The EmoPain@Home Dataset: Capturing Pain Level and Activity Recognition for People with Chronic Pain in Their Homes

Temitayo Olugbade*‡, Raffaele Andrea Buono*, Kyrill Potapov*, Alex Bujorianu*§, Amanda C de C Williams*, Santiago de Ossorno Garcia*, Nicolas Gold*, Catherine Holloway*¶, Nadia Bianchi-Berthouze*†
*University College London, UK, ‡University of Sussex, UK, §Holistic AI, ¶Global Disability Hub, UK,
†n.berthouze@ucl.ac.uk

***Abstract*—Chronic pain is a prevalent condition where fear of movement and pain interfere with everyday functioning. Yet, there is no open body movement dataset for people with chronic pain in everyday settings. Our EmoPain@Home dataset addresses this with capture from 18 people with and without chronic pain in their homes, while they performed their routine activities. The data includes labels for pain, worry, and movement confidence continuously recorded for activity instances for the people with chronic pain. We explored baseline two-level pain detection based on this dataset and obtained 0.62 mean F1 score. However, extension of the dataset led to deterioration in performance confirming high variability in pain expressions for real world settings. We investigated baseline activity recognition for this setting as a first step in exploring the use of the activity label as contextual information for improving pain level classification performance. We obtained mean F1 score of 0.43 for 9 activity types, highlighting its feasibility. Further exploration, however, showed that data from healthy people cannot be easily leveraged for improving performance because worry and low confidence alter activity strategies for people with chronic pain. Our dataset and findings lay critical groundwork for automatic assessment of pain experience and behaviour in the wild.**

***Index Terms*—Activity recognition, affect recognition, body movement, chronic pain, confidence, dataset, worry.**

I. Introduction

AUTOMATIC assessment in the context of pain experience aims to support personalization of care, empowerment of patients, and self-management of chronic conditions through digital technology [1]–[6]. Datasets are a critical resource for such aim. Not only are they fundamental to creating machine learning models for automatic assessment, but they are additionally important for deeper understanding of support needs that can inform how the assessment technology is embedded in care. In this paper, we introduce the extended *EmoPain@Home* dataset that consists of body movement data captured during functional activities at home, from both people with and without chronic pain. Chronic pain is of particular significance due to its effect on the sense of self, engagement in valued activities, and interaction with others [7]–[10]. The *EmoPain@Home* dataset is labelled with self-reported levels of pain and related worry and confidence for the participants with chronic pain. It additionally includes labels of the activities performed, for both participants with and without chronic pain. The current paper is an extended version of our previous paper, [11]. In [11], we presented the first subset of the dataset and preliminary baseline results for pain level classification based on this data. Beyond providing a more thorough discussion of literature and findings, the extension in the current paper covers three areas: 1) inclusion of data from healthy participants, i.e. participants without chronic pain, as controls in addition to further data from some of the same participants with chronic pain; 2) extended exploration of pain level classification baselines for the dataset including use of features extracted from angles, rather than positions, of the anatomical joints captured, to understand their effect on performance; and 3) exploration of activity recognition to understand its feasibility for future use as context information for improving performance for recognition of pain levels.

The latter extension was motivated by findings in studies such as [12], [13] that highlight the role that the activity context can play in automatic assessment of pain behaviour. In the two studies, both the activity type and the presence/absence of protective behaviour were learnt using a hierarchical structure where the recognized activity class was further used as an additional feature for protective behaviour detection. Wang et al. [12] found improvement in protective behaviour detection performance with such inclusion of the activity information. An interesting finding in [13] was that the best performance for protective behaviour detection (F1 score = 0.93) across several modality fusion algorithms was obtained when performance for activity recognition was lowest (F1 score = 0.57, for 6 classes). This led the authors to theorize that the strategies used by people with chronic pain to execute feared movements (protective behaviours) deviates from intuition or natural behaviour and so undermines activity recognition but is in itself useful information for protective behaviour detection. These findings suggest value in exploring activity recognition for movements of people with chronic pain in real everyday settings, i.e. beyond the instructed settings in [12], [13].

In summary, this paper makes the following contributions:

1) The extended EmoPain@Home dataset, which is the first open dataset on functional activities performed by people with chronic pain in real everyday settings. The

Manuscript received xxx; revised xxx.

full dataset consists of motion capture data from 9 people with chronic pain and 9 matched healthy people, captured during normal instances of their home activities. The data for the people with chronic pain further includes self-reports of pain, pain-related worry, and movement-related confidence recorded every minute.
2) Investigation of the feasibility of recognizing pain levels in real world settings as a baseline for the research community. Movement is generally more complex in such settings and so it presents a higher level of challenge than datasets that have been used in previous studies. This is further compounded by differences between home layouts that will introduce variability across people in the execution of the same type of activity and in the strategies used to cope with pain experience.
3) Analysis of the feasibility of automatic recognition of activities of people with chronic pain in real world settings. Since data is typically scarce for such settings, especially for people with chronic pain, we explore the use of data from healthy people to improve performance.

II. Previous Work

A. Datasets for Automatic Detection of Pain

Datasets for automatic recognition of pain and related affect exist for a variety of pain expression modalities, as can be seen in Table I which shows a representative collection. The facial modality of pain expression has been the most widely captured. Findings in [36] indeed highlight the relevance of this modality, showing statistically significant difference in the activation of facial actions between the use of the pain-affected arm of people with shoulder pain and the use of their non-affected arm. A few datasets contain data on vocal/paraverbal expressions, which have also been shown to be valuable for capturing pain experience. For example, Belleni et al. [37] found significant higher base frequencies, insistence and periodicity, as well as duration of maximum pressure in infant cries where there were other expressions of pain, compared to cries associated with mild or no non-vocal expression. Some of the existing datasets include physiological signals that are known to be related to pain experience, e.g. skin conductance level, for which Nickel et al. [38] found significant effect of heat stimuli intensity (higher than subjective pain threshold). Another modality represented in previous datasets is body movement data, which has also been linked to pain expression [39], e.g. lower range of trunk motion in participants with chronic low back pain experiencing high level pain [10].

Although a dataset with multiple nonverbal measures of pain is ideal [4], logistics can be a constraining factor for real world settings. For example, facial capture is more practical for sedentary contexts (e.g. [19], [21]–[23], [29], [33], [34]) or when mobility is limited to a single space (e.g. [20], [27]). The purpose of assessment is also an important consideration. Findings in [40], for instance, suggest that body movement may be particularly valuable when the goal of assessment includes judgement of task demand and coping strategy during physical activity. Existing datasets that include body movement data for people with chronic pain are limited to constrained movement in artificial settings [14], [16], [18], [20], [27], [31]. Thus, there is need for a new dataset of pain experience during everyday activity of people with chronic pain.

We address this gap with our EmoPain@Home dataset of body movement data captured in natural activities in people's homes. It consists of data for both participants with chronic pain and healthy participants. We further investigate the feasibility of pain level classification based on this new dataset that, compared with those used in the studies mentioned above, was captured in the more complex settings of people's homes. The implication of routine activities captured at home is that unlike instructed activities in lab or similar settings:

- they have higher utility for participants, and so are more likely to be performed rather than avoided;
- they are longer in duration, with each activity type involving a wider range of subgoals and motion primitives;
- participants have adaptive strategies that are more applicable at home (personal environment), e.g. sitting for washing up as a strategy for coping with the challenge of this activity, and these strategies can dampen or confound expressions of pain that are salient in other settings.

B. Activity Recognition with Movement Disorders

Activity recognition is an advanced area of research with performance greater than 0.90 accuracy for most benchmark datasets [41]. However, the majority of these datasets only represent data from healthy people (typically young adults), to the exclusion of data from people with movement disorders [42]. There have been very few studies that have investigated the possibility of activity recognition for people with conditions that can affect mobilty or execution of movement.

One of these rare studies is [43] on walking actions in everyday activities of stroke survivors with hemiparesis. Their machine learning model was able to identify all walking periods in both this group of participants and healthy control participants. Zhan et al. [44] who considered a wider variety of conditions (including Parkinson's disease and fracture) and multiple activity types beyond walking (e.g. stair climbing, sit-to-stand, lying down) obtained 0.75 accuracy for 14 instructed activities types. In [45] with participants with cerebral palsy, average F1 score greater than 0.85 was achieved for four categories of instructed activities (including lying down, seated writing, walking). However, their comfortable walking category was challenging for the model to recognize, with true positive rate of 0.66 compared with 0.86-0.97 for the others. In [12], [13], average F1 scores of 0.81 and 0.78 respectively were obtained for classification of 6 instructed activity types (e.g. sit-to-stand, forward reach) for participants with chronic pain. The findings of these studies point to possibility of activity recognition in people with movement disorders.

However, studies where performance for such participant groups is compared with that for healthy participants suggest higher difficulty in recognizing activities where there is a movement disorder. For example, in [46], considerably better recognition of sitting and standing activities was found for the activities of healthy participants compared with stroke survivors, although difference was only statistically significant

TABLE I
EXISTING PAIN RECOGNITION DATASETS

Dataset	Year	Pain	Participant (Number)	Context	Modality (Sensor)
Gioftsos and Grieve [14]	1996	chronic	healthy (14), healthy with past acute pain (12), back pain (10)	instructed exercise movements in the lab	body movement (goniometer), ground force reaction (force plate)
Bishop et al. [15]	1997	acute	healthy (103), back pain (80)	constrained exercise movements in the lab	spine movement (goniometer)
Dickey et al. [16]	2002	chronic	low back pain (9)	instructed exercise movements in the lab	spine movement (marker-based optical)
Brahnam et al. [17]	2006	puncture	infants (26)	hospital neonatal unit	facial (photograph)
Levinger and Gilleard [18]	2007	chronic	healthy females (14), knee pain females (13)	instructed walking in the lab	leg movement (camera, light gates), ground reaction force (force plate)
Hi4D-ADSIP [19]	2011	acted	healthy (80)	seated	facial (RGB camera)
UNBC-McMaster Shoulder Pain Expression [20]	2011	acute, chronic	shoulder pain (129)	instructed exercise movements standing or laid down in the lab	facial (RGB camera)
BioVid Heat Pain [21]	2013	heat	healthy (90)	seated in lab settings	physiological (ECG, EDA, EEG, EOG, EMG), facial & upper body (RGB&D cameras)
BP4D-Spontaneous [22]	2014	cold	healthy (41)	seated in lab settings	facial & head (RGB and grayscale stereo cameras)
Rivas et al. [23]	2015	acute	stroke patients (2)	seated exergaming	hand movement & pressure (game controller), facial (camera)
Infant Cry Sounds [24]	2015	acute	infants (33)	hospital visit	vocal (microphone)
Zhang et al. [25]	2016	cold	healthy (140)	seated in lab settings	physiological (EDA, blood pressure, heart rate, respiration, thermal camera), facial & head (stereo camera)
Triage Pain-Level Multimodal [26]	2016	acute	emergency room patients (182)	triaging in hospital emergency unit	physiological (cardiac, blood pressure), vocal (microphone), facial (camera)
EmoPain [27]	2016	chronic	healthy (28), low back pain (22)	instructed exercise movements in the lab	physiological (EMG), facial (camera), body movement (inertial)
SenseEmotion [28]	2017	heat	healthy (45)	seated looking at affective images augmented with sound	physiological (ECG, EDA, EMG, respiration), vocal (microphone), facial (camera), body movement (markerless)
Multimodal Intensity Pain (MIntPAIN) [29]	2018	electrical	healthy (20)	seated in lab settings	physiological (EMG), facial (RGBD, thermal cameras)
Ubi-EmoPain [10], [30]	2018	chronic	low back pain (12)	instructed exercise & physical task in the lab	physiological (EMG), body movement (inertial)
Hu et al. [31]	2018	chronic	healthy (22), low back pain (22)	instructed standing in the lab	physiological (EMG), ground reaction force (force plate), spine movement (electromagnetic)
Clinical Valid Pain [32]	2018	acute	emergency room patients (140)	hospital emergency unit visit	blood (cyclooxygenase-2, inducible nitric oxide synthase), facial & head (RGB&D cameras)
X-ITE Pain [33]	2019	electrical, heat	healthy (134)	laid down in the lab	physiological (EDA, ECG, EMG, thermal camera), vocal (microphone), facial & body (RGB&D cameras)
Intelligent Sight & Sound Chronic Cancer Pain [34]	2021	cancer	cancer patients (29)	verbal tasks	vocal (microphone), facial (camera)
iCOPEvid [35]	2019	puncture	infants (49)	neonatal hospital unit	facial (camera)
EmoPain@Home [11]	2022	chronic	low back pain (9)	functional home activities	body movement (inertial)
EmoPain@Home extended (current paper)	***2023***	***chronic***	***healthy (9), low back pain (9*)***	***functional home activities***	***body movement (inertial)***

ECG - Electrocardiogram, EDA - Electrodermal activity, EEG - Electroencephalography, EMG - Electromyography, EOG - Electrooculography
*In addition to the data from the 9 participants in [11], additional data was collected from 3 of these participants in the current dataset extension.

for the standing activity. Similarly, in [47], slightly lower recognition accuracy was found for participants with Parkinson's disease compared with healthy participants. When data from healthy participants alone was used to train the model tested on data from participants with Parkinson's disease, accuracy reduced considerably, 0.60 (for 5 instructed activity types and settings) compared to 0.75.

We extend these investigations with our work on activity recognition based on the new EmoPain@Home dataset that is more complex than those considered in the above studies. First, the EmoPain@Home activities, e.g. washing up, vacuuming, are of a higher level of abstraction than sit-to-stand, walking, forward reach. For example, an instance of vacuuming activity could include walking and forward reach simultaneously. Recognition of such composite activities remains a challenge even in the larger human activity recognition area [41], [48]. Although focus on recognition of the lower level actions or gestures could remove this difficulty, knowledge of the wider context of an action/gesture could be useful. For instance, forward reach movements during vacuuming may be different

TABLE II
NORMAL EVERYDAY ACTIVITIES CAPTURED IN THE EMOPAIN@HOME DATASET (SELF-SELECTED BY PARTICIPANTS)

Activity (Number of instances)	**Participant group (Total number of participants)** [P=Pain, H=Healthy]
Bathroom cleaning (9)	P-C, P-NC, H* (6)
Changing bedsheets (6)	P-C**, H* (6)
Cleaning parrot cage (1)	P-NC* (1)
Cleaning windows (11)	P-C, H* (9)
Dusting (1)	P-NC** (1)
Dusting-car (1)	P-NC (1)
Filing documents (1)	P-NC (1)
Hanging clothes to dry (1)	P-C (1)
Ironing (1)	P-C (1)
Loading dishwasher (2)	P-NC, H* (2)
Loading washing machine (9)	P-C**, P-NC**, H* (8)
Organizing boxes (1)	P-NC (1)
Painting a wall (1)	P-C (1)
Painting shelves (1)	P-C (1)
Preparing meal (2)	P-C, P-NC* (2)
Sweeping (2)	P-NC (2)
Tidying up (4)	P-C**, P-NC (3)
Unloading dishwasher (3)	P-C, P-NC, H* (3)
Unloading washing machine (9)	P-C**, P-NC**, H* (7)
Vacuuming (14)	P-C**, P-NC**, H* (10)
Vacuuming-car (1)	P-C (1)
Walking exercise (2)	P-C, P-NC (2)
Walking dogs (1)	P-C (1)
Washing up (17)	P-C**, P-NC**, H* (12)
Watering garden (1)	P-C (1)
Yoga (1)	P-NC (1)

* - without the researcher present; ** - both with and without the researcher present; -C - challenging for $\geq$ 1 participant with pain; -NC - non-challenging for $\geq$ 1 participant with pain

from those performed in bathroom cleaning. Second, there are strong variations for similar activities in the EmoPain@Home dataset due to differences in home settings and in the nuance of what each activity type entails. For example, for some participants, their washing up activity included tidying up the sink area whereas it was just washing up for others. In addition, there were participants for whom vacuuming was for a single room whereas it included vacuuming the stairs for other participants. Even the same home could have differences that have implications for the strategies used to complete the activity, e.g. cleaning windows well within or out of reach, changing sheets for beds set against the wall on one side or not at all on either side, walking with or without a dog.

III. THE EMOPAIN@HOME DATASET

Nine participants (5 female, 4 male) self-identified as living with chronic musculoskeletal pain involving the lower back area took part in the study. They were recruited using adverts on social media as well as by directly contacting community pain support groups across the UK. Four of them specified sciatica as their pain condition; two specified chronic pain resulting from an old spinal injury; and three specified other forms of chronic pain. The participants were between 27 and 59 years old (mean=45, standard deviation=12) and were in the UK at the time of the study (March 2021 - June 2022). We further recruited nine healthy participants (4 female, 5 male; ages 30 to 72 years) as matched controls.

Body movement data was captured using wearable inertia sensor units (Notch [49]) that record 3D joint positions and angles. To limit burden on participants, especially those with chronic pain, only 6 (out of 18) sensor units were used. Findings from pilot tests suggest that sensor attachment time and possibility of technical issues increase with number of units. The discussion section further highlights some of the challenges for people with chronic pain in using such sensors. We captured data for the right elbow and wrist, mid spine, hip, and right knee and ankle (see Figure 3-Left). Findings in previous work [10], [12], [30] suggest that data from only one side of the body can be informative for automatic assessment.

Our studies were approved by the local research ethics committee (reference 5095/001) and all participants gave informed consent for collection and processing of their data as well as for sharing pseudonymized data with the research community. Participants were reimbursed for their time at £10 per hour.

A. Data Capture Settings

Data was captured in the context of everyday physical functioning at home (see Table II) and by the participant themselves, with the sensor system sent to participants by courier or delivery by the researcher. Participants were trained (remotely by the researcher) to use the sensors themselves. For three of the participants, additional data was captured by the researcher physically present in the participant's home.

Rather than recreating activities for the purpose of the study, the participants performed tasks that they needed to do in their own homes. In order to have a good representation of pain experience across the range of these activities for each participant with chronic pain, they were asked to include activities that they usually found particularly challenging as well as those that they did not find challenging. This group of participants was asked to fill in a diary over the days before the first data capture session, to identify these activities beforehand. Data capture sessions were then arranged for the days when the participant planned to engage in the noted activities. Within each session, the participant engaged in a number of the self-selected activities. In order to enable recording of end-of-session interviews as well as to facilitate continuous capture of self-report based on experience sampling, the researcher was present remotely (i.e. via video-conferencing) during the sessions. Three participants took part in further sessions without the researcher present at all. As mentioned above, three participants also took part in additional sessions with the researcher physically present. Each session for a participant with chronic pain was limited to an hour a day to minimize fatigue. We additionally limited capture of each activity in a session to approximately 15 minutes and participants were encouraged to take breaks earlier if needed. These considerations were decided together with a clinician within the research team and discussed with the participant.

The healthy participants were also able to record activities relevant to them. However, to match the most common activities performed by the participants with chronic pain, they

had to choose from: washing up, bathroom cleaning, changing bedsheets, vacuuming, cleaning windows, loading/unloading washing machine, loading/unloading dishwasher. Similar to participants with chronic pain, healthy participants also captured data during their normal completion of the selected activities as part of their routine. The healthy participants could record data whenever was most convenient and for as long as needed for the given activity. To support longer capture, data for the healthy participants was captured at 10Hz instead of the 40Hz used for participants with chronic pain.

There were altogether 103 activity instances across 26 different activity types (see Table II) for all 18 participants. To facilitate[1] activity annotation for the dataset, participants were instructed to start sensor data capture when ready to begin the activity they wanted to record and stop it when concluded. Thus, each activity instance data was pre-segmented. The participant noted the activity type before or after the sensor data capture. This provided activity labels for each activity instance recorded. Participants with chronic pain performed 2 to 17 activity instances across multiple days (median=6), while healthy participants performed 3 to 6 instances (median=4).

B. Labelling of Pain, Worry, and Confidence

In activity instances where the researcher was present, verbal self-report of pain, worry about pain, and confidence about being able to perform the rest of the activity at every minute was recorded from participants with chronic pain.

Pain and worry were assessed on a numeric scale from 0 for 'none' to 10 for 'very severe'. Confidence was assessed on an ordinal scale of: no confidence, less than average confidence, average confidence, more than average confidence, and max confidence. While there is a lot of precedent for the 0-10 scale (especially for pain) [50], there is little evidence that people make up to 11 distinctions between levels of confidence, so we kept the scale for confidence simpler. The fewer distinctions for the confidence scale made it feasible to use an ordinal (rather than numeric) scale, which people prefer [51]. An additional rationale for using the ordinal scale was to help differentiate it from the pain and worry scales for which, unlike confidence, a higher value represents a more negative experience. This minimizes the cognitive effort of self-reporting the three constructs continuously during physical activity.

Participants provided the self-reports on prompts for current pain, worry, and confidence from the researcher. In order to limit disruption of the activities, the verbal prompt was shortened to "time" once the self-report constructs and procedure were extensively described to the participant. The method and frequency of self-report was based on discussion in [52] of the value of the method for both research and the participants themselves. For capture sessions where the researcher was not present, the participant recorded the pain, worry, and confidence experienced (and the type of activity) at the start and end of each activity instance in written form.

[1]Video for later annotation was difficult to capture. Available room space often limited where a camera could be placed, and the camera positions chosen by the participants were sometimes constrained in the view of the participant and the activity being performed especially for activities, e.g. vacuuming, that involve a wide variety of body poses and positions in the home space.

For the activity instances where self-report was provided during the activity, rather than only at the start and end, plot of the pain intensities across activity instance segments (Fig. 1-Top) shows that although most of the pain experiences captured were mid-level on the pain scale, higher pain intensities are well-represented. Further inspection of the distribution of pain intensities within each activity instance (see Figure 2-Top) shows that there are variations in pain intensities even within instances. The findings are similar for worry and confidence although the distribution for worry has a right skew and a mode at 2 (see Fig. 1-Middle). For confidence, there is a skew to the positive side with mode at 4 ('more than average confidence') (see Fig 1-Bottom, labels recoded to integer: 1 for 'no confidence', 5 for 'max confidence'). We found high correlation (Spearman's) between between worry and confidence ($\rho = -0.75$, $p = 1.84e-114$), worry and pain ($\rho = 0.7$, $p = 9.90e-96$), and pain and confidence ($\rho = -0.48$, $p = 1.01e-38$).

IV. Automatic Detection of Pain Levels

A. Methods

Our aim was baseline classification to understand the difficulty of automatic pain level recognition in real world settings. Thus, we used standard machine learning algorithms similar to those that have shown good performance in related studies [10], [30], and we also extracted features informed by these studies. We used EmoPain@Home data for which all 6 body joints of interest were recorded without missing data.

We computed 6 sets of features from the joint positions data, for each self-report instance ($n = 342$). These were average speed, average jerk, average energy, normalized amount of movement, minimum distance from the hand (to capture self-adaptor behaviour or rubbing of the painful region), and angular range of motion. We computed them across two different timescales for each self-report interval: the *segment timescale* defined as the one-minute window up to the given self-report point; and the *activity timescale* defined as the full length of the given activity instance up to the self-report point. This led to a total of 60 features. We further extracted the same features from joint angles data instead of from joint positions data for comparison to understand which input type was more informative. We computed the 3D joint angles data from the joint positions. The 'minimum distance to the hand' features could not be replicated for the joint angles data and so, they were not included for this data type. This led to 40 features.

For the learning algorithm, we used an ensemble of decision trees, specifically *Bagging* [53] which enables (random) selection of a subset of features to build each tree. We set the maximum number of features used to build each learner to 30% of the total number of features with replacement. We did not obtain better performance with other values. The model was evaluated using leave-one-activity-instance-out cross-validation where all segments from the same activity instance are held out in each fold. We did not test for generalization to unseen participants because of differences in the types of activities across participants. The number of trees for the model was selected with nested cross-validation.

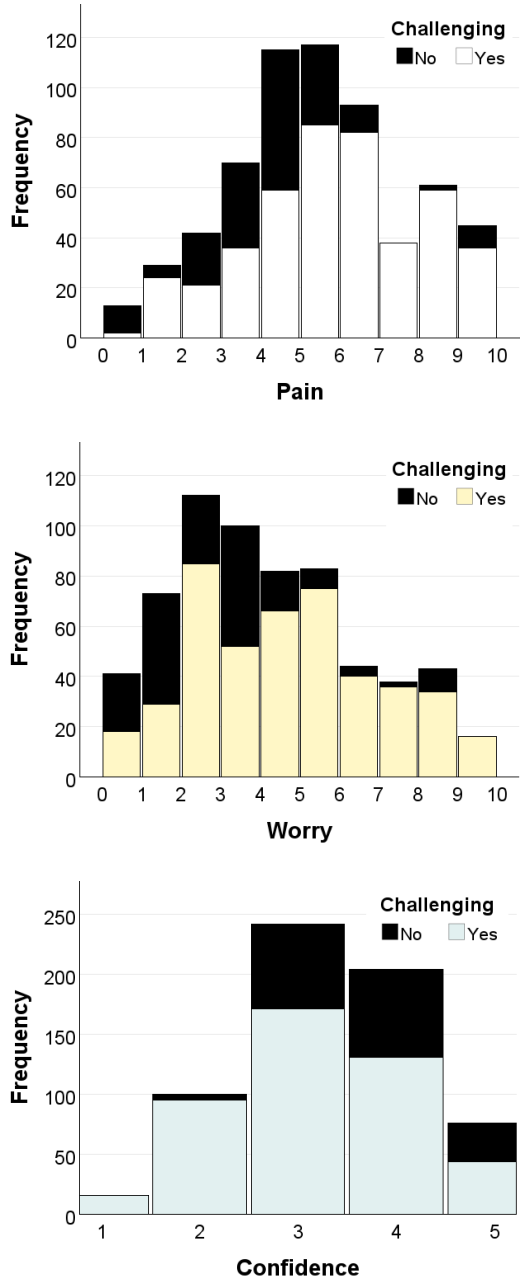

Fig. 1. Distribution of pain, worry, and confidence intensities. See Supplementary material for charts with separate panels for each challenge level.

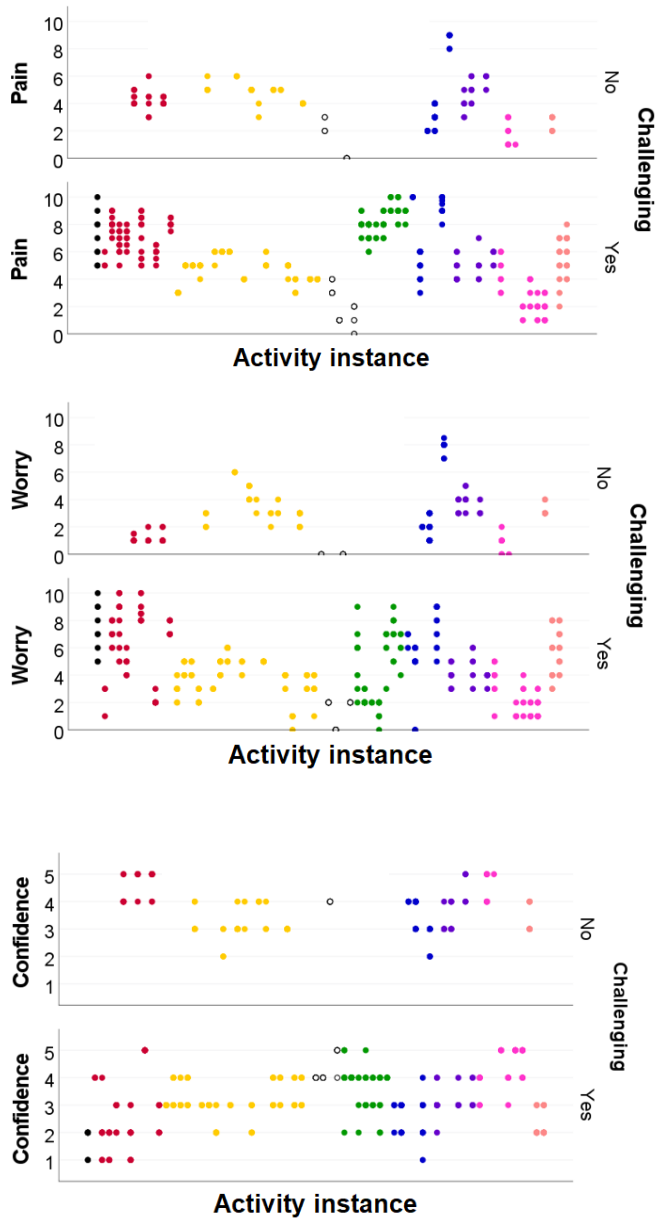

Fig. 2. Distribution of pain, worry, and confidence intensities within activity instances. Instances for the same person are plot in succession with colours used to delineate adjacent instance plots for different people.

TABLE III
PAIN LEVEL CLASSIFICATION RESULTS (F1 SCORES)

Data	Lower level pain	Higher level pain
Joint positions features	0.24	0.66
Joint angles features	0.34	0.73
Joint positions features and [11] data	0.64	0.54
Joint angles features and [11] data	0.63	0.61

B. Results

1) How difficult is pain level detection in real world settings?: We focused on two levels of pain as a starting point, lower level pain defined as pain intensity less than 5 on the pain scale and higher level pain otherwise. We compared joint angles features with joint positions features. We further compared performance with use of the EmoPain@Home data subset reported in [11] ($n = 226$) that does not include data from additional sessions with the researcher physically present.

As can be seen in Table III, the features based on joint angles data performs better than those computed from the joint positions data, for both the extended data and the [11] subset. However, performance for the extended data was poor (mean

TABLE IV
CORRELATIONS OF EXTRACTED FEATURES WITH PAIN

Variable	Joint	Timescale	Correlation	
			Data in [11]	Extended data
Speed	trunk	*Activity*	-0.15	-0.11
	thigh		-0.24*	-0.11
	upper arm		-0.14	-0.04
	lower leg		-0.25*	-0.12
	forearm		-0.06	-0.04
	hip		-0.06	0.02
Jerk	trunk	*Segment*	-0.30**	-0.13
	thigh		-0.08	0.03
	upper arm		-0.28*	-0.07
	lower leg		-0.24*	-0.13
	forearm		-0.18	0.002
	hip		-0.01	0.06
	trunk	*Activity*	-0.36**	-0.16
	thigh		-0.14	0.02
	upper arm		-0.33**	-0.12
	lower leg		-0.32**	-0.19*
	forearm		-0.20	-0.01
	hip		-0.06	0.06
Energy	trunk	*Segment*	-0.15	-0.11
	thigh		-0.09	-0.06
	upper arm		-0.13	-0.07
	lower leg		-0.11	-0.09
	forearm		-0.07	-0.02
	hip		-0.01	0.03
	trunk	*Activity*	-0.26**	-0.2*
	thigh		-0.26**	-0.13
	upper arm		-0.22	-0.14
	lower leg		-0.27**	-0.14
	forearm		-0.07	-0.05
	hip		-0.03	0.03
Amount of movement	trunk	*Activity*	-0.15	-0.07
	thigh		-0.24*	-0.11
	upper arm		-0.14	-0.04
	lower leg		-0.25*	-0.12
	forearm		-0.06	-0.04
	hip		-0.06	0.02
Minimum distance to the forearm	trunk	*Segment*	-0.27**	-0.19*
	thigh		0.04	0.03
	lower leg		0.05	0.08
	hip		-0.01	0.04
	trunk	*Activity*	-0.34**	-0.25**
	thigh		-0.16	-0.19*
	lower leg		-0.10	-0.03
	hip		-0.21	-0.04

* & ** - significant at $p < 0.05$ & $p < 0.01$ respectively with Bonferroni correction. See Supplementary material for detailed table.

F1 score = 0.54). This is interesting given that performance for the [11] subset that consists of data from the same set of participants and similar types of activities is better and well above chance-level classification, mean F1 score = 0.62. This finding suggests that the additional sessions from the three participants at the intersection between the [11] and extended data added more variation to the data and so undermines any gain that one would expect from more data. While the physical presence of the researcher is the primary difference between this additional data and the rest of the dataset, we do not believe that this contributes to such added variation as the researcher was already very familiar to the participants by these sessions and had been present remotely in other sessions included in our experiments. The variation is more likely to a result of other within- and between-subject differences.

2) What is the relevance of each feature used?: We sought to understand relations between pain and the features given the performance for the extended dataset. We focused on linear relationships and employed Spearman's correlation. We used joint positions features and compared relationships for the extended data and [11] subset. Table IV highlights statistically significant correlations. Bonferroni correction was done based on recommendation by [54] and separately for each feature category (speed, jerk, energy, etc) similar to [55].

For the [11] data subset, we found weak but statistically significant ($p < 0.05$) correlations as high as $\rho = -0.36$ (with jerk for the trunk, for the *activity timescale*). Correlation was consistently stronger for the *activity timescale* than the *segment timescale*. This suggests that non-verbal behaviour associated with pain experience may manifest over a longer timescale than the time period that the verbalization of the experience covers. Further, there was generally stronger correlation for jerk than for the others, and correlation with the range of motion (at hip and knee) was not at all statistically significant (at $p = 0.05$). This could be due to the wide variety in activity types that inherently involve different ranges of motion. Indeed, the boxplot in (Figure 3-Right) showed a limited range of motion in *washing up* particularly for the knee, while activities such as *yoga* and *vacuuming* had much larger ranges.

For the extended data, there was similarly no statistically significant correlation with the range of motion features, and correlation was also consistently stronger for the *activity timescale* than the *segment timescale*. However, with this data, the strongest correlation was lower and with minimum distance between trunk and hand ($\rho = -0.25, p = 0.0001$, with Bonferroni correction) rather than jerk, but for the same timescale. Further, only correlation for energy was consistent in significance across the data sets, although with lower strength for the extended set and significance does not remain for the thigh and lower leg with Bonferroni correction. In fact, correlation was consistently lower or less significant for this set except for minimum distance between thigh and hand.

V. AUTOMATIC RECOGNITION OF ACTIVITIES

A. Methods

We report groundwork investigation on the feasibility of automatic activity recognition in real world settings for people with chronic pain, and so we used methods resulting in good performance in related studies [12], [13]. Activity recognition could provide contextual information that encodes within-subject variations and improves pain level classification.

We used a neural network architecture based on graph convolutional layers [56] (1 layer, 26 units) for encoding spatial information and long short-term memory layers (3 layers, 24 units each) [57], [58] for encoding temporal information. The Adam optimizer was used with batch size of 150 and 100 epochs. A validation set was used to determine appropriateness of the number of epochs. The learning rate used was 5e-4 with decay of 1e-5, decided based on experimentation and using the validation set. Unlike Wang et al. [12] who used joint positions data as input into the graph convolution network, we used joint angles data instead. Findings with pilot experiments suggest

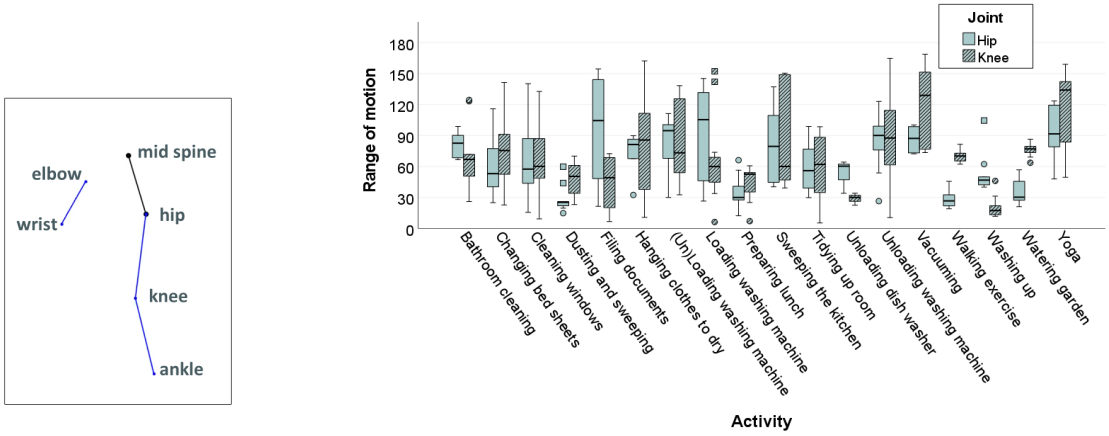

Fig. 3. Left - Anatomical joints captured in the EmoPain@Home dataset; Right - Boxplot of hip and knee ranges of motion (segment timescale).

TABLE V
ACTIVITY RECOGNITION RESULTS (F1 SCORES)

Data	Vacuuming	Washing up	Bathroom cleaning	Washing machine	Changing Bedsheet	Cleaning windows	Walking activities	Dishwasher activities	Tidying up
Pain (40Hz, 3s)	0.26	0.57	0.37	0.43	0.59	0.34	0.56	0	0.71
Pain (10Hz, 3s)	0.32	0.62	0.36	0.32	0.54	0.4	0.6	0	0.79
Pain & Healthy (10Hz, 3s)	0.24	0.54	0.38	0.31	0.38	0.4	0.16	0.33	0.80
Pain & Healthy (10Hz, 12s)	0.18	0.54	0.31	0.19	0.37	0.54	0	0	0.7
Pain & Healthy (10Hz, 30s)	0.30	0.57	0.30	0.27	0.19	0.30	0	0	0.4

that graph convolution on joint angles data leads to slightly better performance than use of joint positions input. Similar to [12], cropping and jittering data augmentations were applied to increase the training data size. We used class-balanced focal categorical cross-entropy as in [12] to address class imbalance.

As some of the 26 activity types were related, e.g. 'loading washing machine', 'unloading washing machine', 'unloading & loading washing machine', we re-structured the activity labels to put such similar activities under the same activity class. Further, we excluded activities that were only performed by one participant and only once (e.g. 'painting a wall', 'yoga'). Activity instances with data capture errors (due to sensor malfunction) were also excluded. In the end, we focused on 9 activity classes that were best represented in the extended EmoPain@Home dataset: *Vacuuming*, *Washing up*, *Bathroom cleaning*, *Washing machine activities*, *Changing bedsheets*, *Cleaning windows*, *Walking*, *Dishwasher activities*, *Tidying up*. The *Washing machine activities* class covers both loading and/or unloading of the washing machine. Similarly, *Dishwasher activities* covers both loading and/or unloading of the dishwasher, and *Walking* covers walking as exercise as well as other forms of walking, e.g. walking dogs. We segmented these activity instances using non-overlapping windows with length of 3 seconds as our default. We explored other window lengths as well. We used hold-out validation for our experiments with random stratified sampling to ensure that each activity class had similar representation in training, validation, and test sets.

B. Results

1) Is activity recognition possible for real-world movements of people with chronic pain?: For our first experiment, we focused on data from participants with chronic pain alone, using the original sampling rate of 40Hz. The results are shown in the first row of Table V. Performance across the activity classes (mean F1 score=0.43) is well above chance level recognition (F1 score = 0.11) as are F1 scores per activity type except for *Dishwasher activities* which has very limited representation in the test data, only 2 instances.

Figure 4 shows the confusion between activity types. For example, although *Washing up* has the highest true positive rate, its false positive rate is high as most of the other activities are misclassified as this activity type. False positive rate is similarly high for *Cleaning windows*. *Vacuuming* has the lowest true positive rate, and it is most often misclassified as *Washing up*. It is also sometimes misclassified as *Cleaning Windows*, *Bathroom cleaning*, and *Changing bedsheets*, but it is never misclassified as *Dishwasher activities* or *Tidying up*.

2) Does a lower time resolution affect performance?: Using a lower sampling rate, e.g. 10Hz instead of 40Hz, can be helpful for real-world deployment of movement sensors and prediction models. First, a lower sampling rate would minimize data storage space allowing saved capture for a

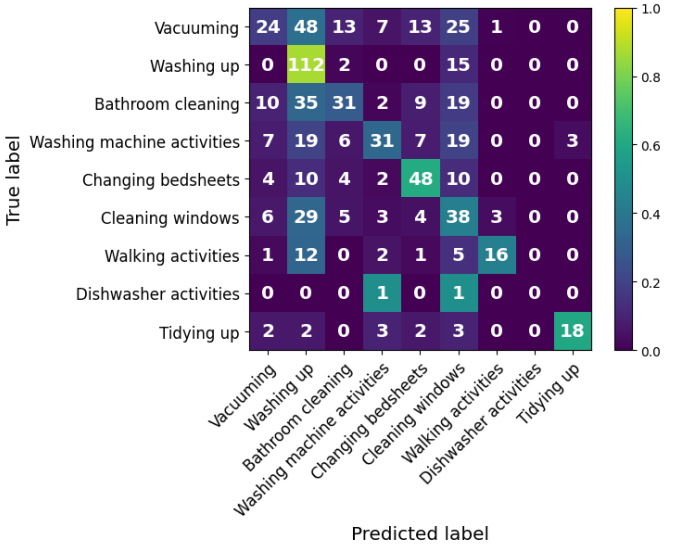

Fig. 4. Confusion matrix for activity recognition based on 3-second, 40Hz segments from participants with chronic pain. Colours represent proportion of instances per row. Numbers in the matrix represent actual number of instances.

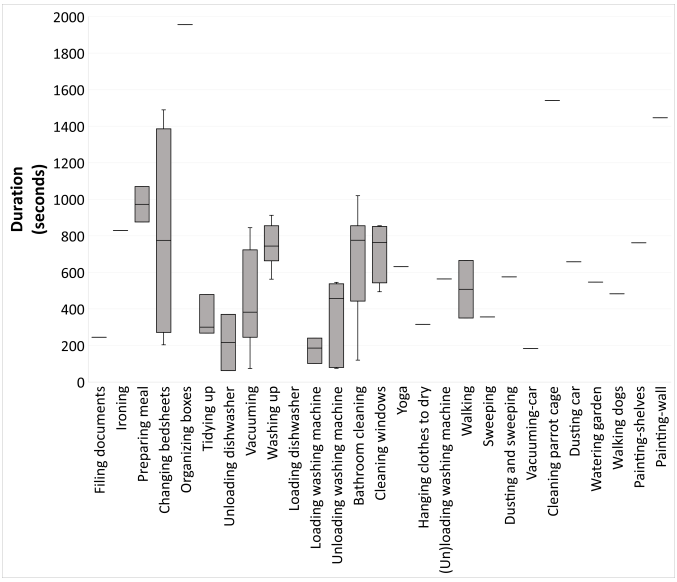

Fig. 5. Duration of activity instances of people with chronic pain across the different recorded activities.

longer duration. Second, a lower sampling rate can additionally limit computational intensity as model input would be smaller in size for the same timescale, and resulting models would be representable by a fewer number of parameters. We evaluated the effect of using a lower sampling rate on activity recognition in movements of people with chronic pain. We ran the same experiment as above but downsampling the data to 10Hz using a moving average window with length and stride of 4 frames.

Results are shown in the second row of Table V. The mean F1 score (0.44) suggests that reducing the sampling rate from 40Hz to 10Hz overall has minimal impact on activity recognition here. However, comparison of performances for each activity type shows improvement with 10Hz data for 5 activity types (*Vacuuming*, *Washing up*, *Cleaning windows*, *Walking activities*, *Tidying up*). There was decrease in performance for *Washing machine activities* and *Changing bedsheets*. For *Washing machine activities* for which performance was especially lower, we found much higher confusions with this activity type misclassified as *Cleaning windows*, *Vacuuming*, and *Washing up* (Figure 6-Left). There was little difference in performance for *Bathroom cleaning* and *Dishwasher activities*.

3) What effect does adding data from healthy participants have on performance?: We further explored influence of additional data from healthy participants on performance. While a larger training set could be valuable, we expected that differences between movement strategies by people with chronic pain and healthy people would undermine any such value and even lead to higher confusion between activities. To investigate this, we combined data from participants with chronic pain resampled to 10Hz with data from healthy participants (captured at 10Hz) for the training, validation, and test sets. For healthy participants, we used data from the 7 participants who had provided data at the time of this study.

The third row of Table V shows the F1 scores obtained per class, with mean F1 score = 0.39. Compared with training and testing on data from participants with chronic pain alone (based on 10Hz data), there is a notable improvement in performance for *Dishwasher activities*, with F1 score higher by 0.33. However, there is deterioration in performance for 4 activity types, with considerable decrease for *Changing bedsheets* (30% decrease) and *Walking activities* (73% decrease) in particular. The confusion matrix (Figure 6-Right) shows that *Walking activities* becomes more strongly misclassified as *Washing up* and *Changing bedsheets* more strongly misclassified as *Washing up* and *Cleaning windows*.

4) Would longer input duration improve performance?: We sought to understand if use of input segments that covered a longer time period would have a positive effect on performance given the timescales of activities in the EmoPain@Home dataset. Several activity instances are well over 15 minutes (900 seconds) in duration (see Figure 5). The disadvantage of longer window segmentation is the reduced number of resulting segments. For this experiment, we used data from the two groups of participants at 10Hz and compared the 3-second segmentation with 12- and 30-second segmentations. For the 12-second segmentation, we used a faster learning rate decay of 3e-5, and for the 30-second segmentation the initial learning rate itself was reduced to 2e-5. This was based on preliminary results with the original learning rate and decay.

The last two rows of Table V shows the results of using 12-second and 30-second segments. The results suggest that larger timescales are not useful for activity recognition in our dataset. The mean F1 score for the 12-second inputs (0.31) is lower than the mean F1 score based on 3 seconds and higher than the mean F1 score based on 30 seconds (0.26). However, with 12 seconds, performance for *Vacuuming* and *Washing up activities* is worse than for both the 3-second and 30-second data. On the other hand, performance for *Cleaning windows* is better with 12 seconds than with 3 or 30 seconds. 30 seconds input additionally lead to worse performance for *Changing bedsheets* and *Tidying up*. *Walking* and *Dishwasher activities*

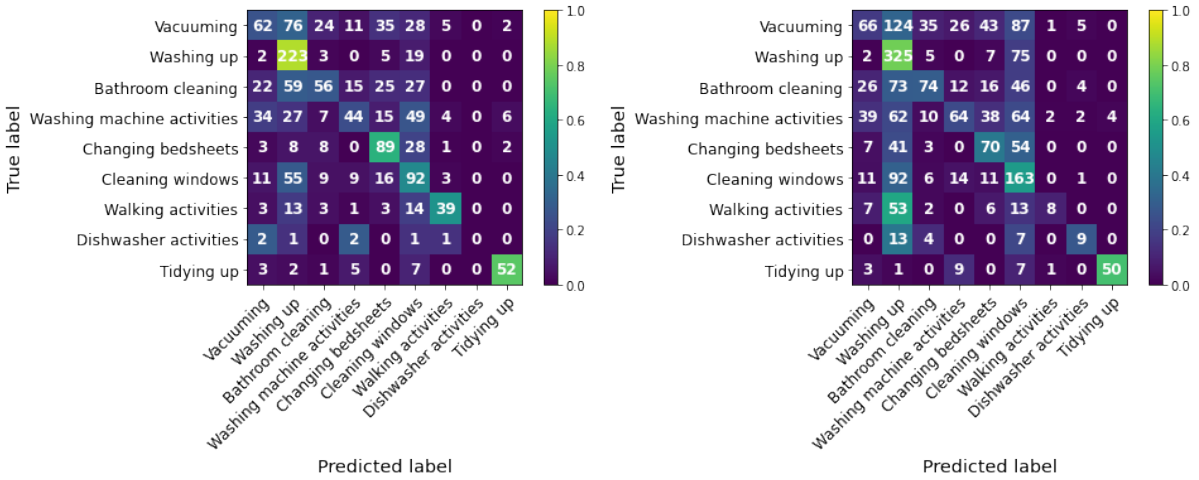

Fig. 6. Confusion matrix for activity recognition based on 3-second, 10Hz segments from participants with chronic pain only (Left) and both participants with and without chronic pain (Right). Colours represent proportion of instances per row. Numbers in the matrix represent actual number of instances.

are not at all recognized at 12- and 30-second timescales. The considerably poorer performance for *Tidying up* is likely a reflection of the resulting few number of instances with 30-second segmentation, only 4 instances in the test set.

VI. Discussion

Our EmoPain@Home dataset captures a wide range of experience (pain, worry, confidence) that people with chronic pain face in everyday functioning at home. The dataset further includes data from healthy people for similar activities. While we were able to obtain mean F1 score of 0.62 for classification of pain levels into two, more data from a subset of participants made pain level classification more difficult. We expect integration of activity recognition to be valuable there. Our investigation of activity recognition with this complex data (with differences in activity types recorded across people, variations in activity type execution between people, limited data size, and minimal degrees of freedom) showed feasibility of such integration. 'Tidying up' in particular was consistently very well recognized in spite of being one of the least represented. Although the others were more difficult for the activity recognition model to differentiate, they were also recognized better than chance level classification, with mean F1 score of 0.43 for 9 classes. Person-dependent modelling will also be crucial. Both can be explored in future work especially as we continue to collect more data and accrue more substantial sizes of data per individual. We acknowledge that our dataset is limited both in the number of participants as well as its overall size. This has implications for the generalization of the above findings. Nevertheless, the dataset is a major step in addressing the critical need for movement data in the wild for chronic pain. Further, our baseline results have significance for pain level and activity recognition in the wild. We discuss this significance in the rest of the section. We additionally discuss recommendations for data capture in home settings that emerge from our data collection, focusing on inclusivity which is particularly relevant for chronic conditions [42].

A. Recommendations for Addressing Inclusivity in Data Capture outside Lab Settings

Despite increasing exploration of sensors for technology for self-management of physical and mental wellbeing and medical conditions, there has been very limited discussion on inclusivity of sensor system designs in the literature. The focus has usually been on preferences for specific population groups in isolation. The experience of participants with chronic pain in our study with the use of a sensor system designed for the general population on their own (i.e. without the researcher) led to insights that highlight important considerations for inclusivity in the creation of real world datasets. We discuss here our recommendations for things to consider for inclusivity in sensor selection/design and participant training.

1) Sensor design: We found that certain kinds of motor manipulations (e.g. turning on the sensor by pinching) that the sensors in our study required were difficult for the participants with (additional) upper limb pain. Such experiences are significant as ethically, the burden on the individual participant must be considered and balanced with the benefits for both the participant and the larger society. Further, sensor use difficulties could contribute to 'data cascades' [59] that negatively affect the performance and use of technology, e.g. by deterring participants from (effectively) engaging with it.

Dataset creators thus need to intentionally incorporate inclusivity in their selection of sensors. We recommend investigation of sensor-related barriers for the specific participant populations in the scenarios of interest. Direct observation in situ may uncover insights that could be missed with other approaches. For example, the study in [60] which relied on participants contacting the researchers to report issues experienced found none with participants with movement impairment. While large scale studies cannot afford direct observation for all data capture sessions, a few sessions with the researcher present could be included in the data collection protocol to enable ethnographic investigation of inclusivity.

We further urge dataset creators to lobby for sensor designs

that are more robustly inclusive. Rather than merely aiming for designs specifically for a given population group, it is more valuable to pursue inclusive designs that address a comprehensive set of requirements for several different groups. This aligns with the accessibility gold standard of catering to the largest possible variety of user abilities [61].

2) Sensor user training: We evaluated the sensors employed as easy to use by people with limited technology experience given training, and so we trained each participant in a half-hour session to enable them capture data on their own. However, we found that they often forgot specific procedures at the time of data capture.

This finding highlights the need for dataset creators to carefully consider learnability in their selection of sensors. While the use of formal post-training tests to check (and require) participant understanding of how to use the sensor(s) can be one approach [60], it comes with inclusivity issues that need to be addressed with care. Ethically, to what extent should participants be excluded if they are unable to grasp how to use the sensors? Also, how many potential participants might drop out from being overwhelmed by a long and cumbersome briefing procedure? We recommend that dataset creators instead aim for sensor systems that are not only easy to use physically but are also of low cognitive demand and so require minimal training to use. Such design would at least ensure that the right amount of information is conveyed at the right time to instruct the user through the process required to set up data capture. The design could further make the sensor system adaptable to available cognitive resources, e.g. through interface options such as 'simple' versus 'enhanced'.

In the long term, it will be valuable to further develop an explicit agenda challenging designers to make sensors easily usable by a broader range of potential users. Learnability has consistently been highlighted across population groups as an important factor of use of relevant systems [62], [63]. Failing to drive engagement with such usability issues will limit the scope and utility of affect-aware systems in the long term.

B. Pain Level Recognition in the Wild

Mean F1 score of 0.62 points to possibility of automatic recognition of lower and higher level pain in real world settings. While the lower performance for the extended data highlights the challenge that automatic detection of pain experience in everyday settings presents, it is not indication of little value in pursuing solutions to this problem. Indeed, the extent of the global burden of musculoskeletal pain and physical rehabilitation needs [64] with the inevitability of self-management at home highlights its necessity. Training data from home settings is clearly critical as movement assessment in clinical contexts does not reflect the same variability and psychological barriers. The work we have presented in this paper provides a baseline on which more sophisticated machine learning approaches or approaches using more data can build.

1) The challenge of limited data: As can be deduced from the sizes of existing datasets for movement disorders, e.g. chronic pain (Table I), Parkinson's disease [65], large sets of data are not trivial to capture, particularly in real world settings. This can be addressed with transfer learning [66], leveraging much larger datasets captured from healthy people or in more controlled movement settings, such as widely done with computer vision and natural language processing [67], [68]. There is little investigation of the efficacy of such transfer (from healthy to chronic pain condition), but our findings in Section V suggest that differences between people with and without chronic pain in the execution of the same activity would need to be specifically addressed. Differences may be direct, such as use of 'protective' behaviours by people with chronic pain to execute the movements involved in an activity that they find challenging (e.g. bending at the knee rather than the hip to reach downward). Differences could also be indirect as in the case of careful setup, by people with chronic pain, of their environment to constrain the postures/movements involved in activities (e.g. sitting for washing up).

2) Unexpected effect of more data: We found that adding more data from a third of the same set of participants (obtained in the extension of the dataset) resulted in poorer generalization especially for the lower level pain class. This finding is likely due to increase in variation introduced with the newer activity instances. Between-subject variation is widely recognized in the affective computing field [4], and we expect variation in pain expression across activity types to also have contributed to the deterioration in model performance. The former can be addressed with the use of personalized models [4], while the latter may be addressed with inclusion of the activity information as context as done in [12]. However, beyond these, variations in environment (e.g. vacuuming the kitchen versus vacuuming the stairs) and within-subject variations in execution strategies of the same activity (e.g. sitting versus standing for washing up) are also implicated. While these sources of variability can be perceived as a limitation, it is an important and critical outcome as capturing homogeneous data would not be valuable for highlighting valid machine learning challenges that need to be addressed. Indeed, findings in [69] suggest that simple data augmentation, which is the widely used strategy, is not enough to build models robust against variations inherent in daily life. The authors further suggest that advanced machine learning approaches will be needed to tackle the problem, and they call for more investigation into the effect of training data (sizes and variations) on robustness.

3) The need for context: Contextual information will be valuable to improve automatic detection performance. One strategy is to use context labels/embeddings that can be learnt from the original input data as additional input [12]. A more robust method could be to apply contextual information to bias pain experience detection through multi-task learning of both context and pain experience constructs simultaneously. The activity labels in the EmoPain@Home dataset could be used as contextual information for pain level recognition. Future data collection studies could employ chatbots to more naturally capture continuous and regular (context) annotations. Findings in [52] suggest that people with chronic pain show openness to use of chatbots in home settings. A different strategy for incorporating contextual information is to use additional sensors to capture other relevant data. For example, given the non-trivial nature of automatic recognition of the composite activities inherent to the home, decibel level sensors could

be explored for capturing sound loudness characteristics that differentiate activities. Vacuuming is, e.g., distinct in loudness.

4) Capturing data at reduced sampling rates: Collecting more data for context would require even more consideration of data storage capacity especially for on-device storage that is usually more practical for small-scale research data collection. In the case of cloud storage, even if at-rest size was of less concern, data transmission between edge device and cloud will be necessary to consider. One strategy that could be explored is reduced sampling rates. Our results suggest that 10Hz can be used for body movement data with little deterioration in activity recognition performance for movements of people with chronic pain. The lower sampling rate will enable longer capture periods and further limit battery drain, which could in turn minimize the recharge burden on participants. Future studies should explore the possibility of such sampling rate for pain level detection. Higher sampling rates have typically been used in the area, e.g. 60Hz in [27], 1024Hz in [31].

C. Activity Recognition for Movement with Pain in the Wild

1) The challenge of composite activities: Our findings provide insight into activity recognition performance for movements of people with chronic pain in real world activities. Such activities are composite, with little clear demarcation between actions or lower level activities within them, e.g. reaching forward and walking during vacuuming, and with strong overlaps across different activities (e.g. walking motions in vacuuming and walking the dogs). There are only a few instances of studies on automatic recognition of composite activities, even for movements of healthy people [48]. In the majority of these studies, the prevailing strategy has been hierarchical recognition where lower-level activities or actions are first classified and then these are used to further differentiate between the higher-level classes [48]. While this is a valuable approach, it depends on fine-grained annotations, which are difficult to obtain. Not only is it expensive to get such labels from human annotators, but in cases where the data captured is not video (e.g. motion capture) and so not an intuitive or natural medium for observation, it may be impossible to get those labels. A more practical strategy with this approach would be to use models trained on existing datasets with lower level labels (e.g. for actions such as reaching downward, or poses such as sitting) for automatic annotation.

In [70], one of the minority studies that does not rely on additional lower-level activity labels [48], contextual information was instead leveraged. They used person-dependent activity recognition models and also employed ambient data (e.g. air temperature) and location data (e.g. Bluetooth beacon) as additional recognition modalities. Thus, the results obtained in [70] are very high, accuracy of 0.92 for 22 activities types averaged across their two participants. The small number of composite activities within the 22 include cooking and washing up. Most of the other activities were simpler, e.g. sitting, standing, lying down. Our own results, although lower (0.48 accuracy for 9 activity types, based on the 10Hz data), show good performance for recognition of complex, composite activities across different participants with chronic pain and different homes with body movement data alone. Performance is particularly good for tidying up, washing up, walking activities, and changing bedsheets (mean F1 score of 0.64 for these 4 activity types). Although vacuuming, washing machine activities, bathroom cleaning, and cleaning windows were more challenging to recognize (mean F1 score of 0.35 for these 4 activity types), performance was still fair for these.

2) The utility of data from healthy people: Our findings further highlight the effect of including data from healthy participants, with walking activities and changing bedsheets showing the strongest (negative) impact. Changing bedsheets is an activity that people with chronic pain find particularly challenging. For example, one of the participants from our dataset reported 'no confidence' (in being able to complete the activity) and worry level of 10 (the highest on the scale) during this movement type. In fact, all 4 participants with chronic pain who performed this activity rated it as typically challenging for them (cleaning windows was the only other such activity that was rated as typically challenging by all participants who performed it). Despite the significance of this activity in everyday life, it is not usually studied and in fact we could not find any studies on either automatic recognition of this activity type or movement strategies used by people with chronic pain in executing it. It is a highly complex activity that can involve very different strategies, especially to deal with worry about the activity or low confidence in the ability to complete it, and levels of both of worry and confidence can change during the activity as found in our dataset. In fact, findings in the literature highlight differences in strategies used by different people with chronic pain for the same movements, compared with healthy people who show less variability [71], [72]. This could explain the poorer performance with inclusion of data from healthy people. The effect on recognition of walking activities is despite the fact that this activity type was only performed by participants with chronic pain. It perhaps suggests that the type of walking motions during washing up (which it was most strongly misclassified as) are very similar in posture and temporal dimension to actual walking activities (e.g. walking outdoor) and this is particularly true for washing up activities for healthy people. The stronger misclassification of walking activities as washing up for larger timescales points to global (rather than temporally local) similarities.

VII. Conclusion

We present the EmoPain@Home dataset of body movement data from people with and without chronic pain captured during everyday physical activities in their homes. The dataset includes labels for pain, worry, and confidence levels for participants with chronic pain. Researchers who wish to use the dataset can contact the corresponding author for access.

Our investigation of pain level classification in real, complex settings of everyday life based on this dataset points to the need for more advanced machine learning strategies than those that have been sufficient for more controlled settings. For example, contextual information that decodes variations in pain expression could also be useful. Additional investigation highlights possibility of automatic recognition of the activity

context for movements of people with chronic pain in such settings. Activity recognition performance could itself be improved, e.g., using transfer learning, although specialized techniques that account for pain versus healthy condition differences may be needed. We aim to extend the dataset in future. This can further improve detection performance.

ACKNOWLEDGMENT

This work was supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent - https://entimement.dibris.unige.it).

REFERENCES

[1] F. R. Avila, C. J. McLeod, M. T. Huayllani *et al.*, "Wearable electronic devices for chronic pain intensity assessment: A systematic review," *Pain Practice*, vol. 21, no. 8, pp. 955–965, 2021.

[2] J. Chen, M. Abbod, and J.-S. Shieh, "Pain and stress detection using wearable sensors and devices—a review," *Sensors*, vol. 21, no. 4, p. 1030, 2021.

[3] A. Leroux, R. Rzasa-Lynn, C. Crainiceanu, and T. Sharma, "Wearable devices: current status and opportunities in pain assessment and management," *Digit. Biomark.*, vol. 5, no. 1, pp. 89–102, 2021.

[4] P. Werner, D. Lopez-Martinez, S. Walter *et al.*, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 530–552, 2019.

[5] S. Walter, S. Gruss, S. Frisch *et al.*, ""what about automated pain recognition for routine clinical use?" a survey of physicians and nursing staff on expectations, requirements, and acceptance," *Front. Med.*, vol. 7, p. 566278, 2020.

[6] S. Felipe, A. Singh, C. Bradley *et al.*, "Roles for personal informatics in chronic pain," in *Int. Conf. Pervasive Comput. Technol. Healthc.*, 2015, pp. 161–168.

[7] J. Smith and M. Osborn, "Pain as an assault on the self: An interpretative phenomenological analysis of the psychological impact of chronic benign low back pain," *Psychol. Health*, vol. 22, no. 5, pp. 517–534, 2007.

[8] C. Murray, C. Groenewald, R. de la Vega, and T. Palermo, "Long-term impact of adolescent chronic pain on young adult educational, vocational, and social outcomes," *Pain*, vol. 161, no. 2, p. 439, 2020.

[9] K. Karos, A. Williams, A. Meulders, and J. Vlaeyen, "Pain as a threat to the social self: a motivational account," *Pain*, vol. 159, no. 9, pp. 1690–1695, 2018.

[10] T. Olugbade, A. Singh, N. Bianchi-Berthouze *et al.*, "How can affect be detected and represented in technological support for physical rehabilitation?" *ACM Trans. Comput. Hum. Interact.*, vol. 26, no. 1, pp. 1–29, 2019.

[11] T. Olugbade, R. Buono, A. Williams *et al.*, "Emopain (at) home: Dataset and automatic assessment within functional activity for chronic pain rehabilitation," in *Int. Conf. Affect. Comput. Intell. Interact.*, 2022, pp. 1–8.

[12] C. Wang, Y. Gao, A. Mathur *et al.*, "Leveraging activity recognition to enable protective behavior detection in continuous data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–27, 2021.

[13] G. Cen, C. Wang, T. Olugbade *et al.*, "Exploring multimodal fusion for continuous protective behavior detection," in *Int. Conf. Affect. Comput. Intell. Interact*, 2022, pp. 1–8.

[14] G. Gioftsos and D. Grieve, "The use of artificial neural networks to identify patients with chronic low-back pain conditions from patterns of sit-to-stand manoeuvres," *Clin. Biomech.*, vol. 11, no. 5, pp. 275–280, 1996.

[15] J. Bishop, M. Szpalski, S. Ananthraman *et al.*, "Classification of low back pain from dynamic motion characteristics using an artificial neural network," *Spine*, vol. 22, no. 24, pp. 2991–2998, 1997.

[16] J. Dickey, M. Pierrynowski, D. Bednar, and S. Yang, "Relationship between pain and vertebral motion in chronic low-back pain subjects," *Clin. Biomech.*, vol. 17, no. 5, pp. 345–352, 2002.

[17] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Machine recognition and representation of neonatal facial displays of acute pain," *Artif Intell. Med.*, vol. 36, no. 3, pp. 211–222, 2006.

[18] P. Levinger and W. Gilleard, "Tibia and rearfoot motion and ground reaction forces in subjects with patellofemoral pain syndrome during walking," *Gait Posture*, vol. 25, no. 1, pp. 2–8, 2007.

[19] B. Matuszewski, W. Quan, and L.-K. Shark, "High-resolution comprehensive 3-d dynamic database for facial articulation analysis," in *Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2128–2135.

[20] P. Lucey, J. Cohn, K. Prkachin *et al.*, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Int. Conf. Automat. Face Gesture Recognit.*, 2011, pp. 57–64.

[21] S. Walter, S. Gruss, H. Ehleiter *et al.*, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *Int. Conf. Cybern.*, 2013, pp. 128–131.

[22] X. Zhang, L. Yin, J. Cohn *et al.*, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[23] J. Rivas, F. Orihuela-Espina, E. Sucar *et al.*, "Detecting affective states in virtual rehabilitation," in *Int. Conf. Pervasive Comput. Technol. Healthc.*, 2015, pp. 287–292.

[24] S. Sharma, S. Asthana, and V. K. Mittal, "A database of infant cry sounds to study the likely cause of cry," in *Proc. Int. Conf. Nat. Lang. Process.*, 2015, pp. 112–117.

[25] Z. Zhang, J. M. Girard, Y. Wu *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3438–3446.

[26] F.-S. Tsai, Y.-L. Hsu, W.-C. Chen *et al.*, "Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions." in *Interspeech*, 2016, pp. 92–96.

[27] M. Aung, S. Kaltwang, B. Romera-Paredes *et al.*, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 435–451, 2016.

[28] M. Velana, S. Gruss, G. Layher *et al.*, "The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, 2017, pp. 127–139.

[29] M. Haque, R. Bautista, F. Noroozi *et al.*, "Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities," in *Int. Conf. Automat. Face Gesture Recognit.*, 2018, pp. 250–257.

[30] T. Olugbade, N. Bianchi-Berthouze, N. Marquardt, and A. Williams, "Human observer and automatic assessment of movement related self-efficacy in chronic pain: from exercise to functional activity," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 214–229, 2018.

[31] B. Hu, C. Kim, X. Ning, and X. Xu, "Using a deep learning network to recognise low back pain in static standing," *Ergonomics*, vol. 61, no. 10, pp. 1374–1381, 2018.

[32] P. Liu, I. Yazgan, S. Olsen *et al.*, "Clinical valid pain database with biomarker and visual information for pain level analysis," in *Int. Conf. Automat. Face Gesture Recogni.*, 2018, pp. 525–529.

[33] S. Gruss, M. Geiger, P. Werner *et al.*, "Multi-modal signals for analyzing pain responses to thermal and electrical stimuli," *JoVE (Journal of Visualized Experiments)*, no. 146, p. e59057, 2019.

[34] C. Ordun, "Intelligent sight and sound: A chronic cancer facial pain dataset," in *Neural. Inf. Process. Syst.*, 2021.

[35] S. Brahnam, L. Nanni, S. McMurtrey *et al.*, "Neonatal pain detection in videos using the icopevid dataset and an ensemble of descriptors extracted from gaussian of local descriptors," *Applied Computing and Informatics*, vol. 19, no. 1/2, pp. 122–143, 2023.

[36] K. Prkachin and P. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[37] C. Bellieni, R. Sisto, D. Cordelli, and G. Buonocore, "Cry features reflect pain intensity in term newborns: an alarm threshold," *Pediatr. Res.*, vol. 55, no. 1, pp. 142–146, 2004.

[38] M. Nickel, E. May, L. Tiemann *et al.*, "Autonomic responses to tonic pain are more closely related to stimulus intensity than to pain intensity," *Pain*, vol. 158, no. 11, pp. 2129–2136, 2017.

[39] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt *et al.*, "Head movements and postures as pain behavior," *PloS one*, vol. 13, no. 2, p. e0192767, 2018.

[40] M. J. Sullivan, P. Thibault, A. Savard *et al.*, "The influence of communication goals and physical demands on different dimensions of pain behavior," *Pain*, vol. 125, no. 3, pp. 270–277, 2006.

[41] L. Dang, K. Min, H. Wang *et al.*, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, 2020.

[42] T. Olugbade, M. Bieńkiewicz, G. Barbareschi *et al.*, "Human movement datasets: An interdisciplinary scoping review," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022.
[43] B. H. Dobkin, X. Xu, M. Batalin *et al.*, "Reliability and validity of bilateral ankle accelerometer algorithms for activity recognition and walking speed after stroke," *Stroke*, vol. 42, no. 8, pp. 2246–2250, 2011.
[44] K. Zhan, S. Faux, and F. Ramos, "Multi-scale conditional random fields for first-person activity recognition," in *Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 51–59.
[45] M. Ahmadi, M. O'Neil, M. Fragala-Pinkham *et al.*, "Machine learning algorithms for activity recognition in ambulant children and adolescents with cerebral palsy," *Journal of neuroengineering and rehabilitation*, vol. 15, no. 1, pp. 1–9, 2018.
[46] N. Capela, E. Lemaire, N. Baddour *et al.*, "Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants," *J. NeuroEng. Rehabil.*, vol. 13, no. 1, pp. 1–10, 2016.
[47] M. Albert, S. Toledo, M. Shapiro, and K. Kording, "Using mobile phones for activity recognition in parkinson's patients," *Front. Neurol.*, vol. 3, p. 158, 2012.
[48] K. Chen, D. Zhang, L. Yao *et al.*, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–40, 2021.
[49] Notch Interfaces Inc, "Notch," 2023, accessed on February 24, 2023. [Online]. Available: https://wearnotch.com/
[50] M. P. Jensen and P. Karoly, "Self-report scales and procedures for assessing pain in adults." in *Handbook of pain assessment*, D. C. Turk and R. Melzack, Eds. The Guilford Press, 2011, pp. 19–44.
[51] N. Cliff and J. A. Keats, *Ordinal measurement in the behavioral sciences*. Psychology Press, 2003.
[52] T. Bi, R. Buono, T. Olugbade *et al.*, "Towards chatbot-supported self-reporting for increased reliability and richness of ground truth for automatic pain recognition: Reflections on long-distance runners and people with chronic pain," in *Companion Publication Int. Conf. Multimodal Interact.*, 2021, pp. 43–53.
[53] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.
[54] R. A. Armstrong, "When to use the b onferroni correction," *Ophthalmic and Physiological Optics*, vol. 34, no. 5, pp. 502–508, 2014.
[55] T. K. Sen, G. Naven, L. Gerstner *et al.*, "Dbates: Dataset for discerning benefits of audio, textual, and facial expression features in competitive debate speeches," *IEEE Transactions on Affective Computing*, vol. 14, no. 02, pp. 1028–1043, 2023.
[56] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[58] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[59] N. Sambasivan, S. Kapania, H. Highfill *et al.*, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–15.
[60] S. Rutkove, K. Qi, K. Shelton *et al.*, "Als longitudinal studies with frequent data collection at home: study design and baseline data," *Amyotroph. Lateral Scler. Front. Degener.*, vol. 20, no. 1-2, pp. 61–67, 2019.
[61] S. Keates and P. Clarkson, "Countering design exclusion: bridging the gap between usability and accessibility," *Univers. Access Inf. Society*, vol. 2, pp. 215–225, 2003.
[62] T. Grossman, G. Fitzmaurice, and R. Attar, "A survey of software learnability: metrics, methodologies and guidelines," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2009, pp. 649–658.
[63] N. Chaniaud, C. Sagnier, O. Megalakaki, and E. Loup-Escande, "Relationship between efficiency, effectiveness, and learnability of home connected medical device in ambulatory surgery," *Telemedicine and e-Health*, vol. 28, no. 6, pp. 904–911, 2022.
[64] A. Cieza, K. Causey, K. Kamenov *et al.*, "Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020.
[65] S. Pardoel, J. Kofman, J. Nantel, and E. D. Lemaire, "Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: a review," *Sensors*, vol. 19, no. 23, p. 5141, 2019.
[66] T. Olugbade, A. C. de C Williams, N. Gold, and N. Bianchi-Berthouze, "Movement representation learning for pain level classification," *IEEE Transactions on Affective Computing*, 2023.
[67] Y. Li, J. Wei, Y. Liu *et al.*, "Deep learning for micro-expression recognition: A survey," *IEEE Trans. Affect. Comput.*, 2022.
[68] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affect. Comput.*, pp. 49–67, 2021.
[69] R. Taori, A. Dave, V. Shankar *et al.*, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 583–18 599, 2020.
[70] P. Vepakomma, D. De, S. Das, and S. Bhansali, "A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities," in *Int. Conf. Wearable Implant. Body Sens. Netw.*, 2015, pp. 1–6.
[71] P. Ippersiel, S. Robbins, and R. Preuss, "Movement variability in adults with low back pain during sit-to-stand-to-sit," *Clin. Biomech.*, vol. 58, pp. 90–95, 2018.
[72] M. Mahdi Dehshibi, T. Olugbade, F. Diaz-de Maria *et al.*, "Pain level and pain-related behaviour classification using gru-based sparsely-connected rnns," *IEEE J. Sel. Top. Signal Process.*, 2022.

Temitayo Olugbade is a Lecturer in Computer Science and AI at University of Sussex and Honorary Fellow at University College London (UCL). Her research pursues development and application of state-of-the-art machine learning methods to new and challenging affective computing contexts.

Raffaele Andrea Buono is an anthropologist, and doctoral candidate at University College London. His research explores robotics engineering projects in Japan, aiming to carve out new approaches to social theory by paying attention to technical development and design. Alongside his main work, he is interested in exploring ways in which social and computer scientists could collaborate, towards the development of inter-disciplinary endeavours.

Kyrill Potapov is a doctoral candidate in HCI and a research fellow in anthropology. His work focuses on personal data and learning.

Alex Bujorianu currently works at Holistic AI as a machine learning engineer; he audits AI models for bias. He also has some background in moral philosophy and has a keen interest in AI applications for medicine, ethical AI, and economic modelling. He graduated with distinction from the University of Twente and Amsterdam, in data science and economics respectively. Alex is a published fantasy author, under the nom de plume Alex Stargazer.

Amanda C de C Williams is a professor of clinical health psychology at UCL, consultant clinical psychologist in pain management at UCL Hospital, UK, and Section Editor for Psychology on the journal PAIN. Her research interests include evidence-based medicine applied to psychologically-informed interventions for pain, including systematic review and meta-analysis; behavioural expression of pain and its interpretation; and responsive wearable technology to extend healthcare into patients' own environments.

Santiago de Ossorno Garcia currently works at NHS as a HPCP registered Counselling Psychologist with Children and Young People. He also has extensive community experience in third-sector organisations with a focus on homelessness and trauma-informed care with adults. He has worked as a digital mental health expert and researcher at Kooth and University College London supporting innovation and evidence-based research. He has a PhD from Universidad Complutense de Madrid exploring the influence of parental context in bullying and victimisation in the classroom.

Nicolas Gold is Associate Professor at UCL Computer Science. His research interests are in the comprehension and comprehensibility of design representations in a range of disciplines including source code, educational resource design, and research ethics design. He has published research in computational musicology, creative music systems, research ethics, and music computing applications in healthcare.

Catherine Holloway is a co-founder and Academic Director of Global Disability Innovation Hub, and Professor of Interaction Design & Innovation. She works to accelerate disability justice through the invention of new devices, knowledge and data.

Nadia Bianchi-Berthouze is Full Professor in Affective Computing & Interaction at the University College London Interaction Centre. Her research focuses on designing technology that can sense the affective state of its users and use that information to tailor interaction. She has pioneered the field of Affective Computing in investigating how body movement and touch behaviour can be used as means to measure quality of user experience.