Learning Frame-Level Classifiers for Video-Based Real-Time Assessment of Stroke Rehabilitation Exercises From Weakly Annotated Datasets

Ana Rita Cóias, *Student Member, IEEE*, Min Hun Lee, Alexandre Bernardino, *Senior Member, IEEE*, Asim Smailagic, *Fellow, IEEE*, Mariana Mateus, David Fernandes, and Sofia Trapola

***Abstract*— Autonomous rehabilitation support solutions, such as virtual coaches, should provide real-time feedback to improve motor function and maintain patient engagement. However, fully annotated dataset collection for real-time exercise assessment is time-consuming and costly, posing a barrier to evaluating proposed methods. In this work, we present a novel framework that learns a frame-level classifier using weakly annotated videos for real-time assessment of compensatory motions in stroke rehabilitation exercises by generating pseudo-labels at a frame level. We consider three approaches: 1) a baseline approach that uses a source dataset to train a frame-level classifier, 2) a transfer learning approach that uses target dataset video-level labels and parameters learned from a source dataset with frame-level labels, and 3) a semi-supervised approach that leverages a target dataset video-level labels and a small set of frame-level labels. We intend to generalize to a weakly labeled target dataset with new exercises and patients. To validate the approach, we use two datasets annotated on compensatory motions: *TULE*, an existing video and frame-level labeled dataset of 15 post-stroke patients and three exercises, and *SERE*, a new dataset of 20 post-stroke patients and five exercises, created by the authors, with video-level labels and a small amount of frame-level labels. We show that a frame-level classifier trained on *TULE* does not generalize well on *SERE* ($f_1 = 72.87\%$), but our semi-supervised and transfer learning approaches achieve, respectively, $f_1 = 78.93\%$ and $f_1 = 80.47\%$. Generating pseudo-labels leads to better frame-level classification results for the target dataset than training a classifier with the source dataset (baseline). Thus, the proposed approach can simplify the customization of virtual coaches to new patients and exercises with low data annotation efforts.**

***Index Terms*— Compensatory motion patterns, dataset, pseudo-labeling, real-time motion assessment, saliency maps, stroke rehabilitation, weakly supervised learning.**

I. Introduction

INDIVIDUALS with neurological conditions (e.g., stroke) need immediate and prolonged rehabilitation therapy [1], [2] with repetitive task-oriented exercises [3], [4], [5]. Therapists assess motor function, guide exercises, and provide feedback [1], [6], [7], [8]. Due to a shortage of therapists and high rehabilitation costs [9], [10], [11], [12], patients are encouraged to exercise autonomously at home or between therapy sessions [13]. Exercising alone leads to challenges in keeping motivation and engagement, hindering recovery [9], [10], [14]. This has sparked interest in developing rehabilitation support systems, as virtual coaches (VCs). VCs should assess exercise performance and offer proper real-time feedback, helping motor function improvement by providing a personalized and pleasant therapeutic experience [15], [16]. As an extra to clinical interventions, they can enhance autonomy and independence, leading to more effective recovery over time.

Advances in Computer Vision and Machine Learning (ML) enabled automated objective assessment of impaired motor function from recorded videos [17], [18], [19], [20]. To deliver real-time feedback, VCs must assess patients' motions in real time. While ML algorithms for performance evaluation after exercise completion use video-level labels (VLL) [20], real-time assessment requires frame-level labels (FLL).

Received 6 January 2025; revised 7 July 2025; accepted 20 August 2025. Date of publication 25 August 2025; date of current version 29 August 2025. This work was supported in part by Portuguese Foundation for Science and Technology-FCT by Laboratory for Robotics and Engineering Systems (LARSyS) FCT under Grant 10.54499/LA/P/0083/2020, Grant 10.54499/UIDP/50009/2020, and Grant 10.54499/UIDB/50009/2020; in part by FCT HAVATAR Project under Grant 10.54499/PTDC/EEI-ROB/1155/2020; in part by a Ph.D. Grant 2021.05239.BD; in part by Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 Grant; and in part by the Portuguese Recovery and Resilience Plan (RRP), Center for Responsible AI under Project 62. *(Corresponding author: Ana Rita Cóias.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Alcoitão Center for Rehabilitation Medicine under Application No. CMRA2023_003.

Ana Rita Cóias and Alexandre Bernardino are with the Institute for Systems and Robotics, LARSyS, and the Instituto Superior Técnico, University of Lisbon, 1649-004 Lisbon, Portugal (e-mail: ana.coias@tecnico.ulisboa.pt; alex@isr.tecnico.ulisboa.pt).

Min Hun Lee is with Singapore Management University, Singapore 188065 (e-mail: mhlee@smu.edu.sg).

Asim Smailagic is with Carnegie Mellon University, Pittsburgh, PA 15289 USA (e-mail: asim@andrew.cmu.edu).

Mariana Mateus is with the NeuroSer Rehabilitation Center, 1600-610 Lisbon, Portugal (e-mail: mmateus@neuroser.pt).

David Fernandes and Sofia Trapola are with Alcoitão Center for Rehabilitation Medicine, 2649-506 Alcabideche, Portugal (e-mail: david.fernandes@scml.pt; sofia.trapola@scml.pt).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2025.3602548, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2025.3602548

However, collecting fully labeled datasets is time-consuming, costly, and impractical for many real-world applications [21], [22]. In addition, data labeling relies on domain experts' availability and experience.

Previous works in real-time feedback generation use **fully supervised models for real-time exercise assessment**, **relying on detailed frame-level video annotation** [23], [24], [25]. Lee and Choy [26] explored a gradient-based Explainable AI (XAI) technique to create frame-level pseudo-labels (FLPL) for compensatory motion detection. Yet, further **analysis on the usability of pseudo-labels for training fully supervised classifiers for frame-level assessment is lacking**. **Thus, real-time motion assessment from video-level annotation was not achieved yet**. **In addition, it is unclear how we can utilize an existing dataset and transfer to develop a tuned model for a new patient.**

In rehabilitation research, testing proposed methods with real patients' data is crucial. Dataset collection is a lengthy procedure, requiring therapists' availability, patient consent for personal data recording, and ethical approvals [21]. As a result, **researchers or healthy volunteers often simulate impaired motions for model evaluation** [25], [27], [28] highlighting the data collection challenges. Additionally, **existing datasets cover a limited number of motions** [29] **or provide general annotations on exercise correctness** [30].

In this work, we present a novel framework that learns a frame-level classifier (FLC) from VLL for real-time video assessment of compensatory motions (e.g., trunk tilt) in rehabilitation exercises, aiming to ease the demands of data labeling when evaluating new patients and exercises. We aim to assess in real-time compensatory motions in a newly collected weakly labeled dataset (target dataset), i.e., only with VLL and a small amount of FLL. To accomplish this, we make use of the knowledge provided by a previously collected dataset (source dataset). We consider three approaches:

1) A baseline approach that uses a source dataset to train a FLC;
2) A transfer learning approach that uses VLL of a target dataset and a threshold parameter learned from a source dataset with FLL to produce FLPL to train a FLC;
3) A semi-supervised approach that leverages the target dataset VLL and a small set of FLL to generate FLPL to train a FLC.

In an exploratory stage with the source dataset we apply our approach to determine the feasibility of our framework. For all approaches, we test the FLC on the test set of the target dataset fully labeled at a frame level. We evaluate which approach yields better results when generalizing to a weakly labeled target dataset with new exercises and patients.

Aiming to explore a broader range of motions for stroke rehabilitation, we collected a new dataset, the StrokE Rehab Exercises (*SERE*)[1], of 20 post-stroke patients performing five functional tasks (e.g., putting on socks) involving the upper limbs, trunk, and legs. We recorded the videos using a ZED Mini stereo camera and physio and occupational therapists annotated the observed compensatory motions.

[1]https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#SERE

To evaluate our approach, we use two datasets: the Three Upper Limb Exercises [20] (*TULE*) of 15 post-stroke patients executing three exercises and the newly collected *SERE*. *TULE* (source dataset) is fully labeled at a video and frame level while *SERE* (target dataset) is fully labeled at a video level but only a small set has FLL .

The baseline approach FLC achieved better True Alarm Rate (TAR=59.02%) and AUC (71.43%) and the transfer learning approach FLC provided improved f_1 score (80.47%) and less false alarms (False Alarm Rate = 19.46%).

We discuss the potential of the proposed approach to simplify the customization of VCs to new patients and exercises, reducing efforts in data labeling, and demonstrate how transferring knowledge across datasets can enhance evaluation on a new weakly labeled dataset. Therefore, our framework contributes to the development of solutions to support rehabilitation exercise training.

This work makes the following contributions:

- We present a **novel framework** that learns a FLC from VLL, easing the demands of data labeling when evaluating new patients and exercises;
- These approaches allow **real-time** video compensatory motions assessment, as VCs' feedback can be given to users at a frame-level;
- We introduce a **new dataset** of five functional tasks with 20 post-stroke patients, enabling the evaluation of our method within several approaches:
- We evaluate FLC generalization to a weakly labeled target dataset with new functional tasks and patients.

II. Related Work

A. Real-Time Exercise Automated Quantitative Assessment After Stroke

Advances in motion capture technology have enhanced the objective assessment of motion impairments [18], [19], with systems categorized into non-vision-based (e.g., inertial sensors) and vision-based solutions (marker-based and marker-free). In contrast with high-precision marker-based systems, marker-free options like MediaPipe [31] provide more convenient and affordable solutions. Kinematic analysis of body pose data is key in assessing biomechanical behavior and tracking motor function improvement [18]. In particular, joint angular motion is crucial in identifying motion limitations [17]. These advancements have stimulated research on rehabilitation exercise training support systems as VCs. VCs should interact with the user maintaining motivation and engagement in therapy while promoting motor function improvement [15], [16]. It should evaluate motion in real time to offer the user real-time feedback.

Researchers have explored real-time exercise assessment using ML and rule-based approaches. Lee et al. [23] proposed an interactive hybrid approach combining supervised ML and rule-based models for frame-level compensatory motion assessment, providing personalized feedback. Using a supervised LSTM architecture, they predicted compensatory patterns (e.g., trunk, shoulder, and head misalignments) at a frame level and employed an ensemble voting method to overcome motion boundary detection challenges. Cóias et al. [24]

developed a real-time assessment method for compensatory motions (e.g., trunk rotations) using fully supervised Neural Networks (NNs) and rule-based models. Their approach framed the problem as multilabel classification, employing two classifiers: a primary classifier to identify frames containing compensatory patterns and a secondary classifier to determine the type of compensation. Mennella et al. [25] introduced a deep learning system for evaluating exercise performance by assessing the range-of-motion (ROM) and compensatory patterns. The system consists of a ROM-classifier, a compensation-classifier, and a module to count valid/invalid motion repetitions. A rehabilitation expert developed an exercise protocol to validate the system, utilizing a dataset labeled at a frame level. However, **fully supervised methods for real-time motion assessment require extensive labeled data (e.g., at a frame level), which is costly and time-consuming to obtain**.

B. Weakly Supervised Learning Based on Feature Saliency

As AI use expands, the associated risks grow too, mainly in critical decision-making areas like healthcare [32], [33], [34]. This encourages research focus on understanding AI decision-making. XAI aims to make AI's "black box" mechanisms more interpretable and transparent, increasing user trust. Saliency maps appeared as explanations highlighting significant areas of an image that influence the model's decision [35], [36]. While primarily used for image data, adaptations of saliency maps have also been used for time-series data [37] identifying significant signal segments [38]. However, these methods often only offer qualitative evaluations of saliencies.

Several works have proposed weakly supervised solutions for semantic segmentation [39] and action recognition [40], [41], [42], [43], using saliency maps to determine object placing or action occurrence in images and videos and assigning a label to relevant pixels or frames using threshold methods. These methods work with weakly labeled data, relying on image-level or VLL denoting the existence of an object or the occurrence of an action without specifying location or timing. Similarly, Class Activation Maps (CAM) [36], [44], a variation of saliency maps, have been used to assign labels to pixels or frames, which are then used to train fully supervised models for more precise object and action location [39], [42]. By evaluating pseudo-labels and fully supervised classifier outcomes trained on them against ground-truth labels, these studies enhance the quantitative evaluation of pseudo-labels and saliency maps concerning quality and usability.

Lee and Choy [26] explored a threshold method combining a weakly supervised ML model with a gradient-based XAI technique, utilizing saliency maps to identify important frames for assessing compensatory patterns. Their goal was to advance research in XAI methods for time-series data, offering explanations for model outcomes to enhance user adoption, particularly in critical healthcare decision-making tasks. They computed saliency maps highlighting key joints and frames involved in compensatory motions, allowing them to identify when compensations occur. Researchers conducted a preliminary analysis to assess whether the saliency scores could be used for frame-level labeling by applying a threshold to the normalized aggregated joint scores at each frame. Yet, **an approach for real-time assessment relying on the training of FLC from video-level annotations, which are easier to obtain, is still lacking**. **In addition, it is unclear how we can utilize an existing dataset and transfer to develop a tuned model for a new patient. Thus, we explore taking advantage of a source dataset to train a FLC (baseline approach) and to generate FLPL (transfer learning approach).**

TABLE I
STUDIED COMPENSATORY MOTION PATTERNS

k	Labels	Compensatory Pattern
1	0/1	Shoulder abnormal/normal alignment
2	0/1	Trunk abnormal/normal alignment
3	0/1	Head abnormal/normal alignment

III. METHODS

We propose a novel framework for a weakly supervised approach that learns a FLC from VLL for real-time video compensatory motion assessment in functional tasks for stroke rehabilitation. It relies on a gradient-based technique (P-CAM) and a threshold method to generate FLPL from salient features and frames. FLPL are used to train a FLC for real-time assessment. With this approach, we aim to ease the labeling efforts to evaluate a weakly labeled dataset with new post-stroke patients and exercises.

A. Problem Definition

We consider a set of N untrimmed videos of post-stroke patients performing a functional task motion trial, $\mathbf{V} = \{v^i\}_{i=1}^N$. Each video has a set of K labels, $\mathbf{Y} = \{\mathbf{y}^i\}_{i=1}^N$, $\mathbf{y}^i \in \{0,1\}^K$ denoting the existence of the compensatory motions as described in Table I and illustrated in Figure 1. Compensation is defined by new patterns patients developed after the stroke to achieve task target [45]. Therapists specifically focus on abnormal trunk displacements (e.g., tilt and excessive flexion), head misalignment (e.g., flexion and tilt), and shoulder elevation and excessive abduction. We treat the detection of a compensatory motion pattern has a multilabel classification problem.

B. Framework for Weakly Supervised Exercise Assessment Overview

Figure 2 describes our framework to learn a FLC from VLL by generating FLPL indicating compensation occurrence in time. First, we train a VLC for video-level assessment. With a trained VLC to determine the existence of compensatory motion patterns (Table I) in a video, we perform a forward pass on the training set to generate video-level predictions (Figure 2.b) For each input video, if the predicted label denotes the absence of a compensatory pattern, we set all FLL as having a normal motion ($y_t^i = 1$). Otherwise, we compute a pseudo-label for each frame through a pseudo-labeler(Figure 2.c) The pseudo-labeler

(a) Shoulder elevation (left) and normal alignment (right). (b) Trunk tilt (left) and normal alignment (right). (c) Head flexion/rotation (left) and normal alignment (right).

Fig. 1. Common types of compensatory motion patterns.

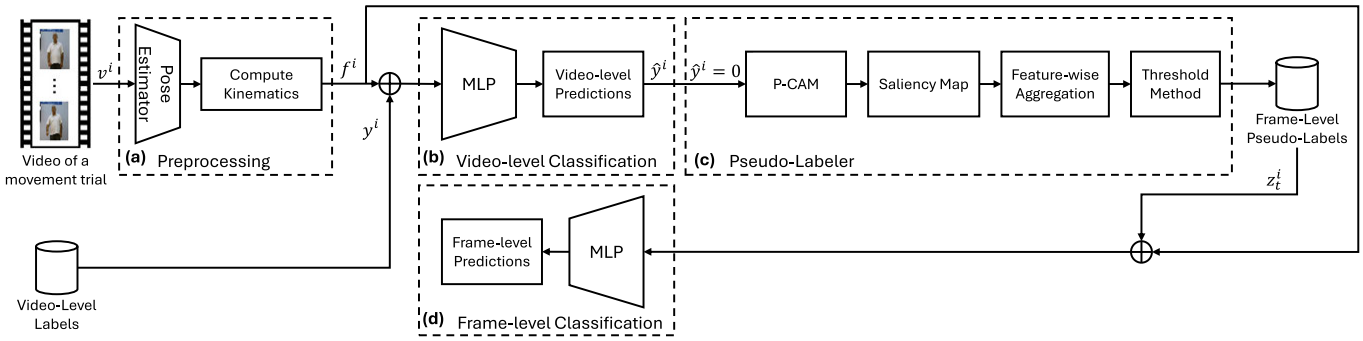


Fig. 2. Frame-level pseudo-labels (FLPL) generation framework: (a) Preprocessing step with pose estimation and calculation of kinematic variables to describe compensation; (b) forward pass in a video-level classifier to generate a prediction; (c) if compensation is detected, the Pseudo-Labeler generates FLPL by applying a threshold method to salient frames.

TABLE II
MEDIAPIPE POSE LANDMARKS

Body Joint	Abbr.	MediaPipe Joint Index
head	*hd*	0
*s*pine/trunk *s*houlder	*ss*	(11+12)/2
*l*eft *sh*oulder	sh^l	12
*l*eft *el*bow	eb^l	14
*l*eft *w*rist	wr^l	16
*r*ight *sh*oulder	sh^r	11
*r*ight *el*bow	eb^r	13
*r*ight *w*rist	wr^r	15
*s*pine/trunk *b*ase (pelvis)	*sb*	(23+24)/2
*l*eft *hip*	hp^l	24
*r*ight *hip*	hp^r	23

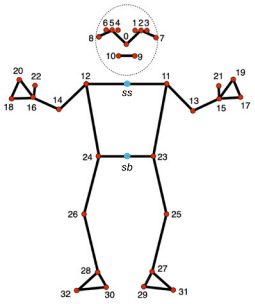


TABLE III
FEATURES DESCRIBING COMPENSATORY MOTIONS [20]

Compensatory Motion	Feature	Notation
Shoulder Abnormal Alignment	· Shoulder elevation angle	· $ja_t(sh_{init}, ss_{init}, sh)$
	· Shoulder abduction angle	· $ja_t(hp, sh, eb)$
	· Shoulder projected trajectory	· $dpt_t(sh_{init}, sh, c)$ for $c \in C$
Trunk Abnormal Alignment	· Tilted angle of the trunk	· $ja_t(ss_{init}, sb_{init}, ss)$
	· Trunk projected trajectory	· $dpt_t(ss_{init}, ss, c)$ for $c \in C$
Head Abnormal Alignment	· Head projected trajectory	· $dpt_t(hd_{init}, hd, c)$ for $c \in C$

applies a gradient-based technique (P-CAM) to each video-level predicted classes to obtain a saliency map of salient features and frames with positive gradients significant for the VLC decision. Then, we threshold the salient frames' gradients to produce FLPL. Finally, we train a fully supervised FLC with the same training data and the FLPL to achieve frame-level compensation assessment (Figure 2.d). In the following sections, we detail each step of the proposed pipeline.

C. Preprocessing

1) Pose Estimation: We use MediaPipe BlazePose,[2] as a Python library, to track post-stroke patients' motions by processing video frames, as it revealed a good alignment with the widely used Microsoft Kinect v2 [46]. MediaPipe provides real-world 3D coordinates, in meters, of 33 pose landmarks with the origin in a midpoint between the hips. Table II shows the landmarks studied. We apply a moving average filter with a window size of five frames to smooth the extracted trajectories.

2) Kinematic Variables Describing Compensation: We use a set of kinematic variables to describe the three compensatory patterns (Table I) summarized in Table III. We compute them at each timestamp for the right and left body sidesThese variables are the input features for our classification models In this work, we adopt the following notation:

- $ja_t(j_1, j_2, j_3)$ stands for Joint Angle computed among three body joints;
- $dpt_t(j_1, j_2, c)$ is the projected trajectory regarding a joint initial to current position in coordinate c;
- j specifies a joint in the set J described in Table II;
- t is the video frame number in a total of T frames;
- c denotes a coordinate in the set $C \in \{x, y, z\}$.

D. Video-Level Classification

Given the positive results in this task [20], [24] [46], we train a Multi-Layer Perceptron (MLP) classifier to assess compensatory motion patterns in a video of an exercise movement trial. Given that a video i is a sample in our training set, we use as input the kinematic variables at each timestamp for

[2]https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker
https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker/python

all video frames, i.e., $f^i \in \mathbb{R}^{TD}$, where T is the maximum number of frames in the videos, D is the number of variables, and i is the video number. To handle videos of varying lengths, we have added variables at $t = 1$ to shorter videos to match the length of the longest sequence (padding).

We implement our models using the *'Pytorch'* library [47], with parameter optimization using the Optuna framework [48]. We explored architectures with one to three fully connected layers with 16 to 4096 hidden units, with the Softmax function in the output layer for class probability calculation, ReLU activation function, and Cross Entropy Loss. We explored *'Adam'* and Stochastic Gradient Descent (*'SGD'*) optimizers with a learning rate between 0.0001 and 0.1.

E. Frame-Level Pseudo-Labeler

We generate FLPL from salient features and frames, significant for VLC decision. We outline the steps for generating the saliency map and FLPL for each video frame (Figure 2.c).

1) P-CAM: Given the video-level classification, we compute the gradients of the predicted class score, before Softmax, w.r.t. the input. The gradients reveal which input features and frames influence class prediction. We inspect positive influences on models' decisions determined by positive gradients. Negative gradients determine influences that drive the model towards the opposing class, generally caused by background information or noise [36]. Thus, we obtain a vector of gradients, $S^{i'}$, matching the shape of the input vector, given by

$$S^{i'} = \begin{cases} \frac{\partial \hat{y}^i_{score}}{\partial f^i}, & \text{if } \frac{\partial \hat{y}^i_{score}}{\partial f^i} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\hat{y}^i_{score}$ is the predicted class score for video i and f^i is the input vector.

2) Saliency Map of Features & Frames: The saliency map is created by reshaping $S^{i'}$ vector into a matrix $S^i \in \mathbb{R}^{T \times D}$. A row has the gradients for a kinematic variable, d, across all frames. A column is the gradient of each kinematic variable for a specific frame, t.

3) Frame-Wise Aggregation: From the saliency map, we perform a frame-wise aggregation of the gradients and a min-max normalization to bring aggregation results for each frame into a value in [0, 1], obtaining a frame pseudo-score by

$$s^i_t = \frac{\sum_d s_{d,t} - min(\{\sum_d s_{d,t}\}^T_{t=1})}{max(\{\sum_d s_{d,t}\}^T_{t=1}) - min(\{\sum_d s_{d,t}\}^T_{t=1})} \quad (2)$$

where $s_{d,t}$ is the gradient of feature d in frame t, and s^i_t is the computed pseudo-score for frame t from video i.

4) Threshold Method: We apply a threshold, τ, to video frame pseudo-scores to classify them as either normal or as a compensatory motion, aiming for high-quality FLPL . Using this threshold, τ, to each frame is assigned a pseudo-label, z^i_t,

$$z^i_t = \begin{cases} 1, & \text{if } \hat{y}^i = 1 \\ \mathbb{I}(s^i_t < \tau), & \text{if } \hat{y}^i = 0 \end{cases} \quad (3)$$

where $\hat{y}^i$ is the predicted class from the video-level classification and $\mathbb{I}$ is an indicator function. For a normal motion trial ($\hat{y}^i = 1$), all FLPL are set to 1. For a video with compensation ($\hat{y}^i = 0$), each frame pseudo-score s^i_t is evaluated by the condition $s^i_t < \tau$. The indicator function determines that if the condition is true, a frame pseudo-label z^i_t is set to 1 or set to 0 otherwise.

TABLE IV
REHABILITATION EXERCISES ON *TULE* AND *SERE* DATASETS AND CORRESPONDING JOINT MOTIONS

Dataset	Exer.	Description	Motions
TULE	E1	*'Bring a Cup to the Mouth'*	• Elbow Flexion
	E2	*'Switch a Light On'*	• Shoulder Flexion
	E3	*'Move a Cane Forward'*	• Elbow Extension
SERE	E1	*'Brushing Hair'*	• Shoulder flexion and elbow flexion/extension
	E2	*'Brushing Teeth'*	• Shoulder flexion and horizontal abduction/adduction and elbow flexion/extension
	E3	*'Wash the Face'*	• Elbow flexion, shoulder flexion/extension and abduction/adduction, and arm coordination
	E4	*'Put on Socks'*	• Trunk flexion and slight right/left rotation, shoulder flexion and elbow flexion/extension
	E5	*'Hip Flexion'*	• Hip flexion

F. Frame-Level Compensation Assessment

We use the training set and the FLPL to train a fully supervised FLC for real-time compensatory motion assessment, easing the need for a costly data labeling process.

We explore architectures with one to three fully connected layers with 3 to 128 hidden units, and dropout at the end of each hidden layer with a probability between 0 and 0.5. Additional implementation details are similar to those used for the VLC, described in Section III-D.

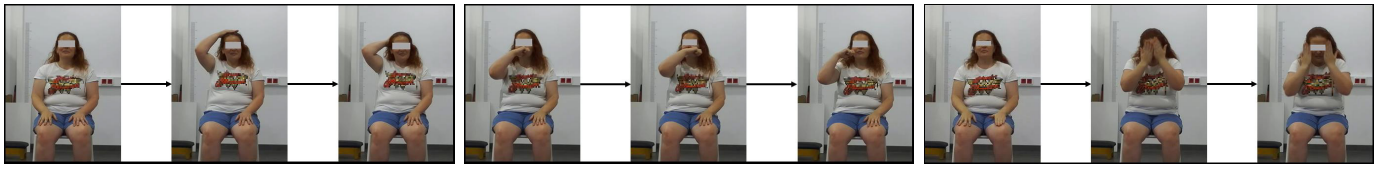
G. Datasets of Functional Tasks for Rehabilitation

1) Three Upper-Limb Exercises (TULE) Dataset: TULE [20] is a dataset of 15 post-stroke patients (63 ± 11.43 years old; 13 males and 2 females) performing three upper limb task-oriented functional tasks described in Table IV. Patients performed, on average, ten motion trials for each exercise. Data was collected with a Microsoft Kinect v2, at a frame rate of 30 fps. In the exercises, patients engaged one of their upper limbs, affected or unaffected body sides[3] Table VI summarizes the number of videos in the dataset and the ratio of videos with the three compensatory patterns. This dataset is fully labeled at a video and frame level.

2) StrokE Rehab Exercises (SERE) Dataset: SERE[4] is the newly collected dataset. Table IV describes the tasks in which post-stroke patients engage with their affected and unaffected limbs separately (E1 and E2), upper limbs simultaneously (E3), trunk (E4), and lower limbs (E5). Figure 3 illustrates the five functional tasks. Therapists suggested these tasks to simulate daily activities that are usually compromised for a post-stroke patient. These movements involve the overall functionality of the limbs and serve as an effective way to assess the quality of movement and its evolution throughout the intervention process.

[3]After stroke patients often describe weakness or loss of movement in one body side (hemiparesis).

[4]https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#SERE



(a) E1. *'Brushing Hair'* (b) E2. *'Brushing Teeth'* (c) E3. *'Wash the Face'*

(d) E4. *'Put on Socks'* (e) E5. *'Hip Flexion'*

Fig. 3. *SERE* functional tasks for rehabilitation.

a) Data collection: We recorded the videos at a frame rate of 30 fps using a ZED Mini Stereo Camera from StereoLabs[5] and the ZED Explorer framework provided by the ZED SDK, operating on a Laptop with 16 GB RAM, 11th Gen Intel(R) Core(TM) i5-11400H 2.70GHz 6 cores CPU, and NVIDIA Geforce RTX 3060 GPU. The camera was placed 0.90m above the floor and 2.5m away from the patient, who performed the tasks while seated in a chair to ensure safety.

Data collection and storage comply with the General Data Protection Regulation (GDPR). The NeuroSer executive board and the Alcoitão Center for Rehabilitation Medicine Ethics Committee and executive board revised and approved all ethical and experimental procedures and protocols. The protocol CMRA2023_003 was approved by the Alcoitão Center for Rehabilitation Medicine Ethics Committee on April 4^{th}, 2023.

b) Participants: 20 post-stroke patients (7 females and 13 males), with 62.3 ± 14.77 years old and 17.46 ± 36.67 months after the stroke event, participated on data collection and performed ten motion trials (repetitions) for each exercise after signing an informed consent authorizing data recording. Table V summarizes patients profiles. Table VI shows the total number of videos in the dataset and the ratio of videos featuring each type of compensatory motion.

c) Annotation: Physio and occupational therapists, with 9.33 ± 1.25 yeas of experience in stroke rehabilitation, assessed compensation during exercise performance and annotated the dataset concerning the presence of compensatory motion patterns, normal or abnormal joint range-of-motion, motion smoothness, and joint spasticity. Therapist made their annotations in agreement. To assess patients overall functionality, therapists applied the Stroke Rehabilitation Assessment of Movement (STREAM) measurement tool [49] (Table V). It evaluates coordination, functional mobility, and range-of-motion of the lower and upper limbs.

[5]https://www.stereolabs.com/

TABLE V
SERE DATASET PARTICIPANTS' PROFILES

Patient ID	STREAM (0-70)	Age	Sex	Affected Side	Type	Time After Stroke [years | months]
P01	36	64	M	Right	Ischemic	12.10 | 145.17
P02	49	66	M	Left	Hemorrhagic	1.33 | 16.03
P03	67	88	M	Right	Ischemic	1.16 | 13.90
P04	55	78	F	Right	Hemorrhagic	8.33 | 104.00
P05	-	70	M	Left	Ischemic	0.19 | 3.80
P06	58	61	M	Left	Ischemic	0.11 | 2.37
P07	46	55	M	Right	Hemorrhagic	0.08 | 0.87
P08	62	59	M	Left	Hemorrhagic	0.42 | 21.86
P09	-	40	F	Left	Ischemic	0.13 | 1.57
P10	62	78	F	Left	Ischemic	0.30 | 3.60
P11	43	55	F	Right	Hemorrhagic	0.41 | 4.90
P12	54	47	F	Left	Ischemic	0.25 | 3.03
P13	-	40	M	Right	Ischemic	0.46 | 5.47
P14	40	77	M	Left	Hemorrhagic	0.28 | 3.37
P15	56	72	M	Left	Ischemic	0.35 | 4.17
P16	46	75	M	Left	Ischemic	0.22 | 2.60
P17	52	36	M	Right	Ischemic	0.34 | 4.10
P18	-	43	M	Left	Ischemic	0.26 | 3.17
P19	69	64	F	Right	Ischemic	0.22 | 2.67
P20	-	78	F	Left	Hemorrhagic	0.22 | 2.60

TABLE VI
DATASETS CHARACTERISTICS

Dataset	Exercise	#videos	% of videos with compensation		
			Shoulder	*Trunk*	*Head*
TULE	E1	300	17.00	13.67	13.00
	E2	298	20.47	15.77	6.71
	E3	299	13.71	20.40	–
	All	897	–	–	–
SERE	E1	400	22.50	9.00	45.50
	E2	400	19.75	10.00	20.00
	E3	200	23.00	20.00	45.00
	E4	200	–	20.00	–
	E5	200	–	65.00	–
	All	1400	–	–	–

IV. EXPERIMENTS

A. Experimental Approaches

We use *TULE* dataset (source dataset), fully labeled at a video and frame levels (15 post-stroke patients), and *SERE* dataset (target dataset) fully labeled at a video level (20 post-stroke patients) but with only a small subset labeled at a frame level (seven post-stroke patients out of 20 - five for test and two for a validation step). Figure 4, illustrates the considered approaches to evaluate the FLC's generalization potential to new patients and exercises. In an exploratory stage, we test our framework with the fully labeled *TULE*

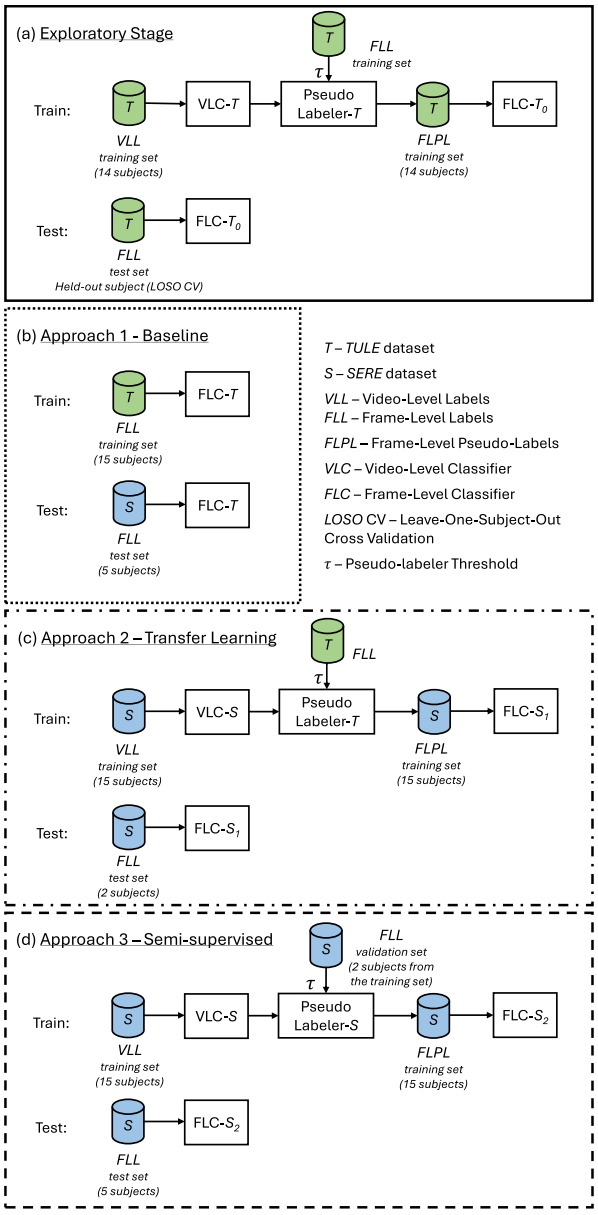


Fig. 4. Description of train and test steps of the exploratory stage with *TULE* and strategies conducted with *SERE* dataset.

dataset and determine a set of pseudo-labeler thresholds for FLPL generation. We consider three approaches to learn real-time FLCs to assess compensatory motions with *SERE*: 1) a baseline approach that uses *TULE* dataset to train a FLC, 2) a transfer learning approach that leverages *SERE* VLL and the pseudo-labeler thresholds determined in the exploratory stage to generate FLPL and train an FLC, 3) a semi-supervised learning approach that uses only *SERE* VLL and a small set of FLL for FLPL generation and train a FLC. From all approaches, the final FLCs are tested with the *SERE* test set fully labeled at a frame level. For 1) and 2) we consider that only *SERE* the test set has FLL for evaluation.

1) Exploratory Stage: In an exploratory stage with *TULE*, we train a VLC (VLC-T in Figure 4.a) for each exercise. We fine-tune the pseudo-labeler threshold, τ, to compute FLPL (Pseudo-Labeler-T in Figure 4.a), as depicted in Section III-E.4, and train a fully supervised FLC (FLC-T_0 in Figure 4.a) with the FLPL. We evaluate this stage through Leave-One-Subject-Out (*LOSO*) cross-validation against a fully supervised FLC trained with ground-truth labeling. With this exploratory stage, we inspect the loss in performance of training the FLC with FLPL instead ground-truth FLL and determine average (across exercises) pseudo-labeler thresholds, τ, for each compensatory motion pattern that are applied afterwards for FLPL generation with *SERE* in the transfer learning approach.

2) Baseline Approach: In the baseline approach, we train the FLC (FLC-T in Figure 4.b) with *TULE* and test it on *SERE* test set of five post-stroke patients, fully labeled at a frame-level. We aim to evaluate how using *TULE* to train the FLC may enable the assessment in real-time of *SERE*.

3) Transfer Learning Approach: In the transfer learning approach, we train a VLC (VLC-S in Figure 4.c) for each exercise with *SERE* training set of 15 post-stroke patients. We use the average pseudo-labeler thresholds (across *TULE* exercises), τ, for each compensatory motion determined in the exploratory stage (Table III of the supplementary materials), to produce FLPL for *SERE* training set (Pseudo-Labeler-T in Figure 4.c). The FLC (FLC-S_1 in Figure 4.c) is trained on *SERE* training set with FLPL. We perform model hyperparameter selection through *LOSO* cross-validation. Equally, we test the FLC on the held-out test set of five post-stroke patients. With this approach, we aim to inspect if the training the FLC on *SERE* with FLPL and the pseudo-labeler thresholds fine-tuned with *TULE* enhance frame-level assessment on *SERE* test set.

4) Semi-Supervised Approach: In the semi-supervised approach, we generation FLPL for *SERE* training set from the same VLC (VLC-S in Figure 4.d) and a pseudo-labeler for training a FLC. In this approach, we fine-tune the pseudo-labeler threshold with a set of two post-stroke patients labeled at a frame level, included in the training set (Pseudo-Labeler-S in Figure 4.d). Finally, we test the FLC (FLC-S_2 in Figure 4.d) using the held-out test set. With this strategy, we aim to determine the benefit of fine-tuning the pseudo-labeler τ with samples of *SERE* dataset for FLPL calculation.

The FLC is trained using the videos correctly classified by the VLC to avoid error propagation to the frame classification step. We selected post-stroke patients for validation and test sets arbitrarily, ensuring a balanced class distribution in those sets and a fair number of samples of each class for training.

B. Model Evaluation Metrics

We use f_1 score, True Alarm Rate (TAR), False Alarm Rate (FAR), and Area Under Curve (AUC) [50] to evaluate our approach. f_1 score is the harmonic mean of precision (model ability to not label as positive negative samples) and recall (model ability to identify all positive samples). TAR (or specificity) measure model ability of identifying all samples of a compensatory motion (negative samples in our problem) and is given by

$$TAR = \frac{tn}{tn + fp} \tag{4}$$

where tn and fp are the number of true negatives and false positives, respectively. FAR is the ratio of samples incorrectly labeled as negative in our problem and is given by

$$FAR = \frac{fn}{fn + tp} \tag{5}$$

where fn and tp are the number of false negatives and true positives, respectively. AUC is independent of the model's decision threshold and indicates the model's ability to differentiate between classes.

C. Pseudo-Labeler Threshold Selection

While determining a suitable pseudo-labeler threshold, τ, we prioritized minimizing false alarms (FAR $< 10\%$) to ensure a reliable real-time assessment experience while maintaining a high TAR (TAR $> 60\%$) for detecting compensation. Given a VC, post-stroke patients should keep exercising while occasional compensations go unnoticed rather than facing frequent inaccurate corrective feedback [24]. With an average 10-second movement trial and a frame rate of 30 fps, a FAR below 10% results in a minimum impact of false alarms, mainly if the feedback produced relies on a window of frames [23], [26]. We identified a threshold that optimally balances TAR and FAR, reflecting a trade-off suitable for the application. If frequently exposed to incorrect feedback, the user might loose engagement and stop exercising, which is of major importance.

For the exploratory stage, threshold selection was based on training set FLPL quality when directly compared with the ground-truth, which was available (FLL training set in Figure 4.a). The average (across exercises) thresholds for each compensatory patterns, determined in the exploratory stage, were applied in the transfer learning approach as described in Section IV-A.3. For the semi-supervised approach, threshold selection was based on the FLPL quality when directly compared with the ground-truth for the two post-stroke patients used for validation (FLL validation set in Figure 4.d). Figures 1 and 2 of the supplementary materials illustrate the TAR and FAR curves that supported our decision. Supplementary Table III presents the selected thresholds.

D. Ablation Study and Computational Latency

We perform an ablation study to evaluate the impact of lower-quality FLPL in the final FLC performance. We consider the use of a generic pseudo-labeler threshold ($\tau = 0.5$) and the threshold used in [26] ($\tau = 0.36$) to generate FLPL for all compensatory patterns, differently from the pseudo-labelers thresholds in the transfer learning and semi-supervised approaches, which are fine-tuned for each compensatory pattern on *TULE* data and *SERE* validation set, respectively.

Additionally, we measure the computational latency to assess real-time applicability. We measure the inference time per frame of pose estimation, preprocessing, and classification that composed the pipeline of real-time assessment.

V. Results

A. Evaluation of Video-Level Compensatory Motion Assessment

For both datasets, all tasks and compensatory motions, VLCs performed with a f_1 score above 95%. As we use trained VLCs to generate FLPL, as described in Section III-E, good VLC performance leads to improved FLPL quality. Models' hyperparameters are detailed in supplementary Table I.

B. Quantitative Evaluation of Frame-Level Pseudo-Labeling Quality

In the exploratory stage with *TULE* dataset, FLPL quality has an overall f_1 score above 90%, TAR over 80%, and FAR under 10%. In the semi-supervised approach, FLPL quality for the validation set achieved similar results except for E1, trunk and head compensation, E3, shoulder and head compensation, and E4. In these cases, TAR is below 60% and FAR above 10%. Detailed FLPL quality results are described in Table III of the supplementary materials.

C. Quantitative Evaluation of Frame-Level Compensatory Motion Assessment

In the exploratory stage with *TULE* dataset, our approach had an average performance (across compensatory motion patterns) of f_1 of 81.93%, 51.02% TAR, a FAR of 24.15%, and an AUC of 63.55%, which is comparable with the fully supervised baseline ($f_1 = 85.85\%$, TAR $= 57.38\%$, FAR $= 19.46\%$, 74.31% AUC).

Table VII details frame-level classification results with *SERE* dataset across experimental approaches. While the baseline approach provides better average TAR (59.02%) and AUC (71.43%) scores, the transfer learning approach achieves higher f_1 (80.47%) and lower FAR (19.46%). For E1, the semi-supervised approach has an average better performance in terms of f_1 score (82.82%), FAR (18.55%), and ability to distinguish between classes (AUC=81.40%), while the baseline was the one in which compensation detection is enhanced (TAR=63.93%). For E2, the transfer learning approach reveals increased f_1 (87.27%) and lower FAR (16.62%), but the semi-supervised approach has better compensation detection performance (TAR=44.55%) while the baseline improved AUC (76.72%). For E3, the baseline has better compensation detection performance (TAR=45.41%), while the transfer learning succeeds in other metrics (f_1=87.62, FAR=8.55%, AUC=63.45%). For E4, the baseline has increased TAR (77.29%) and AUC (64.41%), and semi-supervised and transfer learning approaches reveal improved f_1 (71.09%) and lower FAR (22.30%), respectively. For E5, the semi-supervised approach has a higher f_1 score (77.49%) and lower FAR (21.26%), whereas the transfer learning approach reveals increased TAR (70.88%) and AUC (78.99%). Models' hyperparameters are detailed in Table IV of the supplementary materials. Table V of the supplementary materials details the FLC results for all approaches, exercises, and compensatory motion patterns.

TABLE VII

RESULTS FOR FRAME-LEVEL COMPENSATION ASSESSMENT ON *SERE* TEST SET FOR THE THREE EXPERIMENTAL APPROACHES: 1) BASELINE APPROACH, 2) TRANSFER LEARNING APPROACH, AND 3) SEMI-SUPERVISED APPROACH. THE RESULTS ARE REPORTED AS AVERAGE ± STANDARD DEVIATION EVALUATED PER POST-STROKE PATIENT IN THE TEST SET AND ARE AN AVERAGE ACROSS COMPENSATORY MOTION PATTERNS

Exercises	Metrics	Approach		
		Baseline	*Transfer Learning*	*Semi-Supervised*
E1	f_1	0.7736 ± 0.1023	0.8045 ± 0.0858	**0.8282 ± 0.0774**
	TAR	**0.6393 ± 0.1449**	0.5049 ± 0.1781	0.4298 ± 0.1876
	FAR	0.2903 ± 0.1177	0.2325 ± 0.1450	**0.1855 ± 0.1306**
	AUC	0.7791 ± 0.0825	0.7965 ± 0.0623	**0.8140 ± 0.0566**
E2	f_1	0.8046 ± 0.0665	**0.8727 ± 0.0754**	0.8702 ± 0.0764
	TAR	0.4281 ± 0.1015	0.4408 ± 0.0870	**0.4455 ± 0.0752**
	FAR	0.2608 ± 0.1153	**0.1662 ± 0.0991**	0.1708 ± 0.1100
	AUC	**0.7672 ± 0.0382**	0.7021 ± 0.0360	0.7420 ± 0.0200
E3	f_1	0.7695 ± 0.1043	**0.8762 ± 0.1038**	0.7625 ± 0.1277
	TAR	**0.4541 ± 0.0806**	0.2447 ± 0.0794	0.2219 ± 0.1220
	FAR	0.2932 ± 0.1279	**0.0855 ± 0.0810**	0.2499 ± 0.1511
	AUC	0.6037 ± 0.0425	**0.6345 ± 0.0413**	0.4870 ± 0.0807
E4	f_1	0.5245 ± 0.2806	0.7092 ± 0.3347	**0.7109 ± 0.3477**
	TAR	**0.7729 ± 0.0757**	0.3528 ± 0.0879	0.3856 ± 0.0567
	FAR	0.5441 ± 0.2229	**0.2230 ± 0.1227**	0.2604 ± 0.1842
	AUC	**0.6441 ± 0.0828**	0.4894 ± 0.1082	0.4981 ± 0.1360
E5	f_1	0.7714 ± 0.0681	0.7610 ± 0.0645	**0.7749 ± 0.0680**
	TAR	0.6564 ± 0.0996	**0.7088 ± 0.1070**	0.6471 ± 0.1047
	FAR	0.2211 ± 0.0474	0.2656 ± 0.0359	**0.2126 ± 0.0457**
	AUC	0.7772 ± 0.0665	**0.7899 ± 0.0603**	0.7844 ± 0.0655
All Exercises	f_1	0.7287 ± 0.1244	**0.8047 ± 0.1328**	0.7893 ± 0.1394
	TAR	**0.5902 ± 0.1005**	0.4504 ± 0.1079	0.4260 ± 0.1092
	FAR	0.3219 ± 0.1262	**0.1946 ± 0.0967**	0.2158 ± 0.1243
	AUC	**0.7143 ± 0.0625**	0.6825 ± 0.0618	0.6651 ± 0.0737

TABLE VIII

AVERAGE RESULTS FOR THE ABLATION STUDY EVALUATING THE IMPACT OF LOWER-QUALITY FLPL IN FLC PERFORMANCE BY USING A GENERIC PSEUDO-LABELER THRESHOLD ($\tau = 0.5$) AND THE THRESHOLD FROM [26] ($\tau = 0.36$) TO GENERATE FLPL. THE RESULTS ARE REPORTED AS AVERAGE ± STANDARD DEVIATION

Metrics	Approach			
	$\tau = 0.5$	$\tau = 0.36$ [26]	*Transfer Learning*	*Semi-Supervised*
f_1	0.7222 ± 0.1516	0.8024 ± 0.1494	**0.8047 ± 0.1328**	0.7893 ± 0.1394
TAR	0.2092 ± 0.0581	0.2375 ± 0.0986	**0.4504 ± 0.1079**	0.4260 ± 0.1092
FAR	0.1845 ± 0.0640	**0.1264 ± 0.0831**	0.1946 ± 0.0967	0.2158 ± 0.1243
AUC	0.5745 ± 0.0576	0.6417 ± 0.0688	**0.6825 ± 0.0618**	0.6651 ± 0.0737

1) Ablation Study: For further evaluation of the transfer learning and semi-supervised approaches, we assess the impact of lower-quality FLPL in the final FLC performance, by using a generic threshold for FLPL generation ($\tau = 0.5$ and, from [26] $\tau = 0.36$). Table VIII, shows average results (across exercises and compensatory patterns). Detailed results can be found in the supplementary materials.

2) Computational Latency of Real-Time Assessment: To assess the feasibility of real-time assessment, we measured the computational latency.[6] The pose estimation step averages 71.44ms for a 1920×1080 video frame. Preprocessing, which involves filtering and calculating kinematic variables, takes 0.54ms per frame, and frame classification takes 0.74ms per frame. This results in an average total inference time of 73ms, suitable for a real-time application.

D. Qualitative Evaluation of Saliency Maps

Figure 5 shows an example of a motion trial of a patient describing shoulder compensation and the saliency map of salient features and frames (Figure 5.a). Shoulder compensation is visible through joint markers (Figure 5.b) and image differencing (Figure 5.c). The saliency map captures the frames where the compensation occurs, along with salient shoulder elevation angle and projected trajectories in x and y. We can also observe salient regions where the motion has ended (false saliency) and regions of compensation that are not salient (partial saliency).

[6]Laptop with 16 GB RAM, 11th Gen Intel(R) Core(TM) i5-11400H 2.70GHz 6 cores CPU, and NVIDIA Geforce RTX 3060 GPU.

VI. DISCUSSION

A. Pseudo-Labeler Threshold & Pseudo-Labels Quality

We selected pseudo-labeler thresholds, τ, that reflects a suitable trade-off between TAR and FAR when envisaging the real-world application of our method. We prioritized minimum false alarms (FAR<10%) while maintaining a high TAR (TAR>60%) for compensation detection.

With *TULE* dataset, we evaluate FLPL quality for the training set. With *SERE*, we only evaluate FLPL quality of the validation set of two post-stroke patients, in the semi-supervised approach. The frame-level labeling of two post-stroke patients represented the minimum labeling effort established to gather sufficient information for fine-tuning the pseudo-labeler threshold. Therefore, the reported results of FLPL quality on *SERE*, in Section V-B, reveal a much lower TAR (TAR< 60%) and higher FAR (FAR> 10%) than the FLPL results on *TULE* (TAR> 90% and FAR< 10%). Additionally, Figures 1 and 2 of the supplementary materials, show that the TAR and FAR curves, that support our pseudo-labeler threshold selection, have a steady progression on *TULE* and high variability on *SERE*.

This suggests two potential factors: suboptimal selection of the held-out validation set and inaccuracies in the saliency maps used for FLPL generation. Also, two post-stroke patients are not representative enough to draw solid conclusions about FLPL quality for the entire training set, as we might have performance outliers. In future work, we plan to determine the minimum number of subjects used for pseudo-labeler threshold selection, leading to a more reliable threshold. Additionally, we plan to investigate the effect of fine-tuning a threshold for each post-stroke patient individually, through an adaptive FLPL generation approach, on frame-level outcomes and model adaptability to new patients. Furthermore, techniques to reduce saliency maps noise and label refinement might improve pseudo-labels quality.

Recent follow-up work [51] applies our framework and explores the feasibility of other models for VLC and subsequent FLPL generation to train a FLC.

B. Quantitative Evaluation of Frame-Level Compensatory Motion Assessment

The baseline approach provides better average (across exercises and compensatory patterns) TAR (59.02%) and AUC (71.43%) scores while the transfer learning approach achieves higher f_1 (80.47%) and lower FAR (19.46%). A higher f_1 score is associated with fewer false alarms. Fine-tuning the pseudo-labeler τ with *TULE* (transfer learning approach) leads to a reduced number of false alarms. Conversely, using a FLC trained with *TULE* (baseline approach) provided a solution with an improved ability to distinguish classes (higher AUC)

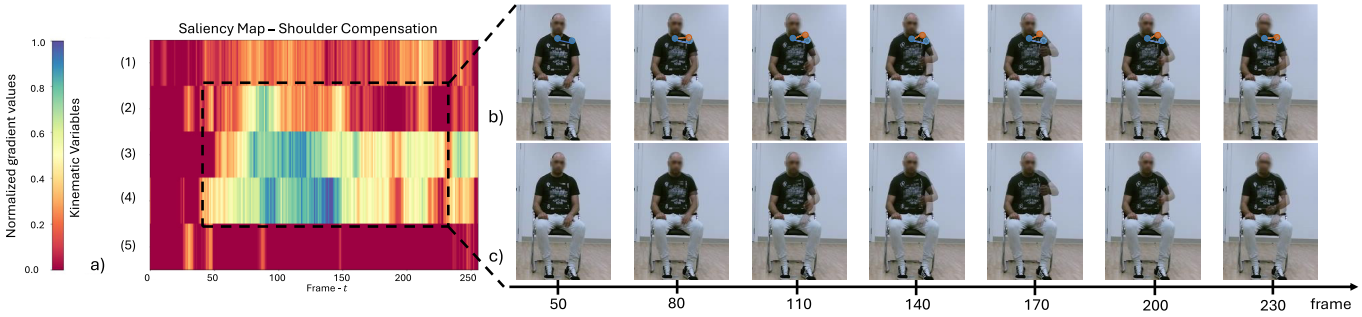


Fig. 5. Salient kinematic variables - shoulder (1) abduction angle, (2) elevation angle, and projected trajectories in (3) *x*, (4) *y*, and (5) *z* - and frames for shoulder compensation detection for an example of a motion trial from *TULE* dataset, E1. a) is the saliency map generated as detailed in Section III-E.2. b) shows the difference of shoulder initial and current positions with joint markers and c) shows the motion the trails of the moving shoulder through frame accumulation.

and identify compensation with greater precision (higher TAR) but with more average false alarms, which are reflected in a lower f_1 score value. These findings show that our transfer learning ($f_1 = 80.47\%$) and semi-supervised ($f_1 = 78.93\%$) approaches generalize better on *SERE* test set than the baseline ($f_1 = 72.87\%$) approach, mainly when we emphasize the importance of a reduced number of false alarms when applying these approaches to a VC to support rehabilitation - avoid offering frequent inaccurate corrective feedback. The positive results show that the proposed approach has the potential of simplifying the customization of VCs to new patients and exercises, reducing efforts in data labeling. We demonstrate that by producing FLPL from VLL for a target dataset (*SERE*) and leveraging the knowledge provided by a source dataset (*TULE*) we can achieve real-time (frame-level) compensatory motion assessment on the target dataset.

Nonetheless, the FLC had lower TAR for E2, E3 and E4 due to unsatisfactory TAR assessing trunk and head compensation (Table V in the supplementary materials). The less desirable results might be due to FLPL noise and saliency maps' inaccuracies for FLPL generation. Also, reduced data samples of different motions impact FLC performance and generalization to new patients.

From this point forward, although compensation is possible to access at a frame level, assessing a set of frames instead of a single frame might enhance our results as it captures motion transitions over time, leading to reduced noise, improving accuracy and generalization across motions. On another note, previous works indicate that training with both clean fully labeled data and weakly labeled data results in better performance and generalization ability instead of only using clean data for validation [52]. Also, the exploitation of methods for pseudo-labels refinement might improve the outcomes of the FLC trained with FLPL [42]. In the future, we aim to explore other weakly supervised techniques and pipelines an compare with the one proposed in this paper.

1) Ablation Study: Table VIII, shows the FLC results from using a generic pseudo-labeler threshold ($\tau = 0.5$) and the threshold from [26] ($\tau = 0.36$) against the proposed transfer learning and semi-supervised approaches. The results demonstrate that lower quality FLPL strongly impacts FLC performance in distinguishing classes (lower AUC) and lower TAR, validating the significance of the applied approaches in achieving good FLC performance.

C. Qualitative Evaluation of Saliency Maps

Figure 5 shows an example of a motion trial in which a patient performs compensation. It displays a saliency map with salient kinematic variables and frames and a sequence of video frames in which shoulder compensation is observed. The saliency map provides us with insights about VLC decision [53]. We determine when compensation occurs and that shoulder elevation angle and displacements in x and y are significant for model decision. Additionally, there are observable inaccuracies in the saliency map (partial and false saliency), which might impact FLPL quality. Methods to overcome saliency inaccuracies, such as noise reduction, and label refinement approaches [39] might improve FLPL quality, leading to enhanced outcomes in the frame-level classification step. In future work, we aim to exploit saliency maps and determine how the information extracted from them can be useful for therapists, increasing their performance in the clinical decision-making process.

VII. Conclusion

In this work, we present a novel framework for a weakly supervised learning approach that learns a frame-level classifier (FLC) from video-level labels VLL for real-time compensation assessment on stroke rehabilitation functional tasks by generating frame-level pseudo-labels (FLPL) to train a frame-level classifier (FLC). We enable real-time video compensatory motion assessment, allowing virtual coaches (VCs) to provide patients with feedback at a frame level. Aiming to explore new motions, we collected a new dataset, the StrokE Rehab Exercises (*SERE*), of videos of 20 post-stroke patients performing five functional tasks for rehabilitation, which is weakly labeled. With *SERE* and a previously available dataset, *TULE*, we evaluate our method under several experimental approaches: 1) baseline approach, 2) transfer learning approach, and 3) semi-supervised approach. We evaluate which achieves better performance on *SERE* test set, testing FLC generalization ability to new patients and exercises.

Our transfer learning ($f_1 = 80.47\%$) and semi-supervised ($f_1 = 78.93\%$) approaches generalize better to the new target

weakly labeled dataset (*SERE*) than the baseline approach ($f_1 = 72.87\%$) in which the frame-level classifier (FLC) is trained fully with a source fully labeled dataset (*TULE*). This analysis shows the great potential of weakly supervised motion impairment assessment relying only on video-level annotations, leveraging saliency maps information, easing the need for detailed labeling, which is harder to obtain due to costs, process length, and expert availability.

Appendix

More study details are available in the Supplementary Materials.

Acknowledgment

The authors would like to thank the NeuroSer Rehabilitation Center and Alcoitão Center for Rehabilitation Medicine for enabling the collection of the SERE dataset and for the advice provided by their teams. In addition, they would like to thank all the participants for accepting the invitation to voluntarily participate in this study.

References

[1] B. Semenk et al., "An evidence based occupational therapy toolkit for assessment and treatment of the upper extremity post stroke," *Screening*, vol. 4, pp. 1–4, Jan. 2015.

[2] E. Lynch, S. Hillier, and D. Cadilhac, "When should physical rehabilitation commence after stroke: A systematic review," *Int. J. Stroke*, vol. 9, no. 4, pp. 468–478, Jun. 2014.

[3] M. Rensink, M. J. Schuurmans, E. Lindeman, and T. B. Hafsteinsdøttir, "Task-oriented training in rehabilitation after stroke: Systematic review," *J. Adv. Nursing*, vol. 65, no. 4, pp. 737–754, 2009.

[4] E. J. Schneider, L. Ada, and N. A. Lannin, "Extra upper limb practice after stroke: A feasibility study," *Pilot Feasibility Stud.*, vol. 5, no. 1, pp. 1–7, Dec. 2019.

[5] S. A. Billinger et al., "Physical activity and exercise recommendations for stroke survivors: A statement for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 45, no. 8, pp. 2532–2553, Aug. 2014.

[6] I. Serrada, M. N. McDonnell, and S. L. Hillier, "What is current practice for upper limb rehabilitation in the acute hospital setting following stroke? A systematic review," *NeuroRehabilitation*, vol. 39, no. 3, pp. 431–438, Aug. 2016.

[7] D. J. Gladstone, C. J. Danells, and S. E. Black, "The fugl-meyer assessment of motor recovery after stroke: A critical review of its measurement properties," *Neurorehabilitation Neural Repair*, vol. 16, no. 3, pp. 232–240, Sep. 2002.

[8] D. M. Morris, G. Uswatte, J. E. Crago, E. W. Cook, and E. Taub, "The reliability of the wolf motor function test for assessing upper extremity function after stroke," *Arch. Phys. Med. Rehabil.*, vol. 82, no. 6, pp. 750–755, Jun. 2001.

[9] K. L. Meadmore, E. Hallewell, C. Freeman, and A.-M. Hughes, "Factors affecting rehabilitation and use of upper limb after stroke: Views from healthcare professionals and stroke survivors," *Topics Stroke Rehabil.*, vol. 26, no. 2, pp. 94–100, Feb. 2019.

[10] T. M. Damush, L. Plue, T. Bakas, A. Schmid, and L. S. Williams, "Barriers and facilitators to exercise among stroke survivors," *Rehabil. Nursing*, vol. 32, no. 6, pp. 253–262, 2007.

[11] A. S. Pollock, L. Legg, P. Langhorne, and C. Sellars, "Barriers to achieving evidence-based stroke rehabilitation," *Clin. Rehabil.*, vol. 14, no. 6, pp. 611–617, Dec. 2000.

[12] K. E. Watkins, W. M. M. Levack, F. A. Rathore, and E. J. C. Hay-Smith, "Challenges in applying evidence-based practice in stroke rehabilitation: A qualitative description of health professional experience in low, middle, and high-income countries," *Disability Rehabil.*, vol. 46, no. 16, pp. 3577–3585, Jul. 2024.

[13] K. Peek, R. Sanson-Fisher, L. Mackenzie, and M. Carey, "Interventions to aid patient adherence to physiotherapist prescribed self-management strategies: A systematic review," *Physiotherapy*, vol. 102, no. 2, pp. 127–135, Jun. 2016.

[14] N. Maclean, "Qualitative analysis of stroke patients' motivation for rehabilitation," *BMJ*, vol. 321, no. 7268, pp. 1051–1054, Oct. 2000.

[15] D. Siewiorek, A. Smailagic, and A. Dey, "Architecture and applications of virtual coaches," *Proc. IEEE*, vol. 100, no. 8, pp. 2472–2488, Aug. 2012.

[16] T. G. Weimann, H. Schlieter, and A. B. Brendel, "Virtual coaches: Background, theories, and future research directions," *Bus. Inf. Syst. Eng.*, vol. 64, no. 4, pp. 515–528, Aug. 2022.

[17] A. Ozturk, A. Tartar, B. Ersoz Huseyinsinoglu, and A. H. Ertas, "A clinically feasible kinematic assessment method of upper extremity motor function impairment after stroke," *Measurement*, vol. 80, pp. 207–216, Feb. 2016.

[18] M. A. Murphy, C. Willén, and K. S. Sunnerhagen, "Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass," *Neurorehabilitation Neural Repair*, vol. 25, no. 1, pp. 71–80, Jan. 2011.

[19] E. V. Olesh, S. Yakovenko, and V. Gritsenko, "Automated assessment of upper extremity movement impairment due to stroke," *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e104487.

[20] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. I. Badia, "Learning to assess the quality of stroke rehabilitation exercises," in *Proc. 24th Int. Conf. Intell. User Interfaces*, Mar. 2019, pp. 218–228.

[21] F. Lanotte, M. K. O'Brien, and A. Jayaraman, "AI in rehabilitation medicine: Opportunities and challenges," *Ann. Rehabil. Med.*, vol. 47, no. 6, pp. 444–458, Dec. 2023.

[22] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," 2022, *arXiv:2203.04291*.

[23] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. Badia, "Towards personalized interaction and corrective feedback of a socially assistive robot for post-stroke rehabilitation therapy," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 1366–1373.

[24] A. R. Cóias, M. H. Lee, and A. Bernardino, "A low-cost virtual coach for 2D video-based compensation assessment of upper extremity rehabilitation exercises," *J. NeuroEngineering Rehabil.*, vol. 19, no. 1, pp. 1–16, Dec. 2022.

[25] C. Mennella, U. Maniscalco, G. D. Pietro, and M. Esposito, "A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions," *Comput. Biol. Med.*, vol. 166, Nov. 2023, Art. no. 107485.

[26] M. Hun Lee and Y. Jing Choy, "Exploring a gradient-based explainable AI technique for time-series data: A case study of assessing stroke rehabilitation exercises," 2023, *arXiv:2305.05525*.

[27] A. Vakanski, H.-P. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, p. 2, Jan. 2018.

[28] R. Aguilar-Ortega et al., "UCO physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises," *Sensors*, vol. 23, no. 21, p. 8862, Oct. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/21/8862

[29] E. Dolatabadi et al., "The Toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy," in *Proc. 11th EAI Int. Conf. Pervasive Comput. Technol. Healthcare*, May 2017, pp. 375–381.

[30] A. Miron, N. Sadawi, W. Ismail, H. Hussain, and C. Grosan, "IntelliRehabDS (IRDS)—A dataset of physical rehabilitation movements," *Data*, vol. 6, no. 5, p. 46, Apr. 2021.

[31] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.

[32] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107413.

[33] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. Bermúdez I Badia, "A human-AI collaborative approach for clinical decision making on rehabilitation assessment," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–14.

[34] T. Isobe and Y. Okada, "Rehabilitation XAI to predict outcome with optimal therapies," in *Proc. 9th Int. Conf. Artif. Intell. Mobile Services*, 2020, pp. 127–139.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[37] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (XAI) on Time-Series data: A survey," 2021, *arXiv:2104.00950*.

[38] S. D. Goodfellow, A. J. Goodwin, R. Greer, P. C. Laussen, M. Mazwi, and D. Eytan, "Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings," in *Proc. Mach. Learn. Healthcare Conf.*, 2018, pp. 83–101.

[39] T. Chen, Z. Mai, R. Li, and W.-L. Chao, "Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation," 2023, *arXiv:2305.05803*.

[40] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Process. Lett.*, vol. 27, pp. 1705–1709, 2020.

[41] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[42] Q. Yu and K. Fujiwara, "Frame-level label refinement for skeleton-based weakly-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 3322–3330.

[43] J. Zhou, L. Huang, L. Wang, S. Liu, and H. Li, "Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23003–23012.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[45] M. F. Levin, J. A. Kleim, and S. L. Wolf, "What do motor 'recovery' and 'compensation' mean in patients following stroke?" *Neurorehabilitation Neural Repair*, vol. 23, no. 4, pp. 313–319, May 2009.

[46] A. R. Cóias, M. H. Lee, A. Bernardino, and A. Smailagic, "Skeleton tracking solutions for a low-cost stroke rehabilitation support system," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Sep. 2023, pp. 1–6.

[47] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–4.

[48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[49] S. Ahmed, N. E. Mayo, J. Higgins, N. M. Salbach, L. Finch, and S. L. Wood-Dauphinée, "The stroke rehabilitation assessment of movement (STREAM): A comparison with other measures used to evaluate effects of stroke and rehabilitation," *Phys. Therapy*, vol. 83, no. 7, pp. 617–630, Jul. 2003.

[50] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[51] G. Mesquita, A. R. Cóias, A. Dubrawski, and A. Bernardino, "Frame-level real-time assessment of stroke rehabilitation exercises from video-level labeled data: Task-specific vs. foundation models," 2025, *arXiv:2506.03752*.

[52] D. Zhu, X. Shen, M. Mosbach, A. Stephan, and D. Klakow, "Weaker than you think: A critical look at weakly supervised learning," 2023, *arXiv:2305.17442*.

[53] M. H. Lee and C. J. Chew, "Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. 2, pp. 1–22, Sep. 2023.