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 functions 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 for real-time video assessment of compensatory motions in stroke rehabilitation exercises using weakly annotated videos. 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 the 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 with labels 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 $f_1 = 80.47\%$. 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.**

This paragraph of the first footnote will contain the date on which you submitted your paper for review.
This work was supported by the Portuguese Foundation for Science and Technology - FCT by LARSyS FCT funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020), FCT HAVATAR Project (DOI: 10.54499/PTDC/EEI-ROB/1155/2020), and a Ph.D. grant (2021.05239.BD), and partially supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 grant.
This work involved human subjects. Approval of all ethical and experimental procedures and protocols was granted by the NeuroSer executive board and by the Alcoitão Center for Rehabilitation Medicine Ethics Committee and executive board. All data collection participants signed an informed consent.

A. R. Cóias and A. Bernardino are with the Institute for Systems and Robotics, LARSyS, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal (e-mail: ana.coias@tecnico.ulisboa.pt; alex@isr.tecnico.ulisboa.pt).
M. H. Lee, is with Singapore Management University, Singapore, Singapore. (e-mail: mhlee@smu.edu.sg).
A. Smailagic is with Carnegie Mellon University, Pittsburgh, PA, USA (e-mail: asim@andrew.cmu.edu).
M. Mateus is with NeuroSer rehabilitation center, Lisbon, Portugal (e-mail: mmateus@neuroser.pt).
D. Fernandes and S. Trapola are with the Alcoitão Center for Rehabilitation Medicine, Alcabideche, Portugal (e-mail: david.fernandes@scml.pt; sofia.trapola@scml.pt).

I. Introduction

INDIVIDUALS with neurological conditions (e.g., stroke) need immediate and prolonged rehabilitation therapy [1], [2] with repetitive task-oriented exercises [3]–[5]. Therapists assess motor function, guide exercises, and provide feedback [1], [6]–[8]. Due to a shortage of therapists and high rehabilitation costs [9]–[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 training 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].

Advances in Computer Vision and Machine Learning (ML) enabled automated objective assessment of impaired motor function from recorded videos [17]–[20]. To deliver real-time feedback, VCs must assess patients' motions in real time. While ML algorithms for post-exercise performance evaluation use video-level (VL) annotations [20], real-time assessment requires frame-level labels (FLL). 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 [23]–[25]. Lee *et al.* [26] explored a gradient-based Explainable AI (XAI) technique to create frame-level pseudo-labels (FLPL)

for compensatory motion detection. Yet, further analysis on pseudo-label usability for training fully supervised classifiers for frame-level (FL) assessment is lacking. Researchers have used Class Activation Maps (CAM) for pixel label assignment, improving outcomes in semantic segmentation tasks when using assigned labels for model training [27]. For action recognition, researchers learned video clip-level pseudo-labels, leveraging attention scores for action localization [28]–[30].

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], [31], [32], highlighting the data collection challenges. Additionally, existing datasets cover a limited number of motions [33] or provide general annotations on exercise correctness [34], limiting their utility.

In this work, we present a novel framework that learns a frame-level classifier (FLC) from video-level labels (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. Following Lee *et al.* [26] work, we create FLPL by applying a threshold method to salient frames (pseudo-labeler). Pseudo-labels are derived from salient features and frames with positive gradients, significant for VL assessment with a video-level classifier (VLC). Given an input video, we compute the gradients of the predicted class score w.r.t. the input, creating a saliency map (CAM). Through feature score aggregation and a threshold method, we produce FLPL, which are then used to train a fully supervised FLC. We consider three approaches:

1) A baseline approach that uses a source dataset to train a frame-level classifier (FCL);
2) A transfer learning approach that uses video-level labels of a target dataset and thresholds learned from a source dataset with frame-level labels to produce frame-level pseudo-labels (FLPL) to train a FLC;
3) A semi-supervised approach that leverages the target dataset video-level labels and a small set of frame-level labels to generate FLPL to train a FLC.

For all approaches, we test the FLC on the target dataset test set 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 video dataset, the StrokE Rehab Exercises (*SERE*), of 20 post-stroke patients performing five functional motions (e.g., putting on socks) involving the upper limbs, trunk, and legs. We recorded the videos using a ZED Mini stereo camera. Physio and occupational therapists annotated the observed compensatory motions.

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 (validation set).

Our VLCs achieved a f_1 score above 95% from *Leave-One-Subject-Out* (*LOSO*) cross-validation for both datasets. In the baseline approach, the FLC achieved a 72.87% f_1 score, 59.02% TAR, and 32.19% FAR, on *SERE* test set. In the transfer learning approach, the FLC achieved 80.47% of f_1 score, 45.04% of TAR, and 19.46% of FAR. For the semi-supervised approach, pseudo-labels quality for the validation set achieved 70.54% f_1 score. The FLC achieved 78.93% of f_1, 42.60% of TAR, and 23.29% of FAR. We show that FLC 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 $f_1 = 80.47\%$. We discuss the potential of the proposed approach to simplify the customization of virtual coaches 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.

This work makes the following contributions:

- We present a **novel framework** that learn a frame-level classifier (FLC) from video-level labels (VLL) by thresholding feature positive gradients from a CAM-based technique to create frame-level pseudo-labels (FLPL) to train a FLC, easing the demands of data labeling when evaluating new patients and exercises;
- These approaches allow **real-time** video compensatory motions assessment, as virtual coaches feedback can be given to users at a frame-level;
- We introduce a **new dataset** of five functional exercises 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 exercises 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 [35] 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 Virtual Coaches (VCs). VCs should interact with the user maintaining motivation and engagement in therapy while promoting motor function improvement [15], [16]. The VC 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 FL 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 require extensive labeled data, 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 [36]–[38]. 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 [39], [40]. While primarily used for image data, adaptations of saliency maps have also been used for time-series data [41] identifying significant signal segments [42]. However, these methods often only offer qualitative evaluations of saliencies.

Several works have proposed weakly supervised solutions for semantic segmentation [27] and action recognition [28], [30], [43], [44], 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 VL labels denoting the existence of an object or the occurrence of an action without specifying location or timing. Similarly, Class Activation Maps (CAM) [40], [45], 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 [27], [28]. 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 *et al.* [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 FL labeling by applying a threshold to the normalized aggregated joint scores at each frame. Yet, an approach for FL assessment relying on VL annotations, which are easier to obtain, is still lacking, and pseudo-labels quality and usability assessment is missing.

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 exercises for stroke rehabilitation. It relies on a CAM-based technique and a threshold method to generate FLPL from salient features and frames. FLPL are used to train a FLC for real-time (at a frame level) 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 exercise motion trial, $\mathbf{V} = \{v^i\}_{i=1}^N$. Each video has a set of labels, $\mathbf{Y} = \{y^i\}_{i=1}^N$, $y^i = \{0,1\}$ denoting the existence of 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 [46]. Therapists specifically focus on abnormal trunk displacements (e.g., trunk tilt and excessive flexion), head misalignment (e.g., flexion and tilt), and shoulder elevation and excessive abduction.

TABLE I: Studied compensatory motion patterns.

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

B. Framework for Weakly Supervised Exercise Assessment Overview

Figure 2 describes our framework to generate FLPL indicating compensation occurrence in time. With a trained, fully supervised, binary VLC to determine the existence of compensation in a video of a movement trial for each compensatory pattern (Table I), we perform a forward pass on the training set to generate VL 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 (Figure 2.c). We compute the gradients of the predicted class score w.r.t. the input to obtain a saliency map of salient features and frames (CAM) significant for the VLC decision. Next, we aggregate and normalize the positive gradients for each feature per frame to have a score for each frame. Then, we fine-tune a threshold to produce FLPL. Finally, we train a fully supervised FLC with the same training data and the

(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.

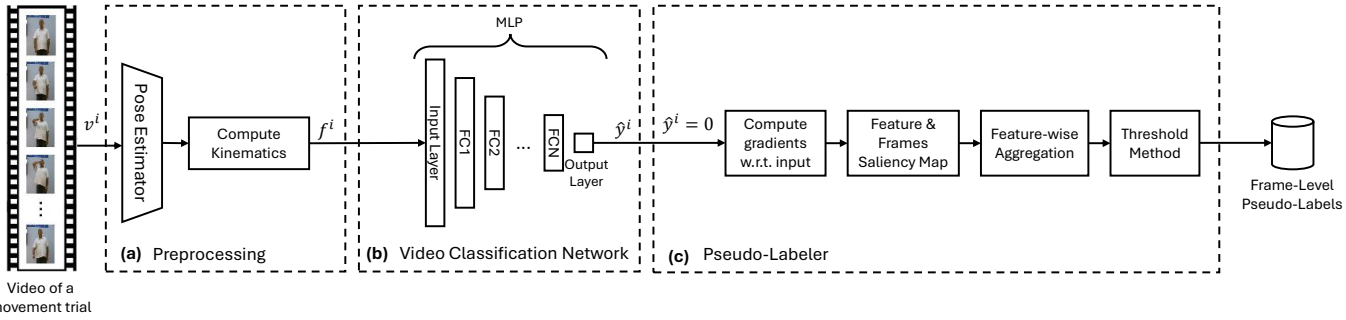


Fig. 2: Frame-level pseudo-labels (FLPL) generation framework: (a) Preprocessing step with pose estimation and calculation of kinematic variables to describe compensation; (b) forward pass in a video-level (VL) classifier to generate a VL prediction; (c) if compensation is detected, the Pseudo-labeler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

pseudo-labels to achieve FL compensation assessment. In the following sections, we describe the subsequent procedures:

- Preprocessing (Figure 2.a)
  - Pose estimation;
  - Kinematic variables describing compensation;
- Video-level (VL) compensation assessment (Figure 2.b);
- Frame-level pseudo-labels (FLPL) generation (Figure 2.c)
  - Saliency maps of features & frames;
  - Frame-level (FL) pseudo-scores;
  - FLPL generation by applying a threshold method;
- FL compensation assessment.

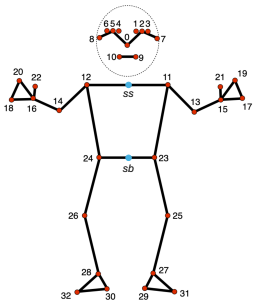
C. Preprocessing

1) Pose Estimation: We use MediaPipe BlazePose [1], 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 [47]. 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: Due to their relevance for impairment assessment, we use joint angles and trajectories to describe the three compensatory patterns (Table I) [17], [18], [20]. Table III summarizes the kinematic variables (features). We compute them at each timestamp for the affected and unaffected body sides[2]. In this work, we adopt the following notation:

TABLE II: MediaPipe Pose Landmarks.

Body Joint	Abbr.	MediaPipe Joint Index
*h*ead	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 *wr*ist	wr^l	16
*r*ight *sh*oulder	sh^r	11
*r*ight *el*bow	eb^r	13
*r*ight *wr*ist	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



- $ja_t(j_1, j_2, j_3)$ 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 $C \in \{x, y, z\}$.

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 $c \in C$

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

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

D. Video-level Compensation Assessment

Given the positive results in this task [24] [20] [47], we train a Multi-Layer Perceptron (MLP) binary classifier to assess each compensatory motion pattern 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 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. As videos have different sizes, we padded variables at $t = 1$ to videos with fewer frames.

1) Implementation Details: We implement our models using the *'Pytorch'* library [48], with parameter optimization conducted through the Optuna framework [49]. 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-label Generation

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) Saliency Maps of Features & Frames: Given the binary VL 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 [40]. Thus, we obtain a vector of gradients, $S^{i\prime}$, matching the shape of the input vector, given by

$$S^{i\prime} = \begin{cases} \frac{\partial \hat{y}^i_{score}}{\partial f^i}, \text{ if } \frac{\partial \hat{y}^i_{score}}{\partial f^i} > 0 \\ 0, \text{ otherwise} \end{cases} \tag{1}$$

where $\hat{y}^i_{score}$ is the predicted class score for video i and f^i is the input vector. The saliency map is created by reshaping $S^{i\prime}$ vector into a matrix $S^i \in \mathbb{R}^{T\times D}$. A row has the gradient for a kinematic variable, d, across all frames. A column is the gradient of each kinematic variable for a specific frame, t.

2) Frame-level Pseudo-Scores: 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

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

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.

3) Frame-level Pseudo-Label Generation Threshold Method: Given the pseudo-scores for each video frame, s^i_t, we fine-tune a threshold, τ, to classify frames as either normal or indicating a compensatory pattern, aiming for high-quality FLPL that aligns closely with ground-truth, achieving strong compensation detection accuracy and low error rates. Using this threshold, τ, each frame receives a pseudo-label, z^i_t,

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

where $\hat{y}^i$ is the predicted class from the VL 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.

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.

1) Implementation Details: 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.1.

G. Datasets of Functional Exercises 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 exercises 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. 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.

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

2) StrokE Rehab Exercises (SERE) dataset: SERE is the newly collected dataset. Table IV describes the exercises 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 exercises.

2.1) Data Collection: We recorded the videos at a frame rate of 30 fps using a ZED Mini Stereo Camera from StereoLabs [3]

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

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 exercises 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.

2.2) Participants: 20 post-stroke patients (7 females and 13 males), with 62.3 ± 14.77 years old, participated on data collection 17.46 ± 36.67 months after the stroke event and performed ten motion trials for each exercise. 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. Post-stroke patients signed an informed consent authorizing data recording.

2.3) 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.

TABLE V: *SERE* Dataset Participants' Profiles.

Patient ID	Age	Sex	Affected Side	Type	Time After Stroke [years|months]
P01	64	M	Right	Ischemic	12.10 | 145.17
P02	66	M	Left	Hemorrhagic	1.33 | 16.03
P03	88	M	Right	Ischemic	1.16 | 13.90
P04	78	F	Right	Hemorrhagic	8.33 | 104.00
P05	70	M	Left	Ischemic	0.19 | 3.80
P06	61	M	Left	Ischemic	0.11 | 2.37
P07	55	M	Right	Hemorrhagic	0.08 | 0.87
P08	59	M	Left	Hemorrhagic	0.42 | 21.86
P09	40	F	Left	Ischemic	0.13 | 1.57
P10	78	F	Left	Ischemic	0.30 | 3.60
P11	55	F	Right	Hemorrhagic	0.41 | 4.90
P12	47	F	Left	Ischemic	0.25 | 3.03
P13	40	M	Right	Ischemic	0.46 | 5.47
P14	77	M	Left	Hemorrhagic	0.28 | 3.37
P15	72	M	Left	Ischemic	0.35 | 4.17
P16	75	M	Left	Ischemic	0.22 | 2.60
P17	36	M	Right	Ischemic	0.34 | 4.10
P18	43	M	Left	Ischemic	0.26 | 3.17
P19	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 level, and *SERE* dataset (target dataset) fully labeled at a video level but with only a small subset labeled at a frame level. Figure 4, illustrates the considered approaches to evaluate the FLC's generalization potential to new patients and exercises. We inspect whether a FLC trained with the source dataset (*TULE*) or trained with the target dataset (*SERE*) FLPL - generated with a pseudo-labeler threshold, τ, fine-tuned on *TULE* or *SERE* small subset with FLL - yields better results for FL compensation assessment on *SERE* test set, given that this dataset is weakly labeled.

In an exploratory stage with *TULE* (source dataset fully labeled at a video and frame level), we train a binary VLC for each exercise and compensatory motion. We fine-tune the pseudo-labeler threshold, τ, to compute FLPL, as depicted in Section III-E.3, and train a fully supervised FLC with the FLPL. We evaluate this stage through Leave-One-Subject-Out (*LOSO*) cross-validation against the fully supervised FLC baseline, trained with ground-truth labeling. With this stage, we determine the loss in performance of training the FLC with ground-truth labels or pseudo-labels. We also determine the average values for the pseudo-labeled threshold, τ, for each compensatory motion pattern.

In approach 1, the baseline (Figure 4.a), we train the FLC with *TULE* and test it on *SERE* test set, fully labeled at a frame-level. We aim to evaluate how using *TULE* to train the FLC may enable the assessment of the *SERE* test set.

In approach 2, the transfer learning approach (Figure 4.b), we train a binary VLC for each exercise and compensatory motion with a training set. We use the average thresholds (τ) for each compensatory motion determined in the exploratory stage with *TULE* (Table VII), to produce FLPL for *SERE* training set with the pseudo-labeler. The FLC 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. 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 FL assessment on *SERE* test set.

In approach 3, semi-supervised (Figure 4.c), we fine-tune the pseudo-labeler threshold with a held-out validation set of two post-stroke patients labeled at a frame level. Finally, we test the FLC 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 FL 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

(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 exercises for rehabilitation.

(model ability to identify all positive samples). TAR 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. Threshold Selection

Figure 5 illustrates TAR and FAR evolution, regarding FLPL quality, for different threshold values in the interval $[0,1]$, for *TULE* E1 in the exploratory stage (Figure 5a) and *SERE* E2 (Figure 5b) and E4 (Figure 5c) validation sets in the semi-supervised approach. In selecting the 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]. Analyzing the TAR and FAR curves, we identified a threshold that optimally balances these metrics, reflecting a trade-off suitable for the application (see Table VII).

In the exploratory stage with *TULE*, for E1, Figure 5a shows a FAR below 10%, while the TAR is highly declining with τ. In the semi-supervised approach with *SERE*, Figures 5b and 5c show TAR and FAR progress regarding validation set FLPL quality. TAR and FAR curves are unstable as we only consider two post-stroke patients. For E3 head compensation, E5 (detailed in Figure 2i of the supplementary materials), and E4 (Figure 5c), TAR and FAR have similar evolution, starting with high values and decreasing with the threshold. We prioritized a TAR higher than the FAR.

TABLE VII: Summary of the selected Pseudo-labeler thresholds, τ, for the exploratory stage with *TULE* and the transfer learning and semi-supervised approaches with *SERE*.

Dataset & Approach	Exercise	Comp. Motion Target Joint	Threshold (τ)
TULE Exploratory Stage	E1	Shoulder	0.10
		Trunk	0.11
		Head	0.06
	E2	Shoulder	0.07
		Trunk	0.16
		Head	0.07
	E3	Shoulder	0.03
		Trunk	0.02
		Shoulder	0.07
	Average	Trunk	0.10
		Head	0.07
SERE Transfer learning approach	E1, E2, E3	Shoulder	0.07
	E1, E2, E3, E4, E5	Trunk	0.10
	E1, E2, E3	Head	0.07
SERE Semi-supervised approach	E1	Shoulder	0.35
		Trunk	0.68
		Head	0.2
	E2	Shoulder	0.04
		Trunk	0.29
		Head	0.23
	E3	Shoulder	0.62
		Trunk	0.41
		Head	0.21
	E4	Trunk	0.52
	E5	Trunk	0.05

V. Results

A. Quantitative Evaluation of frame-level Pseudo-labeling Quality

For both datasets, all exercises 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.

In the exploratory stage with *TULE* dataset, FLPL quality has an overall f_1 score above 90%, TAR over 80%, and FAR

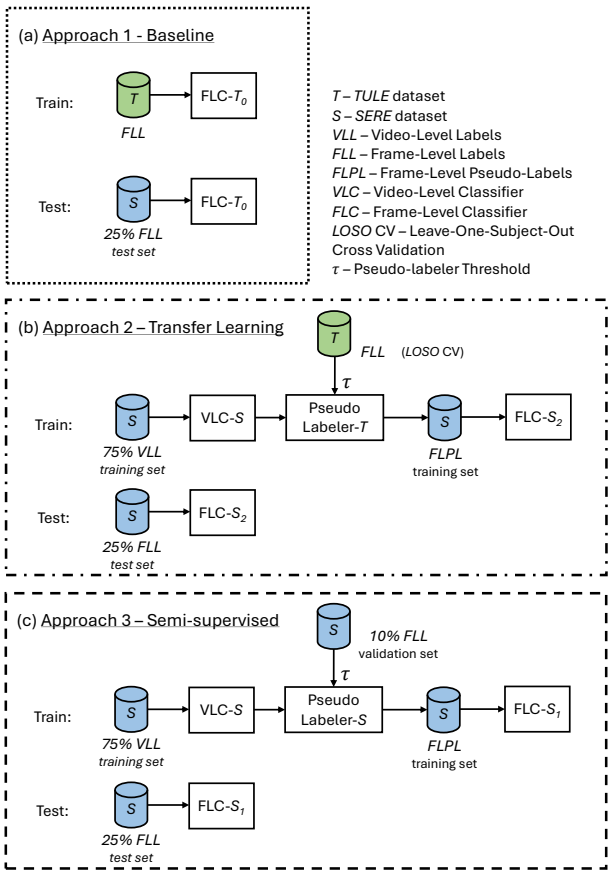


Fig. 4: Description of train and test steps of the exploratory stage with *TULE* and strategies conducted with *SERE* dataset.

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.

B. Quantitative Evaluation of frame-level Compensatory Motion Assessment

In the exploratory stage with *TULE* dataset, our approach had an overall performance of f_1 of 81.93%, 51.02% TAR, a FAR of 24.15%, and an AUC of 63.55%, which is competitive compared with the fully supervised baseline (f_1 = 85.85%, TAR = 57.38%, FAR = 19.46%, 74.31% AUC).

Table VIII details FL classification results with *SERE* dataset across experimental approaches. While the baseline approach provides better overall 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 overall 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.

In more detail, the baseline approach achieved better performance for E1 head compensation ($f_1 = 82.70\%$). The transfer learning approach had increased performance for E1 shoulder compensations ($f_1 = 88.33\%$), E2 shoulder and trunk compensation ($f_1 = 83.05\%$), and E3 trunk ($f_1 = 96.78\%$) and head compensation ($f_1 = 89.38\%$). Our semi-supervised approach achieved better results for E1 trunk compensation ($f_1 = 84.26\%$), E2 head compensation ($f_1 = 92.79\%$), E3 shoulder compensation ($f_1 = 77.84\%$), and E4 and E5 trunk compensation ($f_1 = 71.09\%$ and $f_1 = 77.49\%$, respectively).

C. Qualitative Evaluation of Saliency Maps

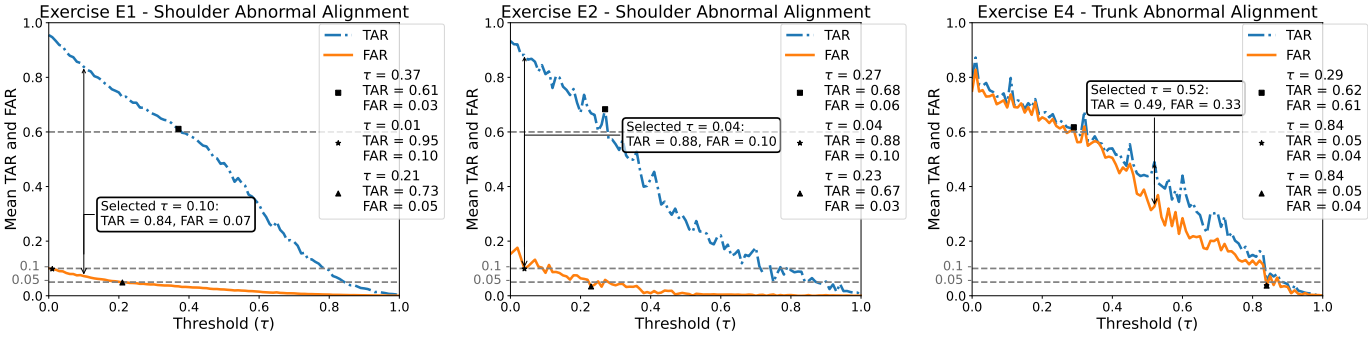
Figure 6 shows an example of a motion trial of a patient describing shoulder compensation and the saliency map of salient features and frames (Figure 6.a). Shoulder compensation is visible through joint markers (Figure 6.b) and image differencing (Figure 6.c). The saliency map captures the frames where 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).

VI. Discussion

A. Threshold Selection & Pseudo-labels Quality

We selected pseudo-labeler thresholds (τ) that reflect a suitable trade-off between TAR and FAR when envisaging the real-world application of our method. By analyzing the TAR and FAR curves (Figure 5) we prioritized minimum false alarms (FAR $< 10\%$) while maintaining a high TAR (TAR $> 60\%$) for compensation detection.

While for *TULE* dataset, we evaluate FLPL quality for the training set through *LOSO* cross-validation, for *SERE*, we only evaluate FLPL quality of the validation set in the semi-supervised approach. From Figure 5a, we can observe that TAR and FAR curves have steady progression, with FAR under 10% while TAR is high and decreases with the threshold value. Figure 5b shows that, unlike in the exploratory stage with *TULE*, the TAR and FAR curves for the *SERE* validation set present more variability. This is due to the small size of the validation set of only two post-stroke patients' motions. For E4, trunk compensation, TAR and FAR curves have similar behavior (Figure 5c), as both metrics are high and decrease with the threshold. This suggests two potential factors: suboptimal



(a) Shoulder compensation in E1 (exploratory stage with *TULE*).

(b) Shoulder compensation in E2 (semi-supervised approach with *SERE*).

(c) Trunk compensation in E4 (semi-supervised approach with *SERE*)

Fig. 5: True Alarm Rate (TAR) and False Alarm Rate (FAR) evolution for several pseudo-labeler threshold (τ) values. It shows the values of TAR and FAR for the selected τ, the τ value for a TAR immediately above 60% (■), and the values of τ for a FAR immediately bellow 10% ($\star$) and 5% (▲).

TABLE VIII: Compensation frame-level assessment results for *SERE* test set fully labeler at a frame level for the three experimental approaches: 1) baseline approach, 2) transfer learning approach, and 3) semi-supervised approach.

E1

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Shoulder	0.6641 ± 0.1662	**0.9102 ± 0.0522**	0.4522 ± 0.2037	0.8592 ± 0.0451
	Trunk	0.8296 ± 0.0797	**0.6187 ± 0.1528**	0.2487 ± 0.1242	**0.8408 ± 0.0480**
	Head	**0.8270 ± 0.0610**	0.3889 ± 0.2298	**0.1700 ± 0.0251**	0.6374 ± 0.1544
	Overall	0.7736 ± 0.1023	**0.6393 ± 0.1449**	0.2903 ± 0.1177	0.7791 ± 0.0825
2	Shoulder	**0.8833 ± 0.0633**	0.7725 ± 0.2561	0.1199 ± 0.1411	**0.9843 ± 0.0163**
	Trunk	0.7943 ± 0.1260	0.1962 ± 0.1919	0.2428 ± 0.2193	0.7873 ± 0.0665
	Head	0.7358 ± 0.0682	0.5461 ± 0.0864	0.3348 ± 0.0747	0.6180 ± 0.1041
	Overall	0.8045 ± 0.0858	0.5049 ± 0.1781	0.2325 ± 0.1450	0.7965 ± 0.0623
3	Shoulder	0.8750 ± 0.0630	0.6546 ± 0.3714	**0.1009 ± 0.1244**	0.9840 ± 0.0162
	Trunk	**0.8426 ± 0.0898**	0.0814 ± 0.0814	**0.1669 ± 0.1665**	0.8030 ± 0.0613
	Head	0.7669 ± 0.0794	**0.5533 ± 0.1099**	0.2886 ± 0.1009	**0.6549 ± 0.0922**
	Overall	**0.8282 ± 0.0774**	0.4298 ± 0.1876	**0.1855 ± 0.1306**	**0.8140 ± 0.0566**

E2

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Shoulder	0.7021 ± 0.0818	0.6053 ± 0.2763	0.3729 ± 0.2013	0.8531 ± 0.0559
	Trunk	0.8139 ± 0.0743	**0.5290 ± 0.0283**	0.2552 ± 0.1094	0.7806 ± 0.0588
	Head	0.8978 ± 0.0435	**0.1500 ± 0.0000**	0.1542 ± 0.0353	**0.6680 ± 0.0000**
	Overall	0.8046 ± 0.0665	0.4281 ± 0.1015	0.2608 ± 0.1153	**0.7672 ± 0.0382**
2	Shoulder	**0.8305 ± 0.1037**	0.7902 ± 0.1268	**0.2239 ± 0.1847**	**0.9680 ± 0.0010**
	Trunk	**0.8706 ± 0.0633**	0.4309 ± 0.1343	**0.1570 ± 0.0467**	0.6639 ± 0.1080
	Head	0.9169 ± 0.0593	0.1014 ± 0.0000	0.1178 ± 0.0659	0.4744 ± 0.0000
	Overall	**0.8727 ± 0.0754**	0.4408 ± 0.0870	**0.1662 ± 0.0991**	0.7021 ± 0.0360
3	Shoulder	0.8284 ± 0.1060	**0.8037 ± 0.1163**	0.2294 ± 0.1864	0.9677 ± 0.0014
	Trunk	0.8542 ± 0.0635	0.4584 ± 0.1094	0.1856 ± 0.0789	**0.7887 ± 0.0870**
	Head	**0.9279 ± 0.0596**	0.0743 ± 0.0000	**0.0973 ± 0.0648**	0.4695 ± 0.0000
	Overall	0.8702 ± 0.0764	**0.4455 ± 0.0752**	0.1708 ± 0.1100	0.7420 ± 0.0200

E3

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Shoulder	0.7179 ± 0.1429	**0.8076 ± 0.1396**	0.3424 ± 0.2163	**0.8692 ± 0.0987**
	Trunk	0.7967 ± 0.0701	**0.2814 ± 0.0000**	0.3144 ± 0.1066	0.4457 ± 0.0000
	Head	0.7938 ± 0.1001	**0.2733 ± 0.1023**	0.2228 ± 0.0609	0.4963 ± 0.0288
	Overall	0.7695 ± 0.1043	**0.4541 ± 0.0806**	0.2932 ± 0.1279	0.6037 ± 0.0425
2	Shoulder	0.7671 ± 0.1362	0.6049 ± 0.2382	0.2219 ± 0.2083	0.8356 ± 0.1239
	Trunk	**0.9678 ± 0.0451**	0.1293 ± 0.0000	**0.0346 ± 0.0347**	**0.5680 ± 0.0000**
	Head	**0.8938 ± 0.1300**	0.0000 ± 0.0000	**0.0000 ± 0.0000**	**0.5000 ± 0.0000**
	Overall	**0.8762 ± 0.1038**	0.2447 ± 0.0794	**0.0855 ± 0.0810**	**0.6345 ± 0.0413**
3	Shoulder	**0.7784 ± 0.1326**	0.3420 ± 0.2399	**0.1253 ± 0.1783**	0.8408 ± 0.1176
	Trunk	0.8724 ± 0.1185	0.1597 ± 0.0000	0.1892 ± 0.1653	0.3664 ± 0.0000
	Head	0.6366 ± 0.1320	0.1639 ± 0.1262	0.4352 ± 0.1096	0.2537 ± 0.1244
	Overall	0.7625 ± 0.1277	0.2219 ± 0.1220	0.2499 ± 0.1511	0.4870 ± 0.0807

E4

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Trunk	0.5245 ± 0.2806	**0.7729 ± 0.0757**	0.5441 ± 0.2229	**0.6441 ± 0.0828**
2	Trunk	0.7092 ± 0.3347	0.3528 ± 0.0879	**0.2230 ± 0.1227**	0.4894 ± 0.1082
3	Trunk	**0.7109 ± 0.3477**	0.3856 ± 0.0567	0.2604 ± 0.1842	0.4981 ± 0.1360

E5

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Trunk	0.7714 ± 0.0681	0.6564 ± 0.0996	0.2211 ± 0.0474	0.7772 ± 0.0665
2	Trunk	0.7610 ± 0.0645	**0.7088 ± 0.1070**	0.2656 ± 0.0359	**0.7899 ± 0.0603**
3	Trunk	**0.7749 ± 0.0680**	0.6471 ± 0.1047	**0.2126 ± 0.0457**	0.7844 ± 0.0655

Overall

App.	Comp. Motion	f_1	TAR	FAR	AUC
1	Overall	0.7287 ± 0.1244	**0.5902 ± 0.1005**	0.3219 ± 0.1262	**0.7143 ± 0.0625**
2	Overall	**0.8047 ± 0.1328**	0.4504 ± 0.1079	**0.1946 ± 0.0967**	0.6825 ± 0.0618
3	Overall	0.7893 ± 0.1394	0.4260 ± 0.1092	0.2158 ± 0.1243	0.6651 ± 0.0737

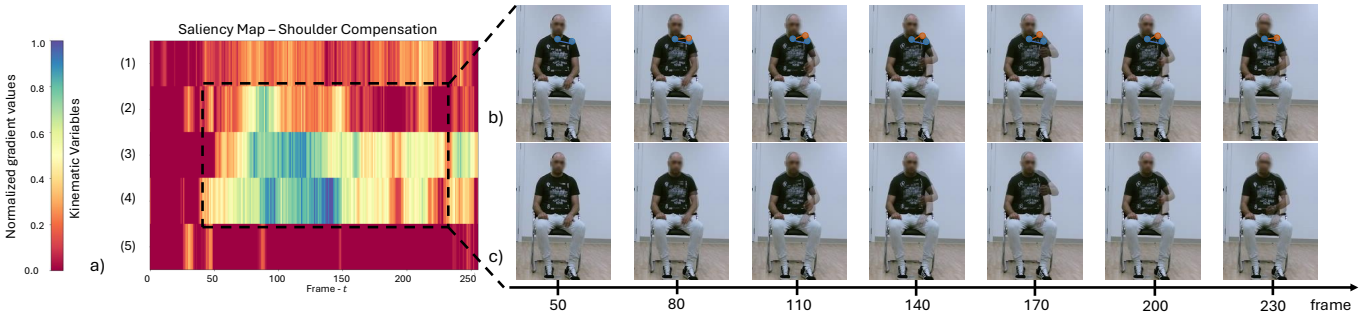


Fig. 6: Salient kinematic variables - shoulder (1) abduction angle, (2) elevation angle, and projected trajectories in (3) x, (4) y, and (5) z - and frames for shoulder compensation detection for an example of a motion trial from *TULE* dataset, E1. a) is the saliency map generated as detailed in Section III-E.1. b) and c) show the difference of shoulder initial and current positions through joint markers and image differencing, respectively.

selection of the held-out validation set and inaccuracies in the saliency maps used for FLPL generation. Also, two 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 validation, 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 FL outcomes and model adaptability to new patients.

B. Quantitative Evaluation of frame-level Compensatory Motion Assessment

The baseline approach achieved an overall 72.87% f_1 score, 59.02% TAR, and 32.19% FAR, on *SERE* test set. The transfer learning approach achieved an overall 80.47% of f_1 score, 45.04% of TAR, and 19.46% of FAR. The semi-supervised approach achieved an overall 78.93% of f_1, 42.60% of TAR, and 23.29% of FAR. The baselined approach provides better overall 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 ability to differentiate normal from compensatory motions (higher AUC) relates to a higher TAR. Meanwhile, a higher f_1 score is associated with fewer false alarms. Fine-tuning the pseudo-labeler τ with *TULE* (transfer learning approach) reduces the false alarms. Conversely, using a FLC trained with *TULE* (baseline approach) provided a solution with an improved ability to distinguish classes (higher AUC) and identify compensation with greater precision (higher TAR) but with more overall 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. Also, 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.

The FLC had unsatisfactory TAR assessing E2 head compensation and E3 trunk and head compensation across approaches. 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.

In the future, 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 [51]. Also, the exploitation of methods for pseudo-labels refinement might improve outcomes [28].

C. Qualitative Evaluation of Saliency Maps

Figure 6 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 an explanation of VLC decision [52]. 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 and label refinement approaches [27] might improve FLPL quality, leading to enhanced outcomes in the FL 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. Conclusions

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 for real-time compensation assessment on stroke rehabilitation exercises using a CAM-based technique and a threshold method to generate frame-level pseudo-labels (FLPL). With this approach, we ease the demands of data labeling when evaluating new patients and exercises. We enable real-time video compensatory motion assessment, allowing Virtual Coaches to provide patients with feedback at a frame level. Aiming to explore new motions, we cured a new dataset, the StrokE Rehab Exercises (*SERE*), of videos of 20 post-stroke patients performing five functional exercises 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 shows that creating pseudo-labels for the weakly labeled dataset to train a fully supervised FLC leads to improved outcomes. 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.

Acknowledgments

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

REFERENCES

[1] B. Semenko, L. Thalman, E. Ewert, R. Delorme, S. Hui, H. Flett, and N. Lavoie, "An evidence based occupational therapy toolkit for assessment and treatment of the upper extremity post stroke," *Screening*, vol. 4, pp. 4–1, 2015.

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

[3] M. Rensink, M. Schuurmans, E. Lindeman, and T. Hafsteinsdottir, "Task-oriented training in rehabilitation after stroke: systematic review," *Journal of advanced 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 and Feasibility Studies*, vol. 5, pp. 1–7, 2019.

[5] S. A. Billinger, R. Arena, J. Bernhardt, J. J. Eng, B. A. Franklin, C. M. Johnson, M. MacKay-Lyons, R. F. Macko, G. E. Mead, E. J. Roth *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, 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, 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 and neural repair*, vol. 16, no. 3, pp. 232–240, 2002.

[8] D. M. Morris, G. Uswatte, J. E. Crago, E. W. Cook III, and E. Taub, "The reliability of the wolf motor function test for assessing upper extremity function after stroke," *Archives of physical medicine and rehabilitation*, vol. 82, no. 6, pp. 750–755, 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 in stroke rehabilitation*, vol. 26, no. 2, pp. 94–100, 2019.

[10] T. M. Damush, L. Plue, T. Bakas, A. Schmid, and L. S. Williams, "Barriers and facilitators to exercise among stroke survivors," *Rehabilitation 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," *Clinical Rehabilitation*, vol. 14, no. 6, pp. 611–617, 2000.

[12] K. E. Watkins, W. 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 and Rehabilitation*, pp. 1–9, 2023.

[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, 2016.

[14] N. Maclean, P. Pound, C. Wolfe, and A. Rudd, "Qualitative analysis of stroke patients' motivation for rehabilitation," *Bmj*, vol. 321, no. 7268, pp. 1051–1054, 2000.

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

[16] T. G. Weimann, H. Schlieter, and A. B. Brendel, "Virtual coaches: background, theories, and future research directions," *Business & Information Systems Engineering*, vol. 64, no. 4, pp. 515–528, 2022.

[17] A. Ozturk, A. Tartar, B. E. 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, 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 and neural repair*, vol. 25, no. 1, pp. 71–80, 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, p. e104487, 2014.

[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 *International Conference on Intelligent User Interfaces*, 2019, pp. 218–228.

[21] F. Lanotte, M. K. O'Brien, and A. Jayaraman, "Ai in rehabilitation medicine: Opportunities and challenges," *Annals of Rehabilitation Medicine*, vol. 47, no. 6, p. 444, 2023.

[22] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning. arxiv 2022," *arXiv preprint 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 *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 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," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, pp. 1–16, 2022.

[25] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions," *Computers in Biology and Medicine*, vol. 166, p. 107485, 2023.

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

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

[28] Q. Yu and K. Fujiwara, "Frame-level label refinement for skeleton-based weakly-supervised action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3322–3330.

[29] Y. Xu, F. Wei, X. Sun, C. Yang, Y. Shen, B. Dai, B. Zhou, and S. Lin, "Cross-model pseudo-labeling for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2959–2968.

[30] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[31] 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, 2018.

[32] R. Aguilar-Ortega, R. Berral-Soler, I. Jiménez-Velasco, F. J. Romero-Ramírez, M. García-Marín, J. Zafra-Palma, R. Muñoz-Salinas, R. Medina-Carnicer, and M. J. Marín-Jiménez, "Uco physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises," *Sensors*, vol. 23, no. 21, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/21/8862

[33] E. Dolatabadi, Y. X. Zhi, B. Ye, M. Coahran, G. Lupinacci, A. Mihailidis, R. Wang, and B. Taati, "The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy," in *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*, 2017, pp. 375–381.

[34] 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, 2021.

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

[36] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognition*, vol. 110, p. 107413, 2021.

[37] 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 *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.

[38] T. Isobe and Y. Okada, "Rehabilitation xai to predict outcome with optimal therapies," in *Artificial Intelligence and Mobile Services–AIMS 2020: 9th International Conference, Held as Part of the Services Conference Federation, SCF 2020, Honolulu, HI, USA, September 18-20, 2020, Proceedings 9*. Springer, 2020, pp. 127–139.

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

[40] 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," *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[41] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (xai) on timeseries data: A survey," *arXiv preprint arXiv:2104.00950*, 2021.
[42] S. D. Goodfellow, A. Goodwin, R. Greer, P. C. Laussen, M. Mazwi, and D. Eytan, "Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings," in *Machine learning for healthcare conference*. PMLR, 2018, pp. 83–101.
[43] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.
[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 003–23 012.
[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
[46] M. F. Levin, J. A. Kleim, and S. L. Wolf, "What do motor "recovery" and "compensation" mean in patients following stroke?" *Neurorehabilitation and neural repair*, vol. 23, no. 4, pp. 313–319, 2009.
[47] A. R. Cóias, M. H. Lee, A. Bernardino, and A. Smailagic, "Skeleton tracking solutions for a low-cost stroke rehabilitation support system," in *2023 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2023, pp. 1–6.
[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Conference on Neural Information Processing Systems (NIPS)*, 2017.
[49] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[51] D. Zhu, X. Shen, M. Mosbach, A. Stephan, and D. Klakow, "Weaker than you think: A critical look at weakly supervised learning," *arXiv preprint arXiv:2305.17442*, 2023.
[52] 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," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–22, 2023.

APPENDIX

More study details are available in the Supplementary Materials.