Frame-Level Real-Time Assessment of Stroke Rehabilitation Exercises from Video-Level Labeled Data: Task-Specific vs. Foundation Models

Gonçalo Mesquita[1], Ana Rita Cóias[1] *Student Member, IEEE*, Artur Dubrawski[2] *Member, IEEE*, and Alexandre Bernardino[1] *Senior Member, IEEE*

***Abstract*— The growing demands of stroke rehabilitation have increased the need for solutions to support autonomous exercising. Virtual coaches can provide real-time exercise feedback from video data, helping patients improve motor function and keep engagement. However, training real-time motion analysis systems demands frame-level annotations, which are time-consuming and costly to obtain. In this work, we present a framework that learns to classify individual frames from video-level annotations for real-time assessment of compensatory motions in rehabilitation exercises. We use a gradient-based technique and a pseudo-label selection method to create frame-level pseudo-labels for training a frame-level classifier. We leverage pre-trained task-specific models - Action Transformer, SkateFormer - and a foundation model - MOMENT - for pseudo-label generation, aiming to improve generalization to new patients. To validate the approach, we use the *SERE* dataset with 18 post-stroke patients performing five rehabilitation exercises annotated on compensatory motions. MOMENT achieves better video-level assessment results (AUC = 73%), outperforming the baseline LSTM (AUC = 58%). The Action Transformer, with the Integrated Gradient technique, leads to better outcomes (AUC = 72%) for frame-level assessment, outperforming the baseline trained with ground truth frame-level labeling (AUC = 69%). We show that our proposed approach with pre-trained models enhances model generalization ability and facilitates the customization to new patients, reducing the demands of data labeling.**

***Index Terms*— Compensatory Motion Patterns, Stroke Rehabilitation, Real-time Motion Assessment, Saliency Maps, Weakly Supervised Learning, Transfer Learning, Transformers, Tasks-Specific Models, Foundation Model.**

I. INTRODUCTION

After a stroke, prompt rehabilitation therapy with task-oriented exercises is crucial for motor function recovery [1], [2], [3]. Therapists evaluate motor function, guide exercises, and provide proper feedback [4]. With the increasing patient load, high rehabilitation costs, and therapist shortages [5], [6], recommendations for autonomous exercise at home or between therapy sessions have grown [7]. Exercising alone is notably challenging as the lack of guidance and feedback highly impacts motivation and engagement, hampering recovery [5], [8]. This has drawn interest in rehabilitation support solutions research, such as Virtual Coaches (VCs). VCs must assess exercise performance and provide real-time feedback helping to improve motor function [9], [10].

Advances in computer vision and machine learning have made it possible to automatically and objectively assess movement ability from videos [11], [12], [13]. While reviewing performance after exercises needs video-level labels [13], giving feedback during exercises requires detailed labels at the clip or frame level. However, collecting fully labeled datasets is expensive, time-consuming, and often impractical in many real-world situations [14], [15].

Past research mainly used fully supervised models to provide real-time (frame-level) feedback on stroke rehabilitation exercises [16], [17], [18]. Lee *et al.* [19] used a gradient-based technique to generate frame-level pseudo-labels from salient features and frames, identifying compensatory patterns (e.g., shoulder elevation) in video frames from a full motion trial assessment. However, further evaluation on pseudo-labels usability for real-time assessment was missing. Later, Cóias et al. [20] introduced a framework using saliency maps and a threshold method to create pseudo-labels from salient kinematic measures and frames. These labels trained a model to detect incorrect movements at the frame level, aiming to work even on weakly labeled datasets. In semantic segmentation, Class Activation Mapping (CAM) for pixel-level label assignment has been used for training semantic segmentation modes, enhancing their performance [21]. In action recognition, video clip-level pseudo-labels have been generated by leveraging attention scores to facilitate action localization [22], [23].

To effectively support rehabilitation exercises or clinical decision-making, solutions should ensure robust model generalization to new patients as a single solution must accommodate a diverse population. However, it is impractical to train models from scratch for each new patient [24]. Using transfer learning, such as leveraging pre-trained models on large datasets, enhances generalization when the source and target domains or tasks are identical or somehow related [24]. These techniques have proven successful in human action recognition tasks [25], [26], [27] with small datasets.

In this work, we introduce a framework that classifies individual frames using video-level annotations for real-

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), by FCT under CMU Portugal, by the Portuguese Recovery and Resilience Plan (RRP), through project number 62, Center for Responsible AI, by NSF awards 2427948 and 2406231, and by DARPA award HR00112420329. Ana Rita Cóias is supported by the FCT doctoral grant [SFRH/BD/05239/2021].

[1]Institute for Systems and Robotics, Laboratory for Robotics and Engineering Systems, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal goncalo.mesquita@tecnico.ulisboa.pt; ana.coias@tecnico.ulisboa.pt; alex@isr.tecnico.ulisboa.pt

[2]Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA awd@cs.cmu.edu

time assessment of compensatory motions in rehabilitation exercises. Our approach leverages pre-trained task-specific models — Action Transformer (AcT) [28] and SkateFormer [29] — trained on large action recognition datasets, and the time-series foundation model MOMENT [30], which performance we compare against a Long Short-Term Memory (LSTM) network baseline for video-level assessment. We employ a gradient-based technique and a pseudo-label selection method to generate frame-level pseudo-labels from video-level predictions, which we use to train a frame-level classifier. We explore Vanilla Gradient and Integrated Gradient techniques and compare different strategies for pseudo-label selection. By using pre-trained models, we aim to enhance video-level generalization outcomes for new patients, improving quality in pseudo-labeling and subsequent frame-level classification. We evaluate our approach using the *SERE* dataset, which includes 18 post-stroke patients performing five functional exercises annotated on compensatory motions.

MOMENT achieves superior video-level assessment (AUC = 73%), surpassing the baseline LSTM (AUC = 58%). For frame-level assessment, AcT delivers the best results using Integrated Gradients for pseudo-label generation (AUC = 72%) outperforming the baseline trained with ground-truth frame-level labels (AUC = 69%). Frame-level ground-truth labels are only used for training the baseline frame-level classifier and for validation purposes. Our approach, leveraging pre-trained models, enhances generalization and simplifies customization to new patients, reducing data labeling efforts.

This work provides the following contributions:

- We explore the advantage of using pre-trained task-specific models, trained on action recognition datasets, and a foundation model for time-series on generalization outcomes to new patients on video-level compensatory motion assessment in stroke rehabilitation exercises;
- We investigate the impact of improved video-level assessment on frame-level pseudo-label generation by evaluating subsequent frame-level classification results across models used in the pipeline;
- We explore Vanilla and Integrated gradient techniques jointly with a single or dual threshold pseudo-label selection for frame-level pseudo-label selection.

II. METHODS

A. Problem Definition

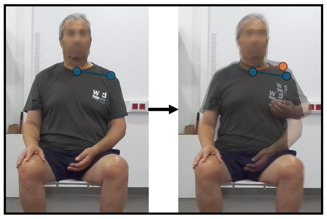


Fig. 1: Example of a shoulder elevation motion.

We consider a set of N untrimmed videos of post-stroke patients performing a functional exercise motion trial, $V = \{v^i\}_{i=1}^N$, i denotes the video number. Each video has a label, $y^i = \{0, 1\}$, denoting the existence of compensation in the described motion (0: normal, 1: compensatory motion). Post-stroke patients may describe compensatory motions — new movement patterns adopted to enable task completion (e.g., reaching an object) [31]. Excessive trunk flexion, shoulder elevatio, and head flexion are common compensatory motions. Figure 1 illustrates shoulder elevation compensatory motion.

Using a gradient-based technique, we generate a saliency map to highlight video frames where compensatory motions occur. The gradients are used for pseudo-label creation. In this paper, we denote as f^i the video features, z_t^i the pseudo-label for the frame t of the video i, and T is the maximum number of frames.

B. Approach Pipeline Overview

Figure 2 illustrates our approach pipeline for real-time video assessment. We consider as features of analysis body pose keypoints extracted by state-of-the-art human body pose detectors (box a) in Figure 2). We preprocess the extracted data to reduce noise artifacts (box a) in Figure 2). From the preprocessed features and video-level labels we fine-tune a video classifier (Model A) - AcT [28], SkateFormer [29], MOMENT [30], or LSTM - for video-level compensatory motion assessment (box b) in Figure 2). From Model A predictions, we apply a gradient-based technique to generate saliency maps [32], [33]. These maps are subsequently used to produce frame-level pseudo-labels by selecting the most salient features and frames using a pseudo-label selection method (box c) in Figure 2). Finally, with the same training data and generated pseudo-labels, we train a Multilayer Perceptron (MLP) for frame-level compensation classification (box d) in Figure 2).

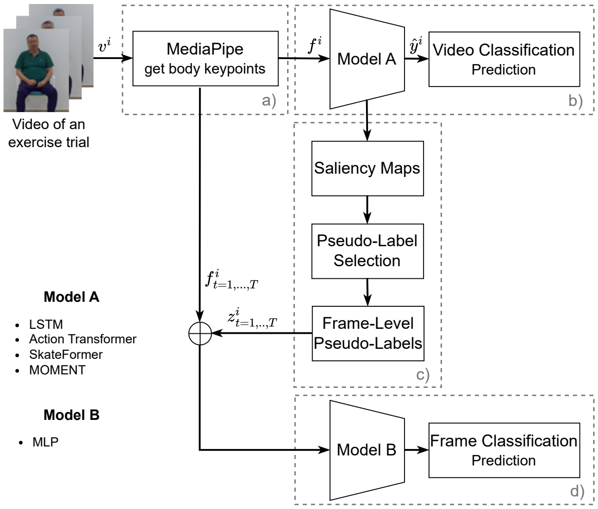


Fig. 2: Approach Pipeline: a) body pose extraction and preprocessing; b) video-level assessment with Model A; c) if a compensatory motion is detected, through the gradient-based technique we generate a saliency map to which we apply a pseudo-labels selection method to generate frame-level pseudo-labels; d) training of Model B, a Multilayer Perceptron (MLP), for frame-level assessment comparing the ground truth with the created frame-level pseudo-labels.

C. Pose Estimation and Preprocessing

The pose estimation and preprocessing steps extract key motion features, reduce noise, and enhance model generalization. MediaPipe[1] is used to track post-stroke patients' movements, providing 33 joint keypoints per frame in real-world 3D coordinates. Figure 3 shows the extracted keypoints. Its efficiency, robustness to occlusions, and real-time capabilities make it well-suited for stroke rehabilitation [34].

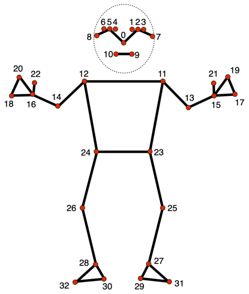


Fig. 3: MediaPipe Pose Keypoints.

Joint positions in each frame were offset by their initial trial positions, highlighting movement changes over absolute positions. This approach reduces starting position variability and improves the detection of subtle compensatory motions by representing motion trajectories as displacement vectors. We applied a moving average filter with a window of five frames to smooth small variations in the extracted signal.

D. Video-Level Compensation Assessment - Task Specific vs. Foundation Models

This study compares task-specific models, trained on action recognition datasets, Action Transformer (AcT) [28] and SkateFormer [29], with the MOMENT [30] time-series foundation model, pre-trained models on large and diverse datasets to tackle multiple tasks. We fine-tuned these models with a rehabilitation exercises dataset. We aim to identify which model yields better video-level assessment outcomes and subsequently improves frame-level assessment results regarding the generalization to new patients by enhancing the quality of pseudo-labels.

The AcT, a fully self-attention based architecture inspired by the Vision Transformer [35], is tailored for short-time, pose-based human action recognition. Its self-attention mechanisms capture temporal dependencies, focusing on key frames where compensatory movements are most pronounced. AcT was pre-trained and evaluated on the MPOSE2021 dataset [28], a large-scale comprehensive collection of human pose data derived from OpenPose and PoseNet across multiple human action recognition datasets.

Likewise, SkateFormer [29] is a skeletal-temporal transformer model designed for skeleton-based action recognition. It uses a partition-specific attention mechanism, classifying skeletal-temporal relationships into four categories based on joint proximity and temporal adjacency. This strategy enables SkateFormer to prioritize key joints and frames for action recognition, effectively capturing complex compensatory movements that can differ across tasks. SkateFormer was pre-trained on the NTU RGB+D 120 dataset [36], which offers a comprehensive collection of skeleton sequences for diverse human actions. Its effectiveness in capturing skeletal-temporal relationships has been demonstrated through evaluations on datasets such as NTU RGB+D 60 [37], NTU RGB+D 120 [36], and NW-UCLA [38].

Unlike task-specific models, MOMENT [30] is pre-trained on diverse time-series tasks such as classification, forecasting, anomaly detection, and imputation, using large-scale datasets from various time-series domains. This extensive pre-training equips MOMENT with a general understanding of time-series dynamics, enabling rapid adaptation to new scenarios, including the classification of compensatory movements in stroke rehabilitation.

The hypothesis is that pre-trained models enhance generalization, reducing the need for extensive data collection and retraining. Fine-tuning foundation models for rehabilitation tasks can provide reliable patient-specific predictions, even with limited data or varying patient characteristics.

For video-level compensatory motion assessment, an LSTM exclusively trained on the rehabilitation dataset serves as the baseline, configured as a Many-to-One model with a single layer and a hidden size of 192. The AcT, SkateFormer, and MOMENT models retain their original architectures.

E. Frame-Level Pseudo-Labels Generation

Following [20], we generate frame-level pseudo-labels from saliency maps highlighting significant features and frames from the video-level predictions. To create the saliency maps, we explore two gradient-based techniques: the Vanilla Gradient and the Integrated Gradient.

The Vanilla Gradient [39] method computes the gradient of the model's output with respect to input features, identifying areas where small input changes significantly impact predictions. The saliency map is derived from the absolute gradient values, highlighting influential features.

The Integrated Gradients [40] method extends Vanilla Gradient by integrating gradients along a path from a baseline input to the actual input, providing a more comprehensive feature attribution. Both methods generate saliency maps that emphasize key frames relevant to compensatory motion detection, even for unseen patients.

1) Pseudo-label Selection Method: From the saliency maps, we aggregate gradients frame by frame and apply min-max normalization to scale the results to a range of $[0, 1]$, yielding a pseudo-score s_t^i for each frame t. We distinguish the frames of a normal motion from the frames of a compensatory motion by thresholding frames pseudo-scores. We explore single threshold and dual threshold approaches.

Single threshold: This approach requires only one threshold τ. Using a threshold, τ, each frame is assigned with a pseudo-label, z_t^i,

$$z_t^i = \begin{cases} 0, \text{ if } \hat{y}^i = 0 \\ \mathbb{I}(s_t^i > \tau), \text{ if } \hat{y}^i = 1 \end{cases} \tag{1}$$

where $\hat{y}^i$ represents the predicted class for video i, and $\mathbb{I}$ is the indicator function. For normal motion video trial ($\hat{y}^i = 0$), all frames are assigned with a pseudo-label $z_t^i = 0$. For videos with compensatory motions ($\hat{y}^i = 1$), each frame's pseudo-score s_t^i is compared against the threshold τ. If $s_t^i >$

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

τ, the indicator function assigns a frame pseudo-label $z_t^i = 1$ and $z_t^i = 0$, otherwise.

Dual threshold: This approach uses two thresholds τ_1 and τ_2, with $\tau_1 < \tau_2$. Therefore each frame is assigned with a pseudo-label, z_t^i,

$$z_t^i = \begin{cases} 0, & \text{if } \hat{y}^i = 0 \vee (s_t^i < \tau_1 \wedge \hat{y}^i = 1) \\ 1, & \text{if } s_t^i > \tau_2 \wedge \hat{y}^i = 1 \\ \text{Not Used}, & \text{if } \tau_1 \leq s_t^i \leq \tau_2 \wedge \hat{y}^i = 1. \end{cases} \quad (2)$$

where, frames with scores below τ_1 are confidently considered normal (labeled 0), while those above τ_2 are confidently considered compensatory (labeled 1). The region $\tau_1 \leq s_t^i \leq \tau_2$ may be treated as uncertain, therefore those labels are excluded and not used. This two-threshold strategy can reduce noisy frame-level assignments by clearly separating highly likely normal or compensatory frames from uncertain ones, however also has less data to feed Model B.

F. Frame-Level Compensation Assessment

Using the training set and the generated pseudo-labels, we train a fully supervised classifier for compensatory motion detection at the frame level, reducing reliance on costly data labeling. In Model B, we train a Multilayer Perceptron (MLP) and explore different model architectures, testing one to two layers with varying hidden units (32,48,64,96,128,192,256) for a binary classification task.

III. EXPERIMENTS

A. StrokE Rehab Exercises (SERE) dataset

SERE [20] is a newly collected dataset of 18 post-stroke patients performing five rehabilitation exercises. Table I details the exercises. In exercise 1 (E1), a patient raises their arm toward the head, simulating the action of combing their hair. In exercise 2 (E2), a patient has to move affected or unaffected arms towards the mouth and move it like brushing the teeth. In exercise 3 (E3), a patient must move both arms, simulating washing the face. In exercise 4 (E4), a patient has to tilt the trunk and move their arms towards the foot as if putting on socks for both feet. In exercise 5 (E5), a subject must lift their knee, flexing the hip with each leg.

Post-stroke patients performed ten repetitions of each exercise, with affected or unaffected arm[2] (E1 and E2), both arms simultaneously (E3 and E4), and both legs (E5). The dataset has 1260 videos in which 538 compensatory motions are displayed (E1: 170, E2: 118, E3: 100, E4: 30, E5: 120).

1) Data Collection: The videos were recorded at a frame rate of 30 fps with a ZED Mini Stereo Camera from StereoLabs[3] and the ZED Explorer framework from the ZED SDK. The camera was positioned 2.5 meters from the patients and 0.90 meters above the floor. To ensure their safety at all times, patients performed the exercises while seated in a chair. Data collection occurred at the NeuroSer Rehabilitation Center and the Alcoitão Rehabilitation Medicine Center.

TABLE I: Functional exercises for stroke rehabilitation and corresponding joint motions.

Exercise	Description	Motions
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

Data collection complies with the General Data Protection Regulation (GDPR) and was approved by NeuroSer and Alcoitão Rehabilitation Center ethics committees.

2) Participants: Data collection involved 18 post-stroke patients (6 females and 12 males), with 61.22 ± 15.06 years old, 12.73 ± 32.49 months after the stroke, whose profiles are detailed in [20]. Due to ethical reasons and consents, P02 and P04 [20] where excluded from this study.

3) Annotation: Two physiotherapists and a occupational therapist, with 9.33 ± 1.25 yeas of experience in stroke rehabilitation, assessed compensation during exercise performance and annotated the dataset.

B. Evaluation

We evaluate our approach using *Leave-One-Subject-Out* (*LOSO*) cross-validation, where for each run, we fine-tune Model A (box b in Figure 2) using data from all post-stroke patients except one, who is held out for testing. From this training set, frame-level pseudo-labels are generated using our pseudo-label selection method, which leverages either Vanilla Gradient or Integrated Gradient saliency maps.

Figure 5 shows a saliency map generated by the AcT using the Integrated Gradients (IG) method, trained on a video motion trial of E1 from a post-stroke patient. The model classifies the motion trial as a compensatory motion. On the vertical axis, we see which joints (features) receive the highest relevance scores throughout the trial.

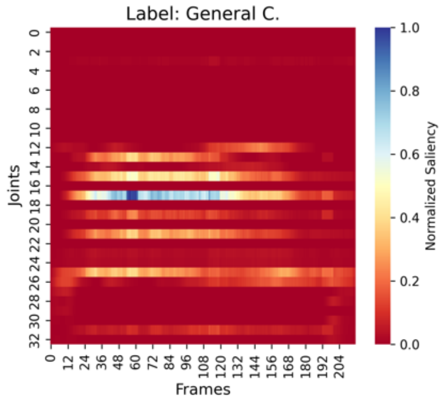
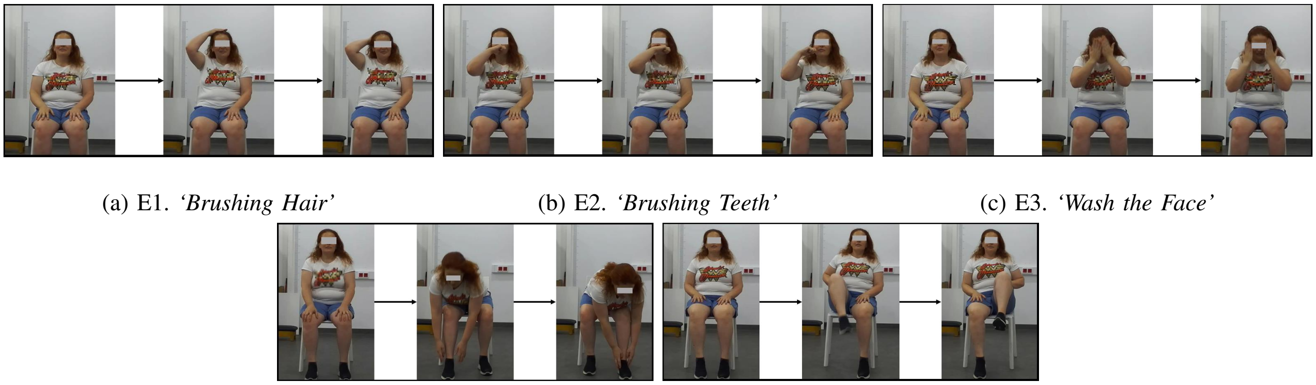


Fig. 5: Saliency map generated by the AcT model using the Integrated Gradients (IG) method, for a video of a post-stroke patient performing E1. The video trial is classified as a compensatory movement.

Since E1 is performed with the left arm, the joints most actively involved are 13, 15, 17, 19, and 21 (Figure 3). In the

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

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



(a) E1. *'Brushing Hair'*

(b) E2. *'Brushing Teeth'*

(c) E3. *'Wash the Face'*

(d) E4. *'Put on Socks'*

(e) E5. *'Hip Flexion'*

Fig. 4: *SERE* functional exercises for rehabilitation.

saliency map, these joints are highlighted more prominently than in the case of a normal (non-compensatory) motion, indicating a compensatory behavior.

Additionally, for the pseudo-label creation, we inspect the benefit of using a single or dual threshold pseudo-label selection based on specific tradeoffs for False Positives and False Negatives ratios. Table II presents the threshold values. These were determined for each model to balance the ratio between False Positives and False Negatives, thereby obtaining pseudo-labels more aligned with the actual distribution of the data and improving the performance of Model B.

TABLE II: Threshold values applied to Vanilla Gradient (VG) and Integrated Gradient (IG) for LSTM, AcT, SkateFormer, and MOMENT.

Model	**Threshold #1**		**Threshold #2**	
	VG	**IG**	**VG**	**IG**
LSTM	4.5	2.5	4; 6.5	2; 3
AcT	5.8	2.7	5.3; 6.3	2.2; 3.2
SkateFormer	3	1.5	1.5; 2.5	0.7; 2
MOMENT	2.5	3	2; 3	2; 3.5

Finally, the training data and the pseudo-labels are used to train Model B, a MLP (box d) in Figure 2), for frame-level assessment. The MLP is tested on the held-out post-stroke patient. This process is conducted 18 times until every post-stroke patient data is used in the test step, assessing all individual motion patterns. We compare the MLP's predictions to frame-level ground-truth labels to validate the approach.

We use the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate our approach and experiments. The AUC measures the model's ability to distinguish classes and is unaffected by the pseudo-label selection threshold.

IV. RESULTS AND DISCUSSION

Models A and B were trained using learning rates (0.001, 0.0001, 0.00001) and dropout probabilities (0.2, 0.3) to optimize generalization and prevent overfitting. The training used the Adam optimizer with a cosine scheduler, binary cross-entropy loss, and ReLU activation, except for the final sigmoid layer. Early stopping prevented overfitting, and batch sizes of 16 and 32 were tested.

A. Video-Level Compensation Assessment - Task Specific vs. Foundation Model

We evaluate AcT, SkateFormer, and MOMENT performance on video-level compensation assessment against a baseline LSTM. Table III summarizes our results. MOMENT yields better results in the task, with an average AUC of $73\pm 20\%$, outperforming the opponent models with no significant difference ($p > 0.05$ using paired t-tests) and the LSTM with statistical significance ($p < 0.05$). All models outperformed the baseline LSTM, trained exclusively in *SERE* dataset (AUC = $58 \pm 24\%$), with the AcT and MOMENT revealing statiscally significant difference ($p < 0.05$).

Although AcT and SkateFormer are pre-trained in large action recognition datasets, the foundation model for time-series MOMENT provides improved performance in video-level assessment, demonstrating that the knowledge acquired from massive collections of non-task-specific time-series, results in an improved generalization capability to new post-stroke patients with distinctive motion patterns. By leveraging pre-trained knowledge and fine-tuning with the small target dataset, we enhance task accuracy and generalization with less data and reduced training costs.

TABLE III: Video-level Classification results. Area Under the ROC Curve (AUC) evaluated through *Leave-One-Subject-Out* (*LOSO*) cross-validation strategy across four models. Results are reported as mean ± standard deviation. All models outperformed the baseline LSTM with MOMENT revealing better performance (pairwise t-tests at 95% significance level).

	LSTM	**AcT**	**SkateFormer**	**MOMENT**
AUC	0.58 ± 0.24	0.68 ± 0.23	0.65 ± 0.23	**0.73 ± 0.20**

B. Frame-Level Compensation Assessment

We use the frame-level pseudo-labels to train a MLP for frame-level compensatory motion assessment. We explore whether a Vanilla Gradient (VG) or an Integrated Gradient (IG) technique benefits frame-level classification. In addition,

TABLE IV: Frame-level Classification results. AUC results (mean ± standard deviation) from *LOSO* cross-validation for Vanilla Gradient (VG) and Integrated Gradient (IG) techniques and video-level assessment models — LSTM, AcT, SkateFormer, and MOMENT. Superscripts indicate statistical comparisons based on paired t-tests at the 95% confidence level.

#Thr.	**LSTM**		**AcT**		**SkateFormer**		**MOMENT**		**Ground Truth**
	VG	**IG**	**VG**	**IG**	**VG**	**IG**	**VG**	**IG**	
1 Thr.	0.64 ± 0.10	0.64 ± 0.11	0.53 ± 0.15	**0.72 ± 0.11**†	0.58 ± 0.08	0.66 ± 0.14	0.51 ± 0.10	0.69 ± 0.09*	0.69 ± 0.11
2 Thr.	0.64 ± 0.10	0.66 ± 0.13	0.55 ± 0.18	**0.72 ± 0.10**†	0.64 ± 0.10	**0.72 ± 0.12**‡	0.52 ± 0.08	0.67 ± 0.12	

Legend: † Not significantly different from Ground Truth ($p > 0.05$), MOMENT IG, or AcT 2 Thr. * MOMENT IG significantly higher than MOMENT VG ($p < 0.05$) ‡ SkateFormer IG (2 Thr.) significantly higher than SkateFormer VG (1 Thr.) ($p < 0.05$)

we investigate if applying a single or dual threshold pseudo-label selection approach leads to improved outcomes. Table IV shows our results for each model, gradient-based technique, and pseudo-label selection approach.

Among all configurations, the AcT model combined with IG techniques achieved the highest frame-level classification performance (AUC = 0.72 ± 0.10), slightly surpassing the performance of the classifier trained with ground-truth labels (AUC = 0.69 ± 0.11). Interestingly, the choice between single and dual thresholding had no impact on AcT's performance. These results highlight the potential of saliency maps to generate pseudo-labels that are more robust than manual annotations.

SkateFormer showed improvements with both IG and dual thresholding. The AUC increased from 0.58 ± 0.08 (VG, 1 Thr.) to 0.72 ± 0.12 (IG, 2 Thr.). In contrast, the LSTM model presented limited sensitivity to both gradient type and Selection Method, with results plateauing around 0.64 ± 0.10, and only minor gains observed using IG and dual thresholding (0.66 ± 0.13).

Although MOMENT led in video-level classification, its frame-level performance was comparatively weaker. The model benefited from the use of IG (improving from 0.51 ± 0.10 to 0.69 ± 0.09 with single thresholding), but dual thresholding did not provide further gains.

The AcT IG and the SkateFormer IG, with a dual threshold, surpassed the MLP trained with ground truth labels. Although MOMENT has achieved better performance on the video-level assessment stage, the AcT provided higher quality pseudo-labeling, enabling a performance on the frame-level assessment even better than using frame-level ground truth labels. Working on top of accurate video-level predictions for frame-level pseudo-label automatic creation by thresholding gradient-based saliency maps can result in more precise labeling, leading to improved frame-level classification results compared with a model trained on ground truth labeling. Ground truth data labeling can be noisy as human annotators are prone to mistakes and, usually, their labeling methodology is based on subjective data evaluation. Labeling is based on therapists' subjective evaluation of post-stroke patients' motions, in which the level of agreement might be low [13], and depends on professionals' experience. Additionally, detecting compensatory motion boundaries is highly challenging [16], leading to labeling inaccuracies.

V. CONCLUSIONS

In this work, we introduced a framework for real-time assessment of compensatory motion in stroke rehabilitation exercises using video-level annotations to train frame-level classifiers. By exploiting gradient-based saliency maps and a pseudo-label selection method, we addressed the challenge of obtaining frame-level labels from video-level annotations, significantly reducing manual labeling efforts. We leverage pre-trained task-specific models and a foundation model to enhance generalization to new patients.

Our experiments demonstrated that the foundation model MOMENT achieves superior performance in video-level classification (AUC = 73 ± 20%), outperforming task-specific models and the LSTM baseline (AUC = $58 \pm 24\%$). For frame-level assessment, the Action Transformer (AcT) model, combined with Integrated Gradients (IG), delivered the best results (AUC = 72 ± 10%), even surpassing a model trained on ground-truth frame-level labels (AUC = $69 \pm 11\%$). These results suggest that pseudo-labels generated from accurate video-level predictions and gradient-based saliency maps can be more reliable than manual annotations, which are susceptible to human bias. Our approach reduces the efforts required for frame-level annotation, facilitating the training of new models while enhancing generalization to new patients, which is crucial in rehabilitation scenarios where individual motor patterns vary significantly.

To further improve the robustness of our approach, future work will focus on enhancing model generalization across different rehabilitation exercises. Additionally, we plan to explore the application of this technique to other datasets beyond the stroke rehabilitation domain, such as general action recognition benchmarks. This will help assess the method's ability to generalize across different types of movement patterns and tasks, providing insights into its broader applicability in human motion analysis.

ACKNOWLEDGMENTS

The authors would like to thank their colleagues from the Robot Vision Laboratory (VisLab), Laboratory of Robotics and Engineering Systems (LARSyS), Instituto Superior Técnico, and Mononito Goswami of the CMU Auton Lab, Carnegie Mellon University, for their insightful discussions, technical support, and valuable feedback throughout this research.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] A. R. Cóias, M. H. Lee, A. Bernardino, A. Smailagic, M. Mateus, D. Fernandes, and S. Trapola, "Learning frame-level classifiers for video-based real-time assessment of stroke rehabilitation exercises from weakly annotated datasets," *TechRxiv*, Jan. 2025.

[21] 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.

[22] 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.

[23] 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.

[24] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *Journal of Big Data*, vol. 9, no. 1, p. 102, 2022.

[25] Y. Abdulazeem, H. M. Balaha, W. M. Bahgat, and M. Badawy, "Human action recognition based on transfer learning approach," *IEEE Access*, vol. 9, pp. 82 058–82 069, 2021.

[26] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," *IEEE Access*, vol. 5, pp. 17 627–17 633, 2017.

[27] U. Kılıç, Ö. Ö. Karadağ, and G. T. Özyer, "Skelresnet: Transfer learning approach for skeleton-based action recognition," in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2024, pp. 1–4.

[28] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, vol. 124, p. 108487, 2022.

[29] J. Do and M. Kim, "Skateformer: skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*. Springer, 2025, pp. 401–420.

[30] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "Moment: A family of open time-series foundation models," *arXiv preprint arXiv:2402.03885*, 2024.

[31] 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.

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

[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[34] 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.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and X. Zhai, "Thomas unterthiner mostafa dehghani matthias minderer georg heigold sylvain gelly jakob uszkoreit and neil houlsby. an image isworth 16× 16 words: Transformers for image recognition atscale," in *International Conference on Learning Representations*, 2021.

[36] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[37] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[38] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.

[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," Apr. 2014, arXiv:1312.6034 [cs]. [Online]. Available: http://arxiv.org/abs/1312.6034

[40] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," Jun. 2017, arXiv:1703.01365 [cs]. [Online]. Available: http://arxiv.org/abs/1703.01365