scientific reports

Check for updates

OPEN

A data-driven approach to detect upper limb functional use during daily life in breast cancer survivors using wrist-worn sensors

Jill Emmerzaal[1], Benjamin Filtjens[2,3], Nieke Vets[1], Bart Vanrumste[2], Ann Smeets[4], An De Groef[1,5,7]✉ & Liesbet De Baets[6,7]

To gain insights into the impact of upper limb (UL) dysfunctions after breast cancer treatment, this study aimed to develop a temporal convolutional neural network (TCN) to detect functional daily UL use in breast cancer survivors using data from a wrist-worn accelerometer. A pre-existing dataset of 10 breast cancer survivors was used that contained raw 3-axis acceleration data and simultaneously recorded video data, captured during four daily life activities. The input of our TCN consists of a 3-axis acceleration sequence with a receptive field of 243 samples. The 4 ResNet TCN blocks perform dilated temporal convolutions with a kernel of size 3 and a dilation rate that increases by a factor of 3 after each iteration. Outcomes of interest were functional UL use (minutes) and percentage UL use. We found strong agreement between the video and predicted data for functional UL use (ICC = 0.975) and moderately strong agreement for %UL use (ICC = 0.794). The TCN model overestimated the functional UL use by 0.71 min and 3.06%. Model performance showed good accuracy, f1, and AUPRC scores (0.875, 0.909, 0.954, respectively). In conclusion, using wrist-worn accelerometer data, the TCN model effectively identified functional UL use in daily life among breast cancer survivors.

The upper limbs (ULs) are essential to our everyday life. To clean our houses, to complete work tasks, to drink, to eat, to scratch, to help or touch other people, etc. Usually, we are unaware of how much we use our ULs until we are unable to properly use them. Life events such as surgery or treatment for breast cancer are moments that might suddenly affect UL function.

After breast cancer treatment, an estimated 30–50% of women suffer from persistent UL dysfunction[1]. This persistent dysfunction goes beyond the natural healing time of tissue after surgery or radiotherapy, and it is considered one of the most troublesome long-term complications after breast cancer treatment[1–3]. Women suffering from UL dysfunction are less able to perform their daily tasks, have an increased risk for chronic pain, and suffer from participation difficulties, which lead to a decrease in their quality of life[1,3].

This impact on everyday life is currently only assessed using questionnaires. While questionnaires are easily administered and provide valuable information on a person's perception of UL function, they have important drawbacks such as recall bias and self-presentation bias (i.e. disclose only what the person wants to or is consciously aware of). Next to questionnaires, clinic-based assessments of body function, e.g. UL range of motion and activity or functional task performance, can measure the capability of a patient but might be poor indicators of actual UL use in daily life[4–6]. Thus, assessment of behaviour in a natural setting is vital, to assess the real-world impact of surgery on UL use and to evaluate UL recovery[7].

To assess UL function, a few definitions need to be addressed. UL function is previously defined by the functional arm activity behavioural observation system (FAABOS) and distinguishes functional and non-functional use[8]. Functional UL use enables interactions with the environment with purpose and can be task-oriented (e.g. opening a jar) or not-task oriented (e.g. touching face)[8]. Non-functional UL use is an action that has no or minimal function (e.g. arm swing during walking)[8]. The total amount of functional arm movements is defined

[1]Department of Rehabilitation Sciences, KU Leuven, 3000 Leuven, Belgium. [2]Department of Electrical Engineering (ESAT), KU Leuven, 3000 Leuven, Belgium. [3]Department of Mechanical Engineering, KU Leuven, 3000 Leuven, Belgium. [4]Department of Surgical Onocology, University Hospitals Leuven, KU Leuven, 3000 Leuven, Belgium. [5]Department of Rehabilitation Sciences, University of Antwerp, 2000 Antwerp, Belgium. [6]Department of Physiotherapy, Human Physiology and Anatomy, Vrije Universiteit Brussel, 1000 Brussels, Belgium. [7]These authors jointly supervised this work: An De Groef and Liesbet De Baets. ✉email: an.degroef@kuleuven.be

by David et al. as UL use' in people with hemiparesis[7]. However, assessing UL use in daily life, which involves distinguishing functional and non-functional UL use from each other, is extremely challenging[7,9–12]. These challenges in daily life encompass the lack of control over the environment and limitations in sensing modalities[7,12]. For example, when using accelerometers, a bus ride will induce many instances of acceleration not associated with functional movement[12]. Distinguishing the accidental acceleration induced by external forces from acceleration signals induced by a decisive action is not a trivial undertaking.

Fortunately, advances in advanced data analysis methods such as machine learning are on the rise within a clinical setting and provide new opportunities regarding the assessment of real-life UL functional use. We applied a pre-trained machine learning model developed by Lum et al.[12], to detect functional UL use in daily life in breast cancer survivors[13]. Lum's model[12] is a decision tree model trained on healthy controls and neurological patients in a controlled setting. Predictions on functional/non-functional UL use are made per 4-s epoch and the label is given on the most common prediction in that epoch[12]. Whilst this method was shown to be accurate in a controlled setting for healthy individuals, the accuracy and especially the f1-score decreases drastically in breast cancer survivors in a real-world setting[13]. The f1-score describes the harmonic mean of the precision and recall—confidence in the true positives while balancing the false positives[14]. We concluded that a pre-trained lab-based model might not be sensitive enough to be used in daily life[13]. This indicates we might need a more powerful tool to correctly estimate functional UL use in daily life. Such a potentially powerful tool for time-series data could be a deep learning model.

The use of deep learning in industry and research has been on the rise since 2012[15]. Machine learning and deep learning are related fields within artificial intelligence that use data to train an algorithm to make predictions or decisions. Deep learning represents more advanced model architectures that can automatically learn patterns and representations directly from the data, in comparison with machine learning which uses more superficial models that rely on predefined features manually extracted from the data[16]. A fundamental aspect of both machine learning and deep learning is the interaction between optimisation and generalisation—i.e. how well can we train a model on the training data (optimisation), and how well the model responds to new, unseen data (generalisation). Machine learning and deep learning models achieve generalisation by learning models that can interpolate between the different training examples—the model can make sense of things that are close to what it has seen before[15]. However, deep learning models have more parameters and hence require more data[15,16].

Studies by Kaggle competitions—a machine learning platform that organises competitions—showed that in the last decade, the winning teams predominantly used deep learning models over machine learning[15]. This rise in using deep learning coincides with the technological advances that have been made in recent years. The long short-term memory (LSTM) algorithm, which is used for time-series analysis, was developed in 1997[17] and had to be run on a room-sized computer. Advances in datasets and technology currently make it possible to run small deep-learning models on laptops, which made this technique more accessible to a wider range of researchers and clinicians. This resulted in a boost of deep learning models applied in health care and health care applications. Using the simple search terms "Deep learning" and "Health Care" in PubMed shows a clear increase in publication count in the last decade (Fig. 1).

One specific deep neural network architecture that may be relevant for detecting functional UL use in daily life is a temporal convolutional neural network (TCN) model[18]. TCN model's leverage hierarchical architectures to automatically learn features at different temporal resolutions. They accomplish this by utilising dilated convolutions[19], which increase the receptive field, or field of view (i.e. the region in the input that affects a certain output value). This hierarchical feature learning capability enables TCN models to capture both short-term and long-term temporal dependencies, allowing them to extract meaningful representations from sequential data. TCN models have shown to outperform recurrent neural network architectures, such as LSTMs[17] and gated recurrent units (GRUs)[20], on various sequence modelling benchmarks[21]. Given the TCN model's previous successes, we believe that this model's architecture might be suited to distinguish functional and non-functional UL use.

Given all this, this research introduces novelty through (1) Addressing the critical need for objective assessment of UL dysfunctions in breast cancer survivors, essential for tailored rehabilitation plans; (2) Proposing

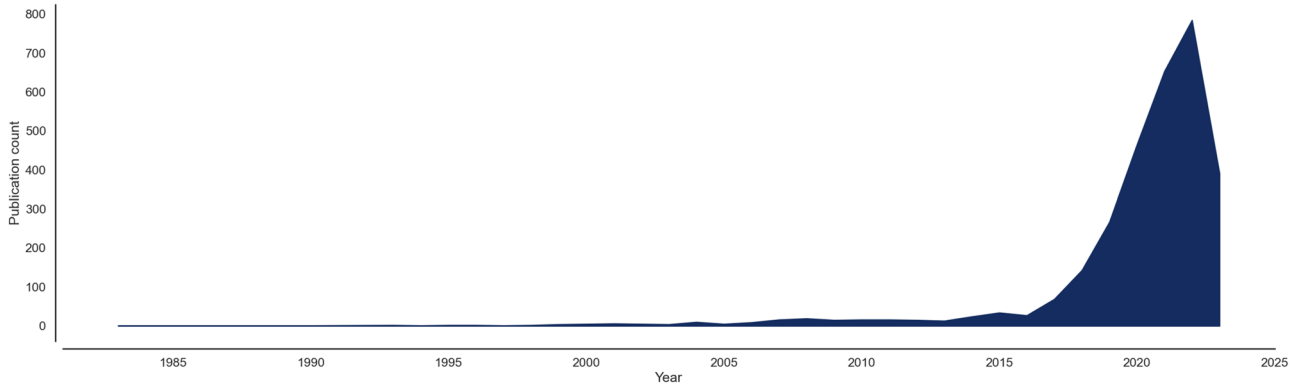

Figure 1. Overview of the number of publications per year using the terms "Deep learning" and "Health care" in PubMed. Data downloaded on June 24, 2023.

accelerometry for minimally disruptive evaluation of UL functional use in daily life settings. However, available existing research suggest that laboratory-developed analysis methods are inadequate for real-life scenarios in breast cancer survivors. Thus, novel methods are needed to accurately detect UL functional use outside laboratory settings. (3) Using a TCN model that incorporates sequential data capable of using both long- and short-term dependencies might be more relevant than looking at individual data points, and (4) Employing deep learning enables the model to learn directly from the data, a crucial capability in this field where determining relevant features to detect UL functional use is challenging.

To summarise, the goal of this study was to develop a TCN model to predict functional UL use in daily life in breast cancer survivors using wrist-worn accelerometer data. If this can be done successfully, it enables clinicians and researchers to gain more insight into the impact of UL dysfunction in daily life in women who received breast cancer treatment, and into their recovery trajectory.

Method

Dataset

This is a secondary analysis from an existing dataset[13]. The study was approved by the local ethical committee of the University Hospital Leuven (s66248) and is part of a larger project at the University Hospital Leuven aiming to identify persistent upper limb dysfunction in breast cancer survivors (UPLIFT-BC, clinicaltrails.com: NCT05297591)[22]. The inclusion and exclusion criteria are in Table 1.

All experiments were conducted following the Declaration of Helsinki, and all participants provided written informed consent before the start of the study.

Test protocol

The participants were instructed to perform four specific activities of daily living (ADL) in their own home environment. The ADL tasks included a laundry activity, a kitchen activity, a shopping activity, and a bedmaking activity. Details about the activities are in Table 2. To mimic daily life as closely as possible, the only instructions that were given are the ones described in Table 2 combined with "perform the activity as you would normally do." No further instructions were given. In between activities, the participants were instructed to walk to a chair and sit down to increase the amount of non-functional UL use. During that time the participant was encouraged to interact with the researcher to provoke habitual arm movements such as hand gestures during talking.

All participants were equipped with an accelerometer on each wrist (ActiGraph wGT3X-BT, sample frequency: 30 Hz, ActiGraph Corporation, Pensacola, FL) and were simultaneously filmed (Sony FDR-AX33, 25 fps). The ADL tasks (inclusive walking bouts between activities and the sitting moments) were performed and filmed as one consecutive measurement. The accelerometers and video were synchronized using 3–5 fast repetitive arm movements at the beginning of the measurement and at the end. To ensure that data remained synchronised throughout the measurement, spot checks were performed when the participants did not move their arms.

Functional UL use was annotated per arm by a single researcher (JE) with Adobe Premier Pro (version 2023) and was considered the ground truth. As described in the introduction, arm use was defined by the FAABOS[8] that distinguishes functional (i.e. task-specific movement with interaction with the environment) from non-functional (e.g. arm swing during walking) UL use. When a functional activity with the upper limbs was noticed, the start and end time were noted using timeline markers indicating whether left, right, or both showed functional UL use. When the upper limbs or hands were not visible, a timeline marker was placed with the label "unknown". The frames with unknown labels were removed from both the video and accelerometer data during model evaluation.

Inclusion	Exclusion
Unilateral breast surgery > 1 month ago	Distant metastases
Proficient in the Dutch language	History of breast surgery
	Planned bilateral surgery
	Neurological or rheumatological disease

Table 1. Inclusion criteria.

Activities	Description
Laundry activity	(1) Move clothes from a closet or basket into a washer, and close the washer, (2) remove the clothes from the washer, put them in the dryer and close the door, and (3) remove the clothes from the dryer and fold them or hang them back in the closet
Kitchen activity	(1) Load and unload four or five items from the dishwasher, (2) cut an apple or equivalent, (3) pick up one item from the floor and (4) use a broom, dust mop, or vacuum to sweep the floor
Shopping activity	(1) Gather four or five items out of the supply closet in their grocery store bag or box, (2) place them into the car, step into the car, then step out, and remove the groceries from the car, and (3) put the groceries back in the supply closet
Bed making activity	(1) Remove the sheets and pillowcases from their bed and (2) replace them

Table 2. Descriptions of performed activities of daily living.

After completing the video annotation, the timeline markers were exported to a CSV file for further analysis in MATLAB (MATLAB 2021b, The MathWorks, Inc. Natick, Massachusetts, USA).

The raw acceleration data was extracted from the accelerometers and the start point was synchronised with the video data by identifying the start of the calibration movement at the beginning of the measurement in both the acceleration and video data. The 3-axis acceleration data were centralised by subtracting the mean value of each signal to remove the constant bias. A single deep-learning model was used for the left and right acceleration data by flipping the transverse axis of the left accelerometer.

Deep neural network architecture

Data-driven deep learning approaches, such as recurrent neural networks (RNNs)[23] and convolutional neural networks (CNNs)[24], have shown great success in many problems that contain temporal information[25]. Unlike RNNs, convolutional models ensure a fixed length for the gradient path between the output and input, regardless of the sequence length. This characteristic helps prevent issues such as vanishing and exploding gradients, which commonly affect RNNs[21]. CNNs can capture long-term dependencies efficiently by utilizing dilated convolutions[19], which has enabled state-of-the-art results in various human activity recognition tasks based on wearable sensor data[26–29].

Our model is a temporal convolutional neural network (TCN)[18] that takes a sequence of acceleration data as input and transforms it through temporal convolutions. The input sequence can be represented as a 2D matrix $X \in R^{T \times C_{in}}$, where T represents the number of samples and C_{in} represents the 3-axis acceleration features. The output can also be represented as a 2D matrix $Y \in R^{T \times L}$, where L represents the two output classes (i.e. functional- and non-functional UL use). The TCN model is tasked to learn the mapping $f : X \rightarrow \widehat{Y}$ so that the predicted output $\widehat{Y}$ closely matches the ground truth labels Y, which was the expert visual assessment of the video data.

The first layer of our TCN model is a non-dilated temporal convolutional layer with a kernel size of k that maps the input to a C-dimensional feature map, with C the number of convolutional filters. This feature map then goes through a series of B ResNet-style TCN blocks, which are supplemented with skip-connections[30]. Each TCN block applies a dilated valid temporal convolution with a kernel size of k and a dilation factor of $d = k^B$, followed by a $1x1$ convolution. Each convolution operation is followed by batch normalisation[31], rectified linear units (ReLU)[32], and dropout[33]. Finally, the last layer implements a $1x1$ convolution and a SoftMax activation function to output a prediction for all samples in the input sequence, exploiting both past and future temporal information.

Implementation details

Our TCN implementation uses dilated and valid temporal convolutions. Dilated convolutions entail that each convolution operation considers a local receptive field determined by the kernel size and dilation rate. Valid convolutions entail that the input is not padded before performing the convolution operation. As a result, the temporal dimension of the output feature maps gradually decreases after each convolutional layer. Therefore, we slice the residuals to match the shape of subsequent feature maps.

During inference, the model efficiently processes entire accelerometer sequences. To avoid losing samples to valid convolutions, replication padding is applied at the input boundaries of a sequence. Our TCN implementation was based on the work of Pavllo et al.[34]. For visualization of the network structure during training and inference we refer to Appendix 6 of their work[34].

Our TCN model was trained for 10 epochs with the Adam optimizer[35]. We used the cross-entropy loss function and set the learning rate to 0.0005. The "unknown" data samples were included during training to not disrupt the temporal consistency of the input data. Though, they were excluded from the calculation of the cross-entropy loss. Our architecture contained four ResNet-style TCN blocks. The convolutional layers had 128 filters, each with a kernel of size 3, resulting in a receptive field of 243 samples or 8.1 s. The dropout was set to 0.5. Our model was implemented using PyTorch version 1.10.1[36]. It was trained and evaluated using a leave-one-subject-out cross-validation approach to assess its ability to generalize to subjects that had not been seen before. A visual overview of our TCN architecture is given in Fig. 2.

Outcome and metrics

The main outcome of this paper is the functional upper limb use expressed in minutes and as a percentage per participant per wrist. As described in the introduction, functional UL use is defined as the total duration of functional arm movements expressed in minutes. Since measuring time is not the same for every participant in the data set, we also calculated the percentage of functional UL use. This is the percentage of functional UL use to the total measuring time of a certain participant per wrist. To assess the agreement between the annotated data and the predicted data by the model of the main outcome measures, we calculated the intraclass correlation coefficient with a two-way random effect (ICC(2,1)). The strength of the agreement was defined as: < 0.5: poor, 0.5–0.75: moderate, 0.75–0.9: good, and ≥ 0.9: excellent[37].

To visualise and analyse the error between the two methods (annotated vs. predicted) we create Bland–Altman plots[38]. These plots show the average error in minutes for functional UL use and the average error in the percentage of functional UL use. The line of equality shows where the error between annotations and prediction would be zero (Annotation – Model = 0)[39]. If the line of equality falls outside of the confidence interval of the mean, this could indicate that the bias is substantial[39]. The limit of agreement shows within which interval 95% of the differences of the second method, compared to the first method, fall[39].

To assess model performance, we calculated accuracy, f1-score, and area under the precision-recall curve (AUPRC). Accuracy is defined as the percentage of all correctly labelled frames as a function of all frames in the time series. The f1-score is the harmonic mean of precision and recall—it shows the model's confidence in the

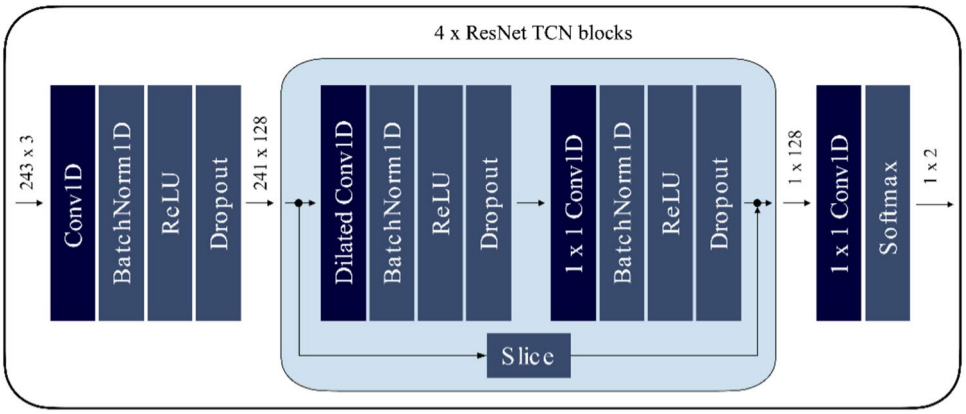

Figure 2. Our temporal convolutional neural network architecture. The input consists of a 3-axis acceleration sequence with a receptive field of 243 samples. The 4 ResNet TCN blocks perform dilated temporal convolutions with a kernel of size 3 and a dilation rate that increases by a factor of 3 after each iteration. As a result, the TCN model covers the entire receptive field of the input. Due to the valid convolutions, we slice the residuals to match the shape of subsequent feature maps.

true positives while balancing the false positives. The AUPRC is a useful metric for imbalanced classes when the focus is to find the positive—in our case, functional activity—classes. Perfect AUPRC equals one and indicates that all positive classes were found (perfect recall), without mislabelling any of the negative classes (perfect precision). The "unknown" samples were excluded from the model's performance evaluation.

Results

Ten women participated in this study; their demographics are in Table 3. All participants received breast cancer surgery, of which six underwent a mastectomy and four breast conservative surgery. In all patients, axillary staging was performed with sentinel lymph node biopsy (n = 7) or axillary lymph node dissection (n = 3). (Neo)-adjuvant treatment consisted of chemotherapy in four patients, radiotherapy in five and endocrine treatment in seven (four tamoxifen, and three aromatase-inhibitor). Most women underwent a combination of different treatment methods (Table 3). Self-reported UL function was measured with the Quick DASH questionnaire[39] with zero meaning no limitations in daily life. Applying a cutoff score of > 15/100 on the QuickDASH to identify patients with upper limb function limitations[40], we found that 30% of our participants experienced such dysfunctions.

In Figs. 2 and 3, we see the scatter plots and Bland Altman plots, respectively, for functional UL use and the percentage functional UL use of the annotated data and the predicted data. From the scatter plots, we see an overall overestimation of the functional UL use (left plot in Fig. 3) for the predicted data by the prediction model. This can also be seen in the larger overestimation of percentage functional UL use (right plot in Fig. 3). From the Bland–Altman plots (Fig. 4) we see the systematic overestimation for functional UL use by the prediction model, with an average bias of − 0.71 min (min) for functional UL use (functional UL use ranges from 12.11 to 24.24 min, see Fig. 3, left) and − 3.06% for percentage functional UL use (percentage functional UL use ranges from 69.21 to 89.13%, see Fig. 3, right and Table S1). The limits of agreement range from − 2.51 to 1.09 min for

Subj ID	Age (years)	Operated side	Surgery	(Neo-) adjuvant treatment	Quick DASH score (0–100)
P01	44	R	ME + SN	TAM	0
P02	48	L	ME + SN	Adj. CT + TAM	9.1
P03	50	R	BCS + SN	Adj. RT + TAM	38.6
P04	53	L	ME + SN	–	4.5
P06	52	L	BCS + SN	Adj. RT + TAM	13.6
P07	45	L	ME + ALND	Neo-adj. CT + Adj. RT + AI	11.4
P08	52	R	ME + ALND	Neo-adj. CT + Adj. RT	15.9
P09	43	R	ME + SN	TAM	0
P10	65	R	BCS + SN	Adj. RT + AI	15.9
P11	72	L	BCS + ALND	Neo-adj. CT + Adj. RT + AI	11.4

Table 3. Participant characteristics. *P* participant, *L* left; *R* right, *ME* mastectomy, *BCS* breast-conserving surgery, *SN* sentinel lymph node biopsy, *ALND* axillary lymph node dissection, *Neo-adj*. neo-adjuvant treatment, *Adj*. adjuvant treatment, *CT* chemotherapy, *RT* radiotherapy, *TAM* tamoxifen, *AI* Aromatase-inhibitor, *DASH* disability of arm, hand and shoulder questionnaire and a higher score indicates more disability.

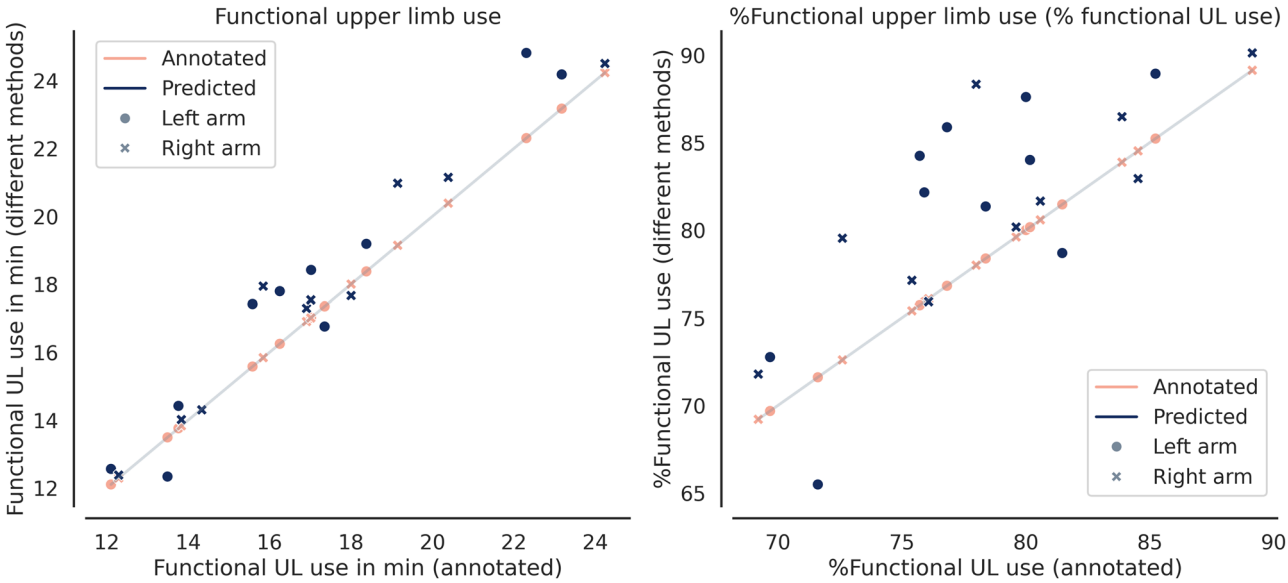

Figure 3. Scatterplot of functional upper limb use (left) and percentage of functional upper limb use (right) for the annotated data and the predicted data. Every marker point represents an individual left, denoted as a circle, or right arm, denoted with x, of each participant.

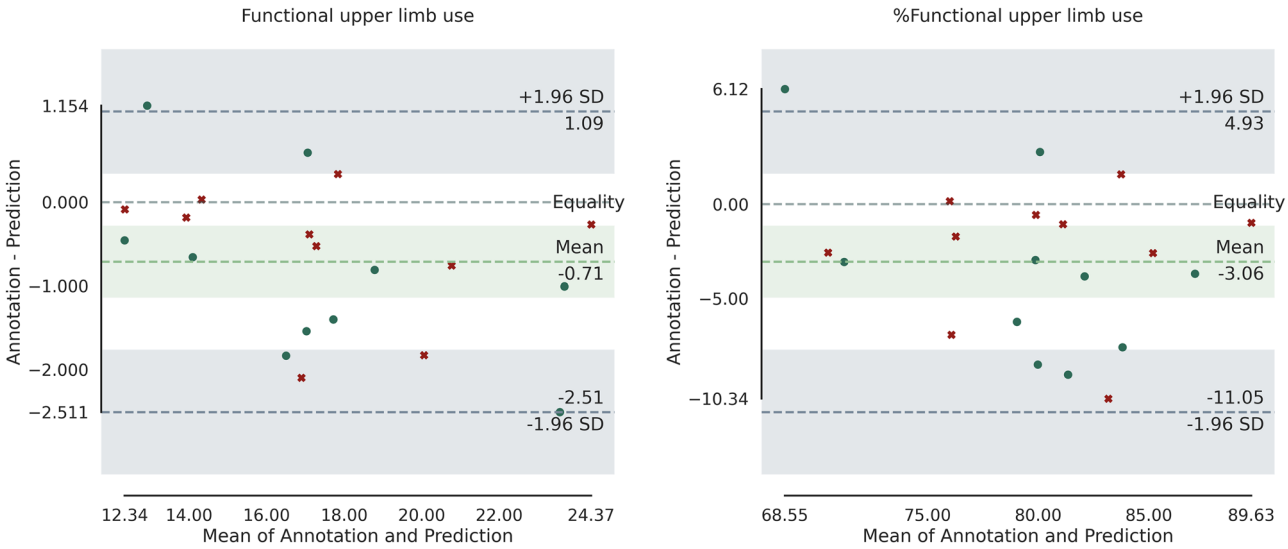

Figure 4. Bland–Altman plot of functional upper limb use (left) and percentage of functional upper limb use (right) for the annotated data and the predicted data. The Bland–Altman plot shows the error between the annotated data and the predicted data. The left arm data is represented with green circles and the right arm with the red crosses. The shaded green area is the average with the confidence interval. Shaded grey areas are the limits of agreement with the corresponding confidence interval. The equality line shows where the difference between annotation and prediction is 0. Every marker point shows the error between the annotated data and the predicted data for each arm of each participant.

functional UL use and − 11.05 and 4.93% for percentage functional UL use (Fig. 4 and Table S1). All individual time data can be found in the supplementary materials (Table S1).

Furthermore, we found an excellent agreement between the predicted data and the annotated data for functional UL use (ICC = 0.975 [0.88, 0.99]) and a good agreement for the percentage functional UL use (ICC = 0.794 [0.33, 0.93]). On average, the accuracy, F1-score, and AUPRC of our model's metrics were good (accuracy = 0.857 ± 0.068; F1-score = 0.909 ± 0.04722, AUPRC = 0.954 ± 0.057).

Discussion

The goal of this study was to develop a TCN model to predict functional UL use in daily life in breast cancer survivors using wrist-worn accelerometer data compared to video annotated data. We found a strong agreement between the annotated video data and the predicted data for the TCN model of functional UL use (0.975) and a

moderately strong agreement for the percentage of functional UL use (0.794). These high values coincide with good accuracy, f1, and AUPRC scores (0.875, 0.909, 0.954, respectively).

However, we found that our TCN model overestimated, on average, the amount of functional UL use by 0.71 min or 3.06%. While this is a small percentage, the limits of agreement ranged from − 2.51 to 1.09 min and − 11.05 and 4.93%, where the negative number is an overestimation of the TCN model's predictions. This indicates that the prediction model works better for some participants than others. More detailed inspection of our participants revealed that the left arms of P02, P04, P07, P09, and P10 and the right arm of P10 had absolute errors greater than 5% for percentage functional UL use, ranging from 6.12 to 10.33% (Fig. 4 and Table S1 in the supplementary materials). All these women, apart from subject 10, were operated on the left side. In those cases, the model predicted large, uninterrupted sections of functional UL use, whereas the video annotations showed a more intermittent functional UL use (example shown in Fig. 4) with functional UL use being interspersed with non-functional UL use. We hypothesise that arm swing during walking is likely misclassified by the prediction model as functional UL use. During the annotation process, the annotator (JE) noticed that some women walked with an abnormal arm swing, i.e. the arms were kept rigid and close to the torso at an elbow angle between 120 and 90 degrees. Inspection of the video data confirmed that this arm position during walking is true for P02, P04, and P10 left arms, but not for P07 and P09. The acceleration signal of such an "arm swing" pattern (classified as non-functional UL use according to FAABOS[8]) will be tough for a TCN model to distinguish from carrying a small item, like keys, in the hand (classified as functional UL use according to FAABOS[8]) (see Fig. 5 for an example).

Despite the overestimation of functional UL use, the TCN model shows promising results in distinguishing functional from non-functional UL use. The average error of the percentage functional UL use is far better using the TCN model (3.04%) than using the pre-trained decision tree model (14%) reported in Vets et al.[13]. While we compared one outcome to the outcome from Vets et al.[13], we cannot comment on the superiority of the TCN model compared to a decision-tree model since the current state-of-the-art TCN model is trained, validated, and tested on the data presented here. The decision tree model is trained and validated on laboratory-based data from healthy controls from the study of Lum et al.[12] and only tested on the data presented here. Thus, a direct comparison of the TCN model vs. the decision tree model and the models' performance (e.g. accuracy, f1-score) would not be fair. However, since TCN models have steadfastly outperformed other deep learning models on various sequence modelling benchmarks[21], we can be confident that a simpler machine learning model, like a decision tree, would not show superior performance over a TCN model in such a complex task.

Though we found promising results, we must consider the data we used and the limitations that are associated with that. Our data was skewed to the functional activity side, where our participants performed roughly 80% functional UL use and only 20% non-functional UL use. With this, two issues arise. First, we need to be careful with the interpretation of the f1-score, as this score completely ignores the true negative class[14]. Therefore, this metric is typically used when we have a positive minority class. Here, we have a positive majority class for the given dataset. However, when evaluating on a full-day basis, the functional activities would rather be the smallest class and the f1-score would be the metric of choice. That is why we opted to use the f1-score in this study although our data distribution is not skewed towards the functional activity. Secondly, 20% of non-functional UL use data might not contain enough complexity to provide meaningful representations for the TCN model. Since we hypothesised that arm swing might be misclassified as functional UL use, we propose that future work could look at creating a more fine-grained model that also predicts arm swing during walking, thus predicting 3 categories: (1) non-functional UL use, (2) functional UL use, (3) arm swing. Moreover, this dataset of 10 women should be enlarged with assessments that have longer measurement sessions including more non-functional UL use. By adding more participants and alternating between functional and non-functional UL use, we create more complexity in the model. Our second major limitation is that we only used specific activities that catered around house-work (i.e. bedmaking, kitchen, laundry activity, and shopping) and did not include any outdoor activities.

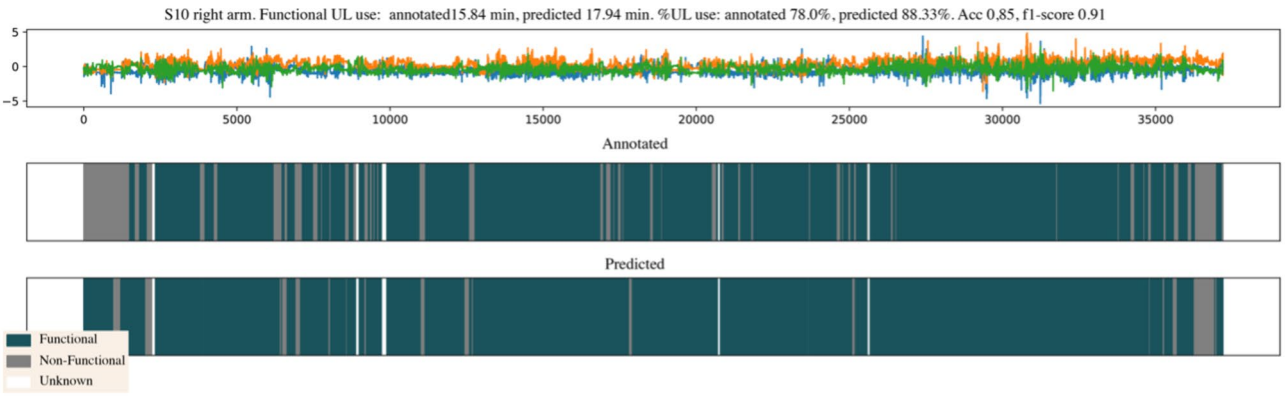

Figure 5. Example figure of the annotated versus predicted results with the corresponding accelerometer data. This person shows the largest difference between annotated %UL use and predicted %UL use. The top plot shows the acceleration signals in the x, y, and z directions. The middle plot shows the annotated labels and the bottom plot shows the predicted labels. Dark green is functional use, grey is non-functional use, and white is unknown.

As mentioned in the introduction, activities such as bus riding that induce many instances of acceleration not associated with functional UL use should be considered in future work to include more complexity in the model. Therefore, we are relatively confident that the model can distinguish UL functional use during the household tasks from non-functional UL use. However, we need more data and research that also incorporates outdoor activities.

Future research

There's a crucial need to better understand, evaluate and manage UL function in breast cancer survivors, given its significance to their well-being[41] However, before developing care pathways, accurate assessment methods for UL dysfunctions must be established. Each survivor's unique needs should be evaluated through a biopsychosocial approach to tailor treatment strategies effectively. In this, both subjective and objective measures of UL function may have value. The use of wrist-worn sensors for collecting objective data on UL function in daily life presents numerous opportunities to enhance rehabilitation outcomes. By providing continuous, detailed insights into a patient's activities, these sensors have the potential to revolutionize rehabilitation practices. Firstly, an objective assessment enables the establishment of a baseline level of UL function, offering a clear understanding of the patient's current functional status. Continuous monitoring of activity levels allows for the easy detection of both improvements and declines over time. Additionally, these sensors can identify compensatory behaviors that may hinder recovery, facilitating timely interventions. Second, with this data, it becomes possible to set realistic and personalized rehabilitation goals, complementing subjective reports with objective empirical data. Rehabilitation programs can then be tailored to address areas of limited use, ensuring a more targeted and effective approach to recovery. Last, ensuring accurate and detailed documentation of patient progress is useful for medical records. However, at this stage, objective assessment tools based on accelerometry data need further development. Our research shows that a TCN model has the potential to encompass various elements to allow for individualization of care. Just as a recent case study demonstrated the potential of a TCN model in personalizing cancer therapy[42], a similar approach could be employed for UL dysfunctions.

Therefore, the next step in research is to further optimize the current available TCN model. Future research should focus on including longer recording times and more various (outdoor) activities, such as riding the bus, that represent in- and outdoor real-life situations in order to integrate more complexity into the TCN model.

Conclusion

We can conclude that a TCN model is suited to detect functional UL use in daily life in breast cancer survivors using wrist-worn acceleration data as input. We found strong agreements between annotated data and data predicted by the TCN model, and we found an average overestimated UL functional use of only 0.71 min or 3.04% for the percentage functional UL use. Nonetheless, our results were not as good for every participant as can be seen by our limits of agreement ranging from − 11.05 to 4.93% for percentage functional UL use. Therefore, future work should consider developing a more fine-grained model, adding more participants with longer data sequences that contain more and longer sections of non-functional UL use together with data captured outdoors, in order that the model can extract meaningful representations from sequential data.

Data availability

The data that support the findings of this study are available from KU Leuven but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of KU Leuven. Requests can be sent to the corresponding author (an.degroef@kuleuven.be).

Received: 26 October 2023; Accepted: 11 July 2024
Published online: 06 August 2024

References

1. De Groef, A. *et al.* The association between upper limb function and variables at the different domains of the international classification of functioning, disability and health in women after breast cancer surgery: A systematic review. *Disabil. Rehabil.* **44**, 1176–1189 (2022).
2. Siqueira, T. C., Frágoas, S. P., Pelegrini, A., de Oliveira, A. R. & da Luz, C. M. Factors associated with upper limb dysfunction in breast cancer survivors. *Support. Care Cancer* **29**, 1933–1940 (2021).
3. De Baets, L., Vets, N., Emmerzaal, J., Devoogdt, N. & De Groef, A. Altered upper limb motor behavior in breast cancer survivors and its relation to pain: A narrative review. *Anat. Rec.* https://doi.org/10.1002/ar.25120 (2022).
4. Mallinson, T. & Hammel, J. Measurement of participation: Intersecting person, task, and environment. *Arch. Phys. Med. Rehabil.* **91**, S29–S33 (2010).
5. Lemmens, R. J., Timmermans, A. A., Janssen-Potten, Y. J., Smeets, R. J. & Seelen, H. A. Valid and reliable instruments for arm-hand assessment at ICF activity level in persons with hemiplegia: A systematic review. *BMC Neurol.* **12**, 21 (2012).
6. Tang, A., Eng, J. J. & Rand, D. Relationship between perceived and measured changes in walking after stroke. *J. Neurol. Phys. Ther.* **36**, 115–121 (2012).
7. David, A., Subash, T., Varadhan, S. K. M., Melendez-Calderon, A. & Balasubramanian, S. A framework for sensor-based assessment of upper-limb functioning in hemiparesis. *Front. Hum. Neurosci.* **15**, 667509 (2021).
8. Uswatte, G. & Hobbs Qadri, L. A behavioral observation system for quantifying arm activity in daily life after stroke. *Rehabil. Psychol.* **54**, 398 (2009).
9. Noorkõiv, M., Rodgers, H. & Price, C. I. Accelerometer measurement of upper extremity movement after stroke: A systematic review of clinical studies. *J. NeuroEng. Rehabil.* **11**, 144 (2014).
10. Dobkin, B. H. & Martinez, C. Wearable sensors to monitor, enable feedback, and measure outcomes of activity and practice. *Curr. Neurol. Neurosci. Rep.* **18**, 1–8 (2018).
11. Sequeira, S. B. *et al.* Machine learning improves functional upper extremity use capture in distal radius fracture patients. *Plast. Reconstr. Surg. Glob. Open* **10**, e4472 (2022).

12. Lum, P. S. *et al.* Improving accelerometry-based measurement of functional use of the upper extremity after stroke: Machine learning versus counts threshold method. *Neurorehabil. Neural Repair* **34**, 1078–1087 (2020).
13. Vets, N. *et al.* Assessing upper limb function in breast cancer survivors using wearable sensors and machine learning in a free-living environment. *Sensors* **23**, 6100 (2023).
14. Powers, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Preprint at http://arxiv.org/abs/2010.16061 (2020).
15. Chollet, F. *Deep Learning with Python* (Simon and Schuster, 2021).
16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
17. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
18. Lea, C., Flynn, M. D., Vidal, R., Reiter, A. & Hager, G. D. Temporal convolutional networks for action segmentation and detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165 (2017).
19. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. Preprint at http://arxiv.org/abs/1511.07122 (2016).
20. Cho, K. *et al.* Learning Phrase representations using RNN encoder-decoder for statistical machine translation. Preprint at http://arxiv.org/abs/1406.1078 (2014).
21. Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Preprint at http://arxiv.org/abs/1803.01271 (2018).
22. De Groef, A. *et al.* Prognostic factors for the development of upper limb dysfunctions after breast cancer: The UPLIFT-BC prospective longitudinal cohort study protocol. *BMJ Open* **14**, e084882 (2024).
23. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. *Biometrika* **71**, 599–607 (1986).
24. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
25. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT press, 2016).
26. Nair, N., Thomas, C. & Jayagopi, D. B. Human Activity Recognition Using Temporal Convolutional Network. In *Proc. of the 5th International Workshop on Sensor-based Activity Recognition and Interaction*, 1–8 (ACM, 2018). https://doi.org/10.1145/3266157.3266221.
27. Zhang, Y. *et al.* Can wearable devices and machine learning techniques be used for recognizing and segmenting modified physical performance test items?. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 1776–1785 (2022).
28. Filtjens, B., Vanrumste, B. & Slaets, P. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Trans. Emerg. Top. Comput.* **12**, 202–212 (2022).
29. Yang, P.-K. *et al.* Freezing of gait assessment with inertial measurement units and deep learning: Effect of tasks, medication states, and stops. *J. NeuroEng. Rehabil.* **21**, 24 (2024).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
31. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *International conference on machine learning* 448–456 (pmlr, 2015).
32. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proc. of the 27th International Conference on Machine Learning (ICML-10)*, 807–814 (2010).
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
34. Pavllo, D., Feichtenhofer, C., Grangier, D. & Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7753–7762 (2019).
35. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
36. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* https://doi.org/10.48550/arXiv.1912.01703 (2019).
37. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
38. Altman, D. G. & Bland, J. M. Measurement in medicine: The analysis of method comparison studies. *J. R. Stat. Soc. Ser. Stat.* **32**, 307–317 (1983).
39. Giavarina, D. Understanding bland altman analysis. *Biochem. Med.* **25**, 141–151 (2015).
40. Beaton, D. E., Wright, J. G., Katz, J. N., UEC Group. Development of the QuickDASH: Comparison of three item-reduction approaches. *JBJS* **87**, 1038–1046 (2005).
41. Ong, W. L. *et al.* A standard set of value-based patient-centered outcomes for breast cancer: the International Consortium for Health Outcomes Measurement (ICHOM) initiative. *JAMA Oncol.* **3**, 677–685 (2017).
42. Mencattini, A. *et al.* Deep-manager: A versatile tool for optimal feature selection in live-cell imaging analysis. *Commun. Biol.* **6**, 1–17 (2023).

Acknowledgements

We would like to acknowledge Professor Dr. Dieter Vanassche and Professor Dr. Nele Devoogdt for their contribution in obtaining funding to complete this work. This work was supported by internal fund of KU Leuven (BOF) grant number C24M/21/046.

Author contributions

Funding acquisition: LDB, ADG, AS. Patient recruitment: NV, AS. Study concept and design: JE, BF, NV, BV, LDB. Data collection and analysis: JE, BF, NV. Results interpretation: JE, BF. Supervision: BV, LDB, ADG. Writing—original draft: JE, BF. Writing—review and editing: All authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67497-6.

Correspondence and requests for materials should be addressed to A.G.