: Adult* (Oct. 2021), 1–8. https://doi.org/10.1080/23279095.2021.1980795
[33] André Pontes-Silva, Almir Dibai Filho, Sulamizia Jesus, Leonardo Oliveira, Daniela Bassi Dibai, Cid Gomes, and Mariana Avila. 2022. The best structure of the Tampa Scale for Kinesiophobia for patients with chronic low back pain has two domains and nine items. *Clinical Rehabilitation* 37 (Sept. 2022), 407–414. https://doi.org/10.1177/02692155221128829
[34] Benjamin Provencher, Stephane Northon, and Mathieu Piché. 2021. Segmental chiropractic spinal manipulation does not reduce pain amplification and the associated pain-related brain activity in a capsaicin-heat pain model. *Frontiers in Pain Research* 2 (Oct. 2021), 733727. https://doi.org/10.3389/fpain.2021.733727
[35] O. Skogberg, L. Karlsson, B. Börsbo, L. Arendt-Nielsen, T. Graven-Nielsen, B. Gerdle, E. Bäckryd, and D. Lemming. 2022. Pain tolerance in chronic pain patients seems to be more associated with physical activity than with depression and anxiety. *Journal of Rehabilitation Medicine* 54 (April 2022), jrm00286. https://doi.org/10.2340/jrm.v54.241
[36] Karsten Specht. 2020. Current challenges in translational and clinical fMRI and future directions. *Frontiers in Psychiatry* 10 (Jan. 2020), 924. https://doi.org/10.3389/fpsyt.2019.00924
[37] Hanne Storm. 2009. Changes in skin conductance as a tool to monitor nociceptive stimulation and pain. *Current Opinion in Anaesthesiology* 21 (Jan. 2009), 796–804. https://doi.org/10.1097/ACO.0b013e3283183fe4
[38] Paul Thiry, Martin Houry, Laurent Philippe, Olivier Nocent, Fabien Buisseret, Frédéric Dierick, Rim Slama, William Bertucci, André Thevenon, and Emilie Simoneau. 2022. Machine learning identifies chronic low back pain patients from an instrumented trunk bending and return test. *Sensors* 22 (July 2022), 5027. https://doi.org/10.3390/s22135027
[39] Thomas Toelle, Daniel Utpadel-Fischler, Katharina-Kristina Haas, and Janosch Priebe. 2019. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: A randomized controlled trial. *NPJ Digital Medicine* 2 (May 2019), 34. https://doi.org/10.1038/s41746-019-0109-x
[40] Emilie van Brummelen, Dimitris Ziagkos, Wadim Boon, Ellen Hart, Robert Doll, Teppo Huttunen, Petteri Kolehmainen, and G. Groeneveld. 2020. Quantification of tremor using consumer product accelerometry is feasible in patients with essential tremor and Parkinson's disease: A comparative study. *Journal of Clinical Movement Disorders* 7 (April 2020). https://doi.org/10.1186/s40734-020-00086-7
[41] Jeanine Verbunt, Ivan Huijnen, and Albère Köke. 2009. Assessment of physical activity in daily life in patients with musculoskeletal pain. *European Journal of Pain* 13 (March 2009), 231–242. https://doi.org/10.1016/j.ejpain.2008.04.006
[42] Xiaoyan Wang, Ming Xia, Huiwen Cai, Yong Gao, and Carlo Cattani. 2012. Hidden-Markov-models-based dynamic hand gesture recognition. *Mathematical Problems in Engineering* 2012 (April 2012). https://doi.org/10.1155/2012/986134
[43] A. D. Wilson and A. F. Bobick. 1999. Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 9 (1999), 884–900. https://doi.org/10.1109/34.790429
[44] Bart Witjes, Sylvain Baillet, Mathieu Roy, Robert Oostenveld, Frank Huygen, and Cecile de Vos. 2021. Magnetoencephalography reveals increased slow-to-fast alpha power ratios in patients with chronic pain. *PAIN Reports* 6 (June 2021), e928. https://doi.org/10.1097/PR9.0000000000000928
[45] Daehyun Yoon, Angela Fast, Peter Cipriano, Bin Shen, Jessa Castillo, Christopher Mccurdy, Carina Aparici, Deirdre Lum, and Sandip Biswal. 2021. Sigma-1 receptor changes observed in chronic pelvic pain patients: A pilot PET/MRI study. *Frontiers in Pain Research* 2 (Oct. 2021), 711748. https://doi.org/10.3389/fpain.2021.711748
[46] Rwei-Ling Yu, Wei-Ju Lee, Jie-Yuan Li, Yung-Yee Chang, Chin-Chung Chen, Juei-Jueng Lin, Yueh-Feng Sung, Tsu-Kung Lin, and Jong-Ling Fuh. 2020. Evaluating mild cognitive dysfunction in patients with Parkinson's disease in clinical practice in Taiwan. *Scientific Reports* 10 (Jan. 2020), 1014. https://doi.org/10.1038/s41598-020-58042-2
[47] Jason Zafereo, Brittany Uhlenbrock, Seth Watson, Sharon Wang-Price, Carl Noe, Robin Jarrett, Karen Meltzer, and Mu Huang. 2022. Using quantitative sensory testing to predict attrition in an interdisciplinary pain management program: A pilot study. *Pain Management* 12 (March 2022), 623–633. https://doi.org/10.2217/pmt-2021-0053
[48] Xiaoping Zheng, Michiel Reneman, Jone Ansuategui Echeita, Rita Preuper, Herbert Kruitbosch, Egbert Otten, and Claudine Lamoth. 2022. Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach. *Computers in Biology and Medicine* 144 (Feb. 2022), 105329. https://doi.org/10.1016/j.compbiomed.2022.105329

Received 15 June 2022; revised 1 November 2022; accepted 16 December 2022