Heuristic Detection of Recovery Progress using Robotic Data

Yujun Lai*, Sheila Sutjipto, Marc Carmichael and Gavin Paul

Centre for Autonomous Systems (CAS)

Faculty of Engineering & Information Technology

University of Technology Sydney (UTS), Sydney, Australia

Abstract—**Assessment methods for rehabilitation and recovery have recently been the focal point of research for medical professionals and engineers alike. Current assessment protocols rely on historical ordinal metrics which have been disputed despite their inter-rater reliability. Contemporary kinematic measures have allowed for new approaches to assess recovery progress. However, the abundance of data has deterred medical professionals from adopting these new protocols. This paper presents a method, based on the RMSE-LWSS (Longest Warping Subsequence) score, to distinguish outliers from systemic change for updating the personalized exercise path for users. By treating change detection as a classification problem, the incorporation of a compromised path based on the user's current capability is possible. Experiments were conducted to verify the efficacy of the method, comparing against statistical techniques for change detection and classification of pre-determined paths. The paper highlights how readily available data, rather than complex sensor systems, can be utilized to improve the robustness of personalization capabilities for robotic rehabilitation systems.**

I. INTRODUCTION

Rehabilitation and recovery of limb capabilities after injuries, such as those arising from sports or stroke, is complex due to the intricate nature of the human body. For the upper limb system, the inherent dexterity and maneuverability magnifies the significance of post-injury rehabilitation.

Injuries, in particular limb impairment, are unique to each patient. However, widely accepted neurological theories have lead to the current protocols for recovery and rehabilitation of limb capability. These protocols involve hierarchical exercise regiments, relying on expert assessments to gate-keep progression in the rehabilitation program.

The assessments performed by medical professionals are standardized based on the outcome measures used, and are currently judged across the field by their inter-rater reliability [1]. Termed "functional measures", due to their measurement of the ability to perform functional movements for Activities of Daily Living (ADL), these ordinal measures have long been scrutinized due to their reliance on subjective opinions.

Recently, there has been heavy scrutiny on the historical functional outcome measures due to their limited granularity and their inability to profile the whole upper limb system during assessment. The introduction of robotic systems in rehabilitation have provided a fresh perspective in measuring progress for recovery [2], most notably by providing access to objective data that is conducive to cardinal measures. However, the most widely used measures for progress still remains with

functional measures such as the Fugl-Meyer Assessment and the Barthel scale [3][4].

Introduced alongside the systems is the burgeoning field of kinematic measures, using objective metrics to provide limb capability assessment with fine granularity. However, there have been concerns over their slow adoption due to the overwhelming amount of data which therapists have to manage and analyze to extract meaningful data [5].

Although it might seem straightforward to use all available kinematic data for analyzing and assessing patient progress, computational bottlenecks impede their feasibility in real-world scenarios. Intuitively, in any recovery scenario, there are unique attributes which provide an indication of progress depending on the technology available. The effective utilization of data is posited to provide similarly accurate results for assessment when compared against using *all* available data. Hence, using robotic systems can potentially lower the adoption barriers for both rehabilitation and recovery assessment. In this paper, a novel metric is introduced and compared with different methods to detect change based on a pre-defined path.

This paper is organized as follows: Section II reviews related works towards utilizing data for rehabilitation assessment and learning models to detect change, Section III outlines the experiment and data processing setup while Section IV presents the results from the experiments. Section V discusses the results and possible limitations, and Section VI provides conclusions and suggests future work.

II. REVIEW OF RELATED WORKS

A. Data Utilization in Rehabilitation and Recovery

The rising trend in kinematic measures used for patient assessment during neurorehabilitation has lead to new frontiers in data utilization for patient progress assessment. With Randomized Controlled Trials (RCT) conducted in the medical field using contemporary kinematic metrics, results are starting to indicate certain data are conducive to rehabilitation and recovery assessment [6].

This is supplemented by earlier studies that suggest profiling the capabilities of the upper limb system requires a multi-lateral approach towards assessment [7]. Some studies even suggest that current measures and therapy facilitation are inappropriate for the assessment of limb progress [8].

Echoing these opinions, a study in personalized stroke rehabilitation found a correlation between the severity of limb impairment and the score obtained from a path-following exercise [9]. It showed how scenario- and task-based selection of metrics provide a more effective indicator for diagnosis.

From an extensive survey of upper-limb robotic rehabilitation systems [10], the most common method of obtaining assessment metrics arise from medical theories and an understanding of the human body. These include surface electromyography activation, kinematic joint observers, and wrench observers at the end-effector of robotic manipulators.

However, most systems do not tackle current barriers to adopt robotic assessment systems for therapists. Factors such as complexity of the data analysis and the metrics used to gauge process are rarely discussed [11]. A popular method to approach complexity reduction is the use of machine learning [12]. Using machine learning to build a model of a user's limb capability facilitates the potential to detect a change in user limb capability.

B. Dataset Shift

In machine learning, the training data set used is assumed to be representative of the distribution from which test samples are generated. We can exploit this to capture the user's limb capability at a specific point in time, detecting any limb capability changes as a shift in the sample generation distribution.

Termed *Dataset Shift*, the field is split into two actions: change detection of the sample generation distribution; and model adaptation to encapsulate the new sample generation distribution. While literature in Dataset Shift is abundant[13], many works rely on benchmark data sets which have been criticized for their detachment from real-world physical applications [14]. A notable exception is the predictive control of interruptions in an industrial plant based on the Exponentially Weighted Moving Average (EWMA) chart [15].

III. METHODOLOGY

This paper proposes a method to detect changes in limb capability. To emulate the user experience during robotic rehabilitation, a path-following exercise was conducted. The nominal movements are extracted using machine learning, allowing the use of statistical methods to detect change.

This paper models the scenario described in [16], targeting the data center. To compare the proposed metric to conventional methods, the collected data sequences are classified into their class-labels and visualized.

A. Experiment Setup

In this experiment, 25 *healthy* subjects are asked to follow 4 paths using the Rethink Robotics Sawyer manipulator affixed with a handle (Figure 1). The path is displayed on a screen, indicating the relative position of the end effector.

Subjects are asked to perform path-following on 4 different paths. In total, 6 unique paths were used: an original set of 4 paths for training the model, 1 modified path (acting as the shifting class), and 1 path, which is novel to the classifier.

One hundred sequences were collected in total. 60 were used for training, 20 as the holdout set during model validation, and 20 for testing (inclusive of the modified *and* novel paths).

Sensible data collection is a key factor to ease the adoption of robotic systems for therapists. Conventional assessment systems record neuro-physiological data from the user which
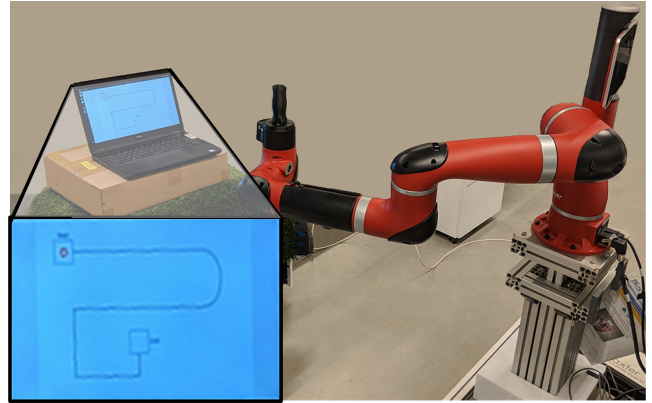

Fig. 1: The setup for the Sawyer manipulator constrained to the X-Y plane with the target path shown on the screen.

are challenging to collect. In this experiment, only the endpoint data from the manipulator was recorded, expediting the data collection process. Updating the model of a user's limb capability is computationally expensive. However, this is assumed to be largely irrelevant since the process is expected to occur between clinical visits.

B. Data Processing

The data collected is temporal due to different velocities at which subjects followed the path. A common method to align time-series data sets is Dynamic Time Warping (DTW) [17]. For two time-series data sets, Q and C, DTW provides a metric distance and a warping matrix to best align the two data sets temporally.

The minimization of the Euclidean distance from one point in data set A to the corresponding point in data set B is used for this experiment. Using DTW for multi-dimensional time-series data set is non-trivial but can be naïvely done by taking the DTW sum of each dimension independently for M number of dimensions.

However, since the collected data is embedded in a multivariate environment (2-D planar points), a dimensional-dependent DTW algorithm [18] is used to warp the data sets (Equation 1). The time-series data set are z-normalized to fit the algorithm's recommendation.

Given M dimensions and N samples in each time series:

$$DTW_{dep} = \sum_{i=1}^{M} \sum_{j=1}^{N} (q_{i,j}, c_{i,j})^2. \tag{1}$$

A covariance matrix of the DTW distances between the training data sets was created and averaged to determine the reference data set for each path during DTW. The warped paths are then temporally standardized and used to train a Gaussian Process (GP) model using the GPML package [19] to generalize the optimal path.

C. Gaussian Process

The non-linear target function f is assumed as:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{Y}$ is the training targets, $\mathbf{X}$ is the training inputs, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma_\epsilon)$ is the observation noise, which is assumed to be distributed normally with $\Sigma_\epsilon = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_{d_y}^2\}$.

To learn such a relationship from data, a GP is fitted with the prior for each dimension of y as follows:

$$\boldsymbol{y}_a \sim \mathcal{GP}(0, K_a), \tag{3}$$

for $a = 1, 2, 3, \ldots, d_y$. The observation noise is $\epsilon_a \sim \mathcal{N}(0, \sigma_a^2)$.

The kernel used to calculate each covariance element is the squared exponential as follows:

$$k_a(\boldsymbol{x}, \boldsymbol{x}') = \sigma_a^2 \exp\left(-\frac{1}{2l_a^2}||\boldsymbol{x} - \boldsymbol{x}'||^2\right), \tag{4}$$

where σ_a^2 is the variance of $\boldsymbol{f}_a$, and l_a is the scaling factor to normalize the distance between x and x'.

As the maximum a-posteriori (MAP) estimate of the kernel function hyperparameter set θ_a occurs where $p(\mathrm{Y}_a \mid \mathrm{X}, \theta_a)$ is greatest, θ_a is optimized as:

$$\log p(\boldsymbol{Y}_a|\boldsymbol{X}, \theta_a) = -\frac{1}{2}\mathcal{D} - \frac{1}{2}\log|K_a| - \frac{N}{2}\log 2\pi, \tag{5}$$

where $\mathcal{D} = \boldsymbol{Y}_a^T K_a^{-1} \boldsymbol{Y}_a$, and N is the number of training data samples.

For regression inference of a Gaussian Process, the joint probability $P(\boldsymbol{Y}_a, \boldsymbol{y}_a^{*T})$, inferring $\boldsymbol{y}_a^*$ at $\boldsymbol{x}^*$, is modelled as a multi-variate Gaussian distribution:

$$P\left(\begin{bmatrix}\boldsymbol{Y}_a \\ \boldsymbol{y}_a^*\end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix}K_a & K_{a*}^T \\ K_{a*} & K_{a**}\end{bmatrix}\right)$$

where $K_{a*} = [k_a(\boldsymbol{x}^*, \boldsymbol{X}_1), k_a(\boldsymbol{x}^*, \boldsymbol{X}_2), \ldots, k_a(\boldsymbol{x}^*, \boldsymbol{X}_N)]$ and $K_{a**} = k_a(\boldsymbol{x}^*, \boldsymbol{x}^*)$, respectively. The conditional distribution, $p(\boldsymbol{y}_a^*|\boldsymbol{Y}_a) = \mathcal{N}(K_{a*}K_a^{-1}\boldsymbol{Y}_a, K_{a**} - K_* K^{-1} K_{a*}^T)$ can be derived from this multivariate Gaussian.

D. Gaussian Process Classification

The joint distribution for Gaussian Process classification is non-Gaussian which is analytically intractable. Thus, approximations or Monte Carlo sampling methods are required.

For binary Gaussian Process Classification, a Laplace Approximation to the posterior is used [20]. A second order Taylor Series expansion around the maximum a posteriori provides the approximation as follows:

$$q(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}}, A^{-1}), \tag{6}$$

where $\hat{\boldsymbol{f}} = \text{argmax}_{\boldsymbol{f}}\, p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y})$, and A is the Hessian of the negative log posterior at that point $(-\nabla\nabla\log p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y})|_{\boldsymbol{f}=\hat{\boldsymbol{f}}})$.

Expanding the Laplace Approximation to multi-class classification extends the binary algorithm by utilizing the softmax output $\boldsymbol{\pi}$ during training (Equation 7). The class probabilities are calculated in a one vs. rest fashion:

$$\boldsymbol{\pi}_i^c = \frac{\exp y_c^i}{\sum_{c'}\exp y_{c'}^i}, \tag{7}$$

for $i = 1, 2, \ldots, n$ for n training samples, and $c = 1, 2, \ldots, C$ for C classes.

Similar to the regression inference of GP, calculating the approximate joint posterior distribution will allow for the predictive mean for class c as follows:

$$\mathbb{E}_q[f^c(x_*)|X, \boldsymbol{y}, x_*] = k_c(x_*)^T K_c^{-1}\hat{\boldsymbol{f}}^c, \tag{8}$$

where $k_c(x_*)$ is the covariances between the test point and each training point of the cth covariance function, and $\hat{\boldsymbol{f}}^c$ is the subvector of $\hat{\boldsymbol{f}}$ in class c.

Unlike binary classification, a simple mean threshold does not take variability between classes into account. The most common way to overcome this is through Monte Carlo sampling of the Gaussian Laplace Approximation, $q(\boldsymbol{f}_*|y)$, performing the softmax, and averaging the probabilities.
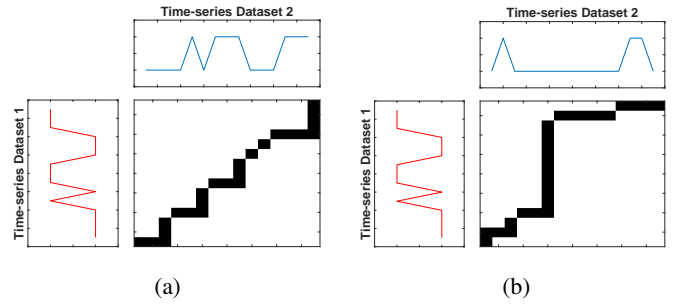

(a) (b)

Fig. 2: A comparison of the warping path between two time-series data sets: (a) the two near-identical data sets produce a near-diagonal warping path representing a clear 1-to-1 match between the data samples; and (b) the two different data sets will cause long sequences in the warping path indicating inability to find data sample matches.

E. Change Detection

Given two multi-dimensional data sets, conventional comparison methods compute a single scalar metric of similarity between the two data sets. Common methods include the paired t-test, Root Mean Square Error (RMSE), and the Autoregressive Moving Average. However, these methods do not indicate *when* and *where* the divergence occurs.

DTW provides a distance metric and a warping path, describing how data samples are matched, using one data set as a reference. Thus, change detection is possible using the longest sequence in the warping matrix. A long sequence, as shown in Figure 2, indicates a region of change.

The Longest Warping Subsequence (LWSS) is the ratio of the longest sequence normalized to the data set sizes:

$$LWSS(w, z) = \frac{n \times L_{max}(w, z)}{m \times \text{argmax}\, size(w, z)}, \tag{9}$$

where w is a $n \times D$ data set, z is a $m \times D$ data set, and L_{max} is the longest sequence within the warping matrix.

In scenarios where two data sets are discordant, ascertaining the magnitude of the difference is challenging with the warping path. Thus, RMSE can be used as a complementary measure of the difference between the two data sets.

When a changed path is detected, a buffer, W is initialized to accumulate subsequent test paths. The buffer looks for

changes in a window of test paths, using the RMSE and the LWSS ratio as follows:

$$\boldsymbol{g}(W,c) = \sum_{a=1}^{A} \left\{ RMSE(w_a, z_c) \times LWSS(w_a, z_c) \right\}, \quad (10)$$

where A is the number of accumulated test paths and z is the model of the trained path for class, c.

When the joint RMSE-LWSS score, $\boldsymbol{g}(W,c)$ is consistently less than a specified threshold, the changes are assumed to be a deviation from the learned model. Thus, the path's model can be updated to incorporate the changes.

Conversely, when the RMSE-LWSS score for all classes is over an empirically specified threshold, the path is assumed to be novel, allowing for a new model to be learned and incorporated for future classifications.

IV. RESULTS

A. Model Validation

The paths performed by the subjects are indicated as Path 1-4. The 15-subject training data set produced a GP model for each of the original 4 paths (with one seen in Figure 3). The 5-subject holdout set used the original paths while the 5-subject test set was obtained from 2 original paths, 1 modified path, and 1 novel path.

All data sets were temporally normalized to 200 samples through interpolation. Each of the 4 GP models in the training data set was then trained on a GP Laplace Approximation classifier in a one vs. rest fashion.

To avoid class imbalance, training samples were reduced to keep a 1:1 ratio of positive and negative training samples. The holdout set was tested against the classification models built to confirm the validity of the trained GP classifiers.

During data processing, one of the data sets in the holdout set was found to be erroneous, since it only consisted of half a second's worth of data, and was thus discarded. The results
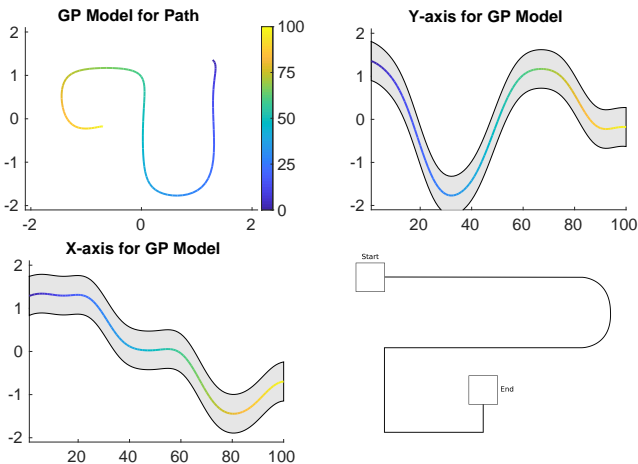

Fig. 3: The time-series GP model built for classification based on the data sets for the path shown in the bottom right.

from the holdout set are shown in Table I along with the calculated RMSE-LWSS score for comparison.

TABLE I: The posterior probabilities from the GP classification and RMSE-LWSS score from the holdout set. Note that the outlier sample from path 1 has been excluded due to human error during data collection.

Path	Mean Class Posterior Probabilities			
No.	Class 1	Class 2	Class 3	Class 4
1	**0.6873**	0.1644	0.2563	0.2706
2	0.2742	**0.7942**	0.1830	0.2225
3	0.2809	0.1795	**0.7283**	0.3063
4	0.3736	0.2978	0.3371	**0.7172**
Mean RMSE-LWSS Score				
1	**0.0019**	0.2378	0.0957	0.0682
2	0.5341	**0.0036**	0.1451	0.0369
3	0.1162	0.1209	**0.0014**	0.0162
4	0.1426	0.0534	0.0164	**0.0018**

B. Novel and Drifting Class Detection

In the 5-subject test data set, Paths 1 & 2 are identical to the learned models (Class 1 & 2), Path 3 was an amended path of Class 3, and Path 4 was a novel path which has not been seen by the GP classifiers.

The posterior probabilities from the GP Laplace Approximation classification and the RMSE-LWSS scores can be seen in Table II with the best results highlighted in bold. The component measures of mean RMSE and mean LWSS score is also presented in the table.

TABLE II: The mean posterior probabilities, RMSE-LWSS score, RMSE, and LWSS score of each class/path in the testing data set.

Path	Mean Class Posterior Probabilities			
No.	Class 1	Class 2	Class 3	Class 4
1	**0.7974**	0.2288	0.3524	0.3573
2	0.2024	**0.8117**	0.2591	0.2971
3	0.2398	0.1848	**0.5093**	0.4052
4	0.2558	0.2678	0.4564	**0.4926**
Mean RMSE-LWSS Score				
1	**0.0042**	0.2709	0.0211	0.1079
2	0.1143	**0.0030**	0.1331	0.2117
3	0.0416	0.1525	**0.0230**	0.0343
4	**0.0177**	0.0477	0.0511	0.0681
Mean RMSE				
1	0.1641	1.0022	**0.1370**	0.4282
2	0.8564	**0.1564**	0.9740	1.2646
3	0.2143	1.1225	**0.1776**	0.2784
4	**0.1144**	0.8973	0.2803	0.4128
Mean LWSS				
1	**0.0249**	0.2716	0.1801	0.2458
2	0.1333	**0.0179**	0.1363	0.1672
3	0.1930	0.1363	0.1303	**0.1154**
4	0.1552	**0.0527**	0.1841	0.1622

V. DISCUSSION

A. Classification vs. Calculation

From the holdout set in Table I, the RMSE-LWSS scores supported the posterior class probabilities from the GP classification. We highlight that the RMSE-LWSS score is a viable alternative to statistical methods for change detection.

In this experiment, a threshold RMSE-LWSS score of 0.01 was sufficient to classify the paths. While this threshold is heuristically obtained, the reduced complexity allows for online classification and threshold tuning, and even the possibility of real-time classification.

The mean class posterior probabilities in Table II shows the classifier's capability to identify incorrect paths as seen by the low probabilities for Path 3 & 4. However, there is no discriminating factor between the shifted and the novel path.

While the RMSE-LWSS score provides the same capability as that of the GP classifier, a cursory glance at the individual measures of RMSE and the LWSS score highlights a distinction. The modified path is distinguishable by the combination of low RMSE and a high RMSE-LWSS score. The novel path can also be identified by both a high RMSE-LWSS score and a high LWSS score.

Thus, using a combination of the joint score and its sub-scores, the proposed method can detect and differentiate between a shifted path and a novel path. This provides an advantage over conventional comparison methods.

B. Effective Data Utilization

Intuitively, it may be assumed that using *all* available data would improve results. However, with the abundance of data generated from robotic systems, utilizing all available data does not add insight into the results obtained from less complex measures. Furthermore, with an increasing amount of data to analyze and process, computational cost will need to be considered for the application.

To test this hypothesis, another set of binary GP classification models were built using position *and* force data. The 5-subject test set was then fed into the GP classifiers, with the results shown in Table III.

The results indicate that the inclusion of force data did mildly improve the classifier's confidence for known paths. However, the inclusion was detrimental to the classifier's confidence of shifted and novel paths. The inclusion also provides no new insight when compared against results from the position only GP classifiers.

C. Data Representation

The experiments conducted followed author recommendations for the dimension-dependent DTW [18], Z-normalizing all data sequences. To determine whether data representation will affect the use of RMSE-LWSS, another set of GP classifiers were built, validated, and tested using raw data obtained from the end effector.

Classification results from using raw data shows similar behaviour with the Z-normalized experiments. Including all available raw data during classification, however, is seen to be even more detrimental to classifier confidence and results as shown by the results for Path 4 where the posterior probabilities for all 4 classes converge towards 50%.

A closer inspection into visualizing the two representations of the paths in Figure 4c and 4d highlights the differences between them. It can be seen that using raw data can improve

TABLE III: The GP Laplace Approximation classification posterior probabilities for Z-normalized and raw test set, along with position only data and force-inclusive position data.

Path No.	Mean Class Posterior Probabilities			
	Class 1	Class 2	Class 3	Class 4
Z-normalized Position Data				
1	**0.7974**	0.2288	0.3524	0.3573
2	0.2024	**0.8117**	0.2591	0.2971
3	0.2398	0.1848	**0.5093**	0.4052
4	0.2558	0.2678	0.4564	**0.4926**
Z-normalized Position & Force Data				
1	**0.8694**	0.2128	0.3991	0.3231
2	0.2113	**0.8810**	0.3766	0.3084
3	0.2834	0.1436	0.4431	**0.5137**
4	0.3054	0.2787	**0.4241**	0.4052
Raw Position Data				
1	**0.7733**	0.2002	0.3600	0.4138
2	0.2276	**0.8152**	0.2473	0.3450
3	0.2815	0.2715	**0.5570**	0.4429
4	0.3191	0.3784	**0.4393**	0.4349
Raw Position & Force Data				
1	**0.8499**	0.5000	0.3338	0.2831
2	0.4064	**0.8554**	0.2978	0.2128
3	**0.5881**	0.4366	0.5828	0.4934
4	0.5478	0.5554	0.5490	0.5568

confidence when a correct path is presented. However, the loss of size-invariance and robustness outweighs any confidence gains during classification.

In the context of progress assessment, the normalization of data will provide a more agnostic approach towards the integration of RMSE-LWSS score in robotic systems for rehabilitation and recovery.

VI. CONCLUSION

The rising trend to incorporate robotic systems in rehabilitation and recovery necessitates the current need to overcome adoption barriers for recovery assessment.

One such area is the efficient use of data to find insights into a patient's limb capability. The conventional approach is to use *all* available data to provide assessment and advice. However, the presented results show that efficient utilization of data can be posited to provide similar, if not the same, amount of meaningful information to aid therapists.

A method to detect change using a novel metric (RMSE-LWSS) on incoming data has been proposed. Experiments, involving a path-following exercise, were conducted to test its efficacy for classifying the paths. Further experiments demonstrate that a combination of the RMSE-LWSS score and its sub-scores can distinguish outliers from systemic change.

The ability to identify modified, novel, and prescribed paths provides supplementary information for therapists during progress assessment. Future works will include different ways to utilize the RMSE-LWSS score for current assessment protocols, and incorporate temporal data to further improve the metric. Efficient data utilization will complement current works related to Assistance-As-Needed (AAN) frameworks [21], with the potential incorporation of musculo-skeletal models for recovery assessment [22].
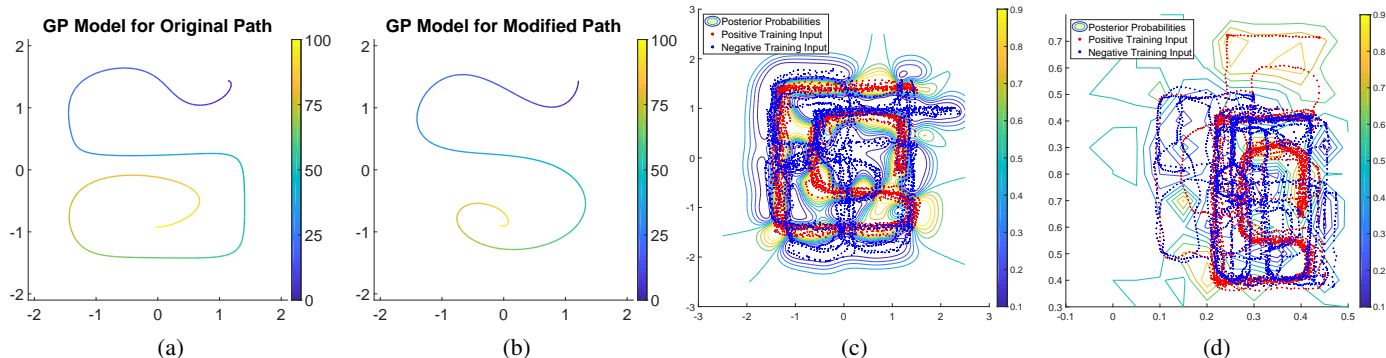
Fig. 4: A comparison of (a) the original trained model and; (b) the modified path emulating a patient's inability to follow the path during recovery exercises. (c-d) Contour plot comparisons of how data representation can affect the posterior probabilities from the learning model: (c) using z-normalized data; and (d) using raw data.

REFERENCES

[1] T. Platz, C. Pinkowski, F. van Wijck, I.-H. Kim, P. di Bella, and G. Johnson, "Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study," *Clinical Rehabilitation*, vol. 19, no. 4, pp. 404–411, 6 2005.

[2] H. I. Krebs, M. L. Aisen, B. T. Volpe, and N. Hogan, "Quantization of continuous arm movements in humans with brain injury," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4645–9, 4 1999.

[3] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance," *Scandinavian journal of rehabilitation medicine*, vol. 7, no. 1, pp. 13–31, 1975.

[4] F. I. Mahoney and D. W. Barthel, "Functional evaluation: The Barthel Index: A simple index of independence useful in scoring improvement in the rehabilitation of the chronically ill," *Maryland State Medical Journal*, vol. 14, pp. 61–65, 1965.

[5] J. Brackenridge, L. V. Bradnam, S. Lennon, J. J. Costi, and D. A. Hobbs, "A Review of Rehabilitation Devices to Promote Upper Limb Function Following Stroke," *Neuroscience and Biomedical Engineering*, vol. 4, no. 1, pp. 25–42, 5 2016.

[6] A. Schwarz, C. M. Kanzler, O. Lambercy, A. R. Luft, and J. M. Veerbeek, "Systematic Review on Kinematic Assessments of Upper Limb Movements After Stroke," *Stroke*, vol. 50, no. 3, pp. 718–727, 3 2019.

[7] F. J. Valero-Cuevas, V. Klamroth-Marganska, C. J. Winstein, and R. Riener, "Robot-assisted and conventional therapies produce distinct rehabilitative trends in stroke survivors," *Journal of NeuroEngineering and Rehabilitation*, vol. 13, no. 1, p. 92, 12 2016.

[8] N. Schweighofer, C. Wang, D. Mottet, I. Laffont, K. Bakhti, D. J. Reinkensmeyer, and O. Rémy-Néris, "Dissociating motor learning from recovery in exoskeleton training post-stroke," *Journal of NeuroEngineering and Rehabilitation*, vol. 15, no. 1, p. 89, 12 2018.

[9] M. J. Johnson, X. Feng, L. M. Johnson, and J. M. Winters, "Potential of a suite of robot/computer-assisted motivating systems for personalized, home-based, stroke rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 4, no. 1, p. 6, 12 2007.

[10] P. Maciejasz, J. Eschweiler, K. Gerlach-Hahn, A. Jansen-Troy, and S. Leonhardt, "A survey on robotic devices for upper limb rehabili-

tation," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 3, 1 2014.

[11] W. Wu, D. Wang, T. Wang, and M. Liu, "A personalized limb rehabilitation training system for stroke patients," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 12 2016, pp. 1924–1929.

[12] Y. Lai, J. Poon, G. Paul, H. Han, and T. Matsubara, "Probabilistic Pose Estimation of Deformable Linear Objects," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*. Munich, Germany: IEEE, 2018, pp. 471–476.

[13] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. MIT Press, 2009.

[14] I. Zliobaite, "How good is the Electricity benchmark for evaluating concept drift adaptation," Bournemouth University, Poole, Tech. Rep., 1 2013.

[15] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern Recognition Letters*, vol. 33, no. 2, pp. 191–198, 1 2012.

[16] Y. Lai, S. Sutjipto, M. D. Clout, M. G. Carmichael, and G. Paul, "GAVRe2: Towards Data-driven Upper-Limb Rehabilitation with Adaptive-Feedback Gamification," in *Proceedings of IEEE International Conference on Robotics and Biomimetics (ROBIO)*. Kuala Lumpur, Malaysia: IEEE, 2018, pp. 164–169.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 2 1978.

[18] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing DTW to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 1–31, 1 2017.

[19] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[20] C. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[21] M. Carmichael, "A musculoskeletal model-based Assistance-As-Needed paradigm for assistive robotics," Ph.D. dissertation, University of Technology Sydney (UTS), 2013.

[22] R. Khonasty, M. G. Carmichael, D. Liu, and K. J. Waldron, "Upper Body Pose Estimation Utilizing Kinematic Constraints from Physical Human-Robot Interaction," in *Proceedings of Australasian Conference on Robotics and Automation*, Sydney, Australia, 2017.