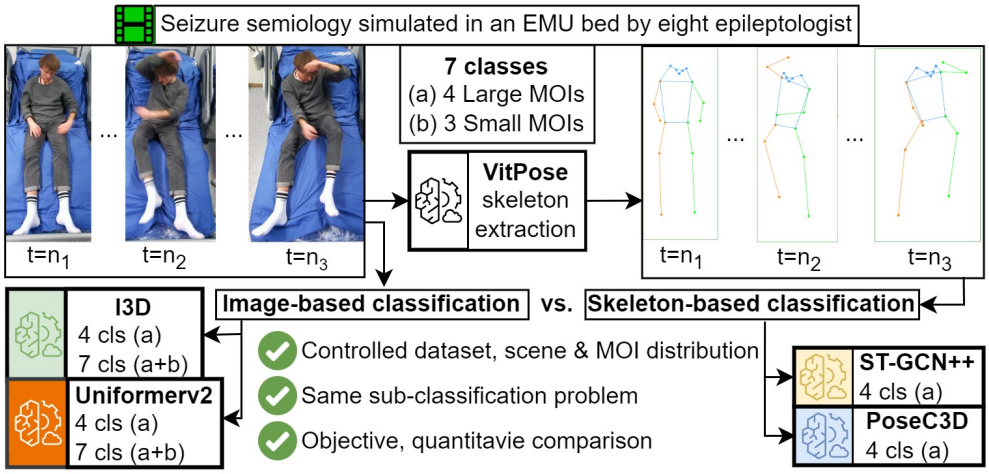
Graphical Abstract

Exploring Image and Skeleton-Based Action Recognition Approaches for Clinical In-Bed Classification of Simulated Epileptic Seizure Movements

Tamás Karácsony, Nicholas Fearns, Denise Birk, Selina Denise Trapp, Katharina Ernst, Christian Vollmar, Jan Rémi, László Attila Jeni, Fernando De la Torre, João Paulo Silva Cunha



Highlights

Exploring Image and Skeleton-Based Action Recognition Approaches for Clinical In-Bed Classification of Simulated Epileptic Seizure Movements

Tamás Karácsony, Nicholas Fearns, Denise Birk, Selina Denise Trapp, Katharina Ernst, Christian Vollmar, Jan Rémi, László Attila Jeni, Fernando De la Torre, João Paulo Silva Cunha

- Novel study to classify simulated Movements of Interest (MOIs) by action recognition.
- Acquired a novel 7-class simulated seizure MOI dataset acted by 8 epileptologists.
- Image-based vs. skeleton-based action recognition are compared for MOI classification.
- Highlights benefits of skeleton-based action recognition with transfer learning.
- Future work should integrate skeleton-based methods with hand gesture recognition.

Exploring Image and Skeleton-Based Action Recognition Approaches for Clinical In-Bed Classification of Simulated Epileptic Seizure Movements

Tamás Karácsony[a,b,c,1], Nicholas Fearns[d,1], Denise Birk[d], Selina Denise Trapp[d], Katharina Ernst[d], Christian Vollmar[d], Jan Rémi[d], László Attila Jeni[c], Fernando De la Torre[c], João Paulo Silva Cunha[a,b,*]

[a]*Center for Biomedical Engineering Research, Institute for Systems' Engineering and Computers, Technology & Science (INESC TEC), Porto, Portugal*
[b]*Faculty of Engineering (FEUP), University of Porto, Porto, Portugal*
[c]*Robotics Institute at Carnegie Mellon University, Pittsburgh, 15213, PA, USA*
[d]*Epilepsy Center, Department of Neurology, University of Munich, Munich, Germany*

Abstract

Epileptic seizure classification based on seizure semiology requires automated, quantitative approaches to support the diagnosis of epilepsy, which affects 1% of the world's population. Current approaches address the problem on a seizure level, neglecting the detailed evaluation of the classification of the underlying action features, also known as Movements of Interest (MOIs), which are critical for epileptologists in determining their classifications. Moreover, it hinders objective comparison of these approaches and attribution of performance differences due to datasets, intra-dataset MOI distribution, or architecture variations.

*Corresponding author
Email addresses: tamas.karacsony@inesctec.pt (Tamás Karácsony), nicholas.fearns@med.uni-muenchen.de (Nicholas Fearns), denise.birk@med.uni-muenchen.de (Denise Birk), selina.trapp@med.uni-muenchen.de (Selina Denise Trapp), katharina.ernst@med.uni-muenchen.de (Katharina Ernst), christian.vollmar@med.uni-muenchen.de (Christian Vollmar), jan.remi@med.uni-muenchen.de (Jan Rémi), laszlojeni@cmu.edu (László Attila Jeni), ftorre@andrew.cmu.edu (Fernando De la Torre), joao.p.cunha@inesctec.pt (João Paulo Silva Cunha)
[1]These authors contributed equally and share first authorship.

Objective evaluation of action recognition techniques is crucial, with MOIs serving as foundational elements of semiology for clinical in-bed applications to facilitate epileptic seizure classification. However, until now, there were no MOI datasets available nor benchmarks comparing different action recognition approaches for this clinical problem. Therefore, as a pilot, we introduced a novel, simulated seizure semiology dataset carried out by 8 experienced epileptologists in an EMU bed, consisting of 7 MOI classes. We compare several computer vision methods for MOI classification, two image-based (I3D and Uniformerv2), and two skeleton-based (ST-GCN++ and PoseC3D) action recognition approaches.

This study emphasizes the advantages of a 2-stage skeleton-based action recognition approach in a transfer learning setting (4 classes) and the multi-scale challenge of MOI classification (7 classes), advocating for the integration of skeleton-based methods with hand gesture recognition technologies in the future. The study's controlled MOI simulation dataset provides us with the opportunity to advance the development of automated epileptic seizure classification systems, paving the way for enhancing their performance and having the potential to contribute to improved patient care.

Keywords: Action recognition, Transfer Learning, Epilepsy, Semiology Dataset, Diagnosis Support

1. Introduction

Monitoring and classification of patient activity in epilepsy monitoring units (EMUs) play a pivotal role in the diagnosis of epileptic seizures, a condition affecting approximately 1% of the global population (Fiest et al., 2017; Begley et al., 2022). Automated diagnostic support systems are increasingly crucial, particularly in EMUs. Our group has been a pioneer in R&D of such systems, where quantitative seizure semiology-based diagnosis and evaluation hinge on identifying specific seizure Movements of Interest (MOIs) that are symptomatic of different epilepsy syndromes. The quantitative MOI concept was introduced in 2012 within our first 2D vs. 3D approach report (Cunha et al., 2012), where for instance, Frontal Lobe Epilepsy (FLE) is often characterized by hyperkinetic MOIs with rapid and large proximal movements, whereas Temporal Lobe Epilepsy (TLE) typically involves automotor MOIs with slower, distal, repetitive motions (e.g. hand, and finger movements) (Lüders et al., 1998; Noachtar & Borggraefe, 2009; Noachtar &

Peters, 2009).

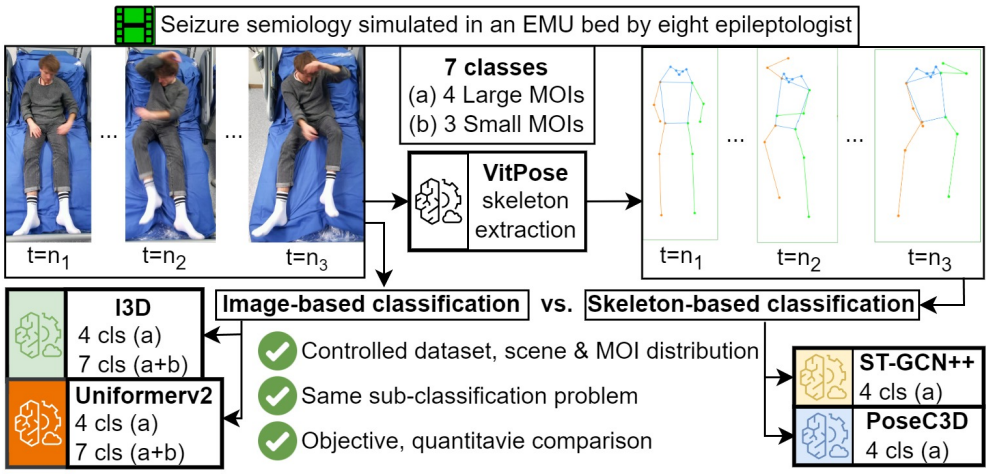


Figure 1: Overview of this work: comparing image-based (Uniformerv2 (Li et al., 2022)), I3D (Carreira & Zisserman, 2017)) and skeleton-based (ST-GCN++ (Duan et al., 2022a), PoseC3D (Duan et al., 2022b)) Movement Of Interest (MOI) classification methods on a seven-class simulated seizure semiology dataset with VitPose(Xu et al., 2022) skeleton extraction. (EMU - Epilepsy Monitoring Unit)

The development of automated epilepsy classification systems is hindered by the absence of standardized benchmark datasets, leading to a fragmented research landscape (Karácsony et al., 2024; Ahmedt-Aristizabal et al., 2024). Current top-performing approaches consider a deep learning action recognition approach with transfer learning (Karácsony et al., 2022; Pérez-García et al., 2021; Karacsony et al., 2020; Hou et al., 2021), most of them focusing on full seizure analysis (Pérez-García et al., 2021; Hou et al., 2021; Moro et al., 2023), classifying only smaller snippets of the seizures (Karácsony et al., 2022; Hou et al., 2022; Pérez-García et al., 2021; Moro et al., 2023), and then aggregating it to seizure level (Pérez-García et al., 2021; Moro et al., 2023), without specifically addressing MOI classification, critical for distinguishing between epilepsy types such as FLE and TLE (Noachtar & Peters, 2009).

The utilized clinical datasets (Ahmedt-Aristizabal et al., 2024) are private and relatively small, as only a part of each seizure contains the MOIs with classification value, with an unknown distribution between the seizures, introducing an unknown imbalance of the underlying temporal-spatial features. They address slightly different setups and classification problems, such

as FLE vs TLE (Karácsony et al., 2022), or Epileptic Seizures (ES) vs Psychogenic Non-Epileptic Seizures (PNES) (Hou et al., 2021, 2022), introducing dataset-specific biases. Besides seizure and patient-specific features, these include a different distribution of clinical challenges, such as blankets on the patients (TLE likely stays on, at least partially, FLE the patient likely kicks off most of the blanket), and clinicians surrounding the scene (PNES likely less interaction with the patient, ES likely more interaction with the patient). This complexity hinders objective, quantitative comparisons of automated classification approaches, making it challenging to discern performance improvements attributable to new architectural approaches from those arising from dataset representation variances, occlusions, and other challenges. Thus in this study, a controlled MOI simulation dataset was collected, as described in Sec. 2.1, enabling rapid development and objective comparison of MOI detection and classification approaches. Furthermore, this dataset enables the development of more explainable seizure classification methods in the future, as it enables the core differentiating features, MOIs, to be classified, following the clinical practice.

State-of-the-art Deep Learning (DL) RGB video-based action recognition is categorized into image-based, dominated by vision transformers (ViTs) and Convolutional Neural Networks (CNNs), and skeleton-based, primarily using Graph Convolutional Networks (GCNs) or CNN-based methods (Sun et al., 2022). Image-based approaches excel in automatically learning from raw videos, potentially enhancing accuracy by leveraging comprehensive video data, but suffer from high computational demands and extensive data requirements (Sun et al., 2022). Conversely, skeleton-based methods are efficient, require less computational power and data, and are robust to environmental changes, yet rely heavily on skeleton extraction quality and may miss non-skeletal details like facial expressions or gestures (Sun et al., 2022). To extract these skeletons, one of the best performing 2D Human Pose Estimation (HPE) methods from monocular videos, complying with the common EMU setup, are ViT-based approaches such as VitPose (Xu et al., 2022).

2. Materials and Methods

2.1. Dataset description and acquisition

A simulated seizure semiology dataset was recorded in an EMU bed, without blankets or other occlusions obstructing the camera, with our NeuroKinect4K system described in (Karácsony et al., 2024). This controlled

scenario enables the determination of the most effective architecture for MOI classification without the influence of additional environmental features. The movement sequences were designed and simulated by eight experienced epileptologists who have already seen and classified these MOIs countless times, ensuring a high level of accuracy in imitating these semiologies.

This dataset, created within a controlled clinical setting, aims to ensure a good balance between its limitations and advantages. It is designed to simulate seizure semiologies while removing unrelated clinical challenges, such as the presence of clinicians and blankets that often obstruct views in clinical settings. This provides clearer, unobstructed observations of MOI manifestations. Moreover, the dataset maintains a balanced representation of samples of the semiologies, a task usually unachievable in clinical settings due to the limited availability and visibility of certain seizure and movement types, such as hyperkinetic movements. This control is crucial as it can significantly influence performance outcomes.

Simulating a complete seizure pattern could pose challenges; however, MOIs, being smaller and well-defined segments, allow for effective emulation. Conducted by a group of seasoned clinicians, the dataset ensures high relevance along with considerable complexity and variability in movements. Although this dataset has its strengths, it's crucial to recognize that it does not encompass the full spectrum of variability in epileptic seizures. Instead, it primarily focuses on representing the key aspects of movement-based classification in this field, capturing variations in movement size and speed. Nevertheless, when it is compared to other real-world epileptic seizure datasets, it offers a great number and high variability of MOIs. This is significant because gathering comprehensive data on real-world epileptic seizures is both time-consuming and resource-intensive, as no publicly available clinical datasets currently exist. As the first dataset and approach to MOI classification, the novelty and benefits of this acquired dataset substantially outweigh the potential drawbacks of a slightly reduced complexity and variability when compared to clinical MOIs.

The video data was acquired with an Azure Kinect camera, in vertical 4K resolution (2160x3840, 30 fps), following our early geometric EMU video-EEG scene study from 2009 (Cunha et al., 2010), where the camera is mounted from the ceiling above the patient's feet facing the bed, frequently also used by other groups (Karácsony et al., 2024; Hou et al., 2022; Moro et al., 2023). The recorded simulated semiologies consisted of seven classes of MOIs, in two groups, (a) large MOIs included (1) 95 Hyperkinetic, (2) 85

Bilateral tonic, and (3-4) sign of four, with 92 right or 88 left arms extended and the (b) hand MOIs included (5-7) manual automatism with 88 left, 90 right or 113 both hands. In total 651 samples (26,048 frames total) were utilized, with a minimum of 9, a maximum of 217, and on average of 39.95 frame per sample. For the ground truth labels a neurologist (N.F.) labeled the start and end times of each simulated MOIs as if they were part of a clinical seizure. The included MOIs are some of the more common MOIs present in epileptic seizures, thus chosen to evaluate the feasibility of MOI classification and to guide future studies. While significantly more MOIs exist, we had to limit our selection to these due to the time availability of the staff, still, these classes provide a solid foundation, as most automatic seizure classifications, the closest comparable methods, only include 2 or 3 classes. The recording parameters and dataset metrics are summarized in Tab. 1, and all participants signed an informed consent. This data acquisition setup provided a controlled dataset in the target domain, both in terms of actions and scenery.

Table 1: Table of dataset main metrics, (So4 - Sign of 4, ext. - extended, Autom. - Automotor, # - Number of)

(a) Dataset

# Subjects	8
# Classes	7
Total # samples	651
Length of actions (frames) average±std (min/max)	39.95±24.60 (9/217)
Total # frames	26,048
Sensor	Azure Kinect
fps	30
Resolution	2160x3840 (vertical 4K)
Size of video data	1.35 GB

(b) Class distribution

classes	#Samples	group
So4, right arm ext.	92	a
So4, left arm ext.	88	
Hyperkinetic	95	
Bilateral tonic	85	
Autom., both hands	113	b
Autom., left hand	88	
Autom., right hand	90	

2.2. 2D Human Pose Estimation and object detection

For the skeleton-based approaches individuals were detected in every frame with YOLOX-x (Ge et al., 2021; Chen et al., 2019), a well-established object detection architecture. Each frame was resized to an input resolution of (640x640), keeping the original aspect ratio and padding, and bounding box detections were provided on the original frames. Then from inside these

detected bounding boxes the human poses were estimated with VitPose-H, in the standard 17 keypoints COCO format (Lin et al., 2014). For the HPE the frames were extracted and resized to the input resolution of (256x192), still keeping the original aspect ratio and padding (Xu et al., 2022; MMPose Contributors, 2020) (Fig.2). This processed the 4K video segments (NxHxWxC; Nx2160x3840x3) to a 2D skeleton sequence (NxKPxXYP; Nx17x3), where N is the number of frames, H and W are the height and width of the video, C is the 3 RGB channels, KP is the number of keypoints, XYP is the x,y coordinates and keypoint probabilities. .

2.3. Action recognition architectures

Four action recognition architectures were evaluated on the semiology simulation dataset representing the two main approaches for action recognition, UniFormerV2 (Li et al., 2022) and Inflated 3D ConvNet (I3D) (Carreira & Zisserman, 2017) representing image-based, ST-GCN++ (Duan et al., 2022a) and PoseC3D (Duan et al., 2022b) representing skeleton-based approaches. The implementations were adapted and extended from the pre-existing mmaction2 codebase (MMAction2 Contributors, 2020). The details of the training parameters are provided in Tab. 3 and the details of the pre-training datasets in Tab. 2

Table 2: Main dataset metrics of the pre-training data of the action recognition architectures (NC - Not Controlled; NTU RGB+D 60 Xsub - The NTU RGB+D 60 dataset in a cross-subject test setting)

Dataset	Modality	# Classes	# Videos	# Subjects
NTU RGB+D 60 Xsub (Shahroudy et al., 2016; Duan et al., 2022a)	RGB-D, skeleton	60	57K	40
Kinetics 400 (Kay et al., 2017)	RGB	400	306K	NC
Kinetics 710 (Li et al., 2022)	RGB	710	660K	NC

I3D. is a well-established 3D CNN based architecture, that was already utilized in epileptic seizure action classification. It was pre-trained on ImageNet (Deng et al., 2009) and Kinetics 400 (Kay et al., 2017). It samples uniformly from the input video 32 frames with a 224x224 input resolution.

UniFormerV2. is a modern, ViT-based, state-of-the-art performance architecture, that utilizes a CLIP-ViT (Radford et al., 2021) backbone. It was pre-trained on Kinetics-710 an action recognition dataset composed of Kinetics 400/600/700 (Kay et al., 2017; Carreira et al., 2018; Smaira et al., 2020)

as described in (Li et al., 2022). In this study specifically, the UniFormerV2-L/14 (CLIP-ViT-L/14 backbone) was utilized with a uniform 32-frame sampling from the input video with a 224x224 input resolution.

ST-GCN++. is one of the top performing Spatial-Temporal Graph Convolutional Network pre-trained on NTURGB+D-60-Xsub skeleton action recognition benchmark (Shahroudy et al., 2016; Duan et al., 2022a), it was trained with a uniformly sampled 2D COCO joint skeleton sequences (Lin et al., 2014; MMAction2 Contributors, 2020), with a sequence length of 100 (MMAction2 Contributors, 2020).

PoseC3D. is a 3D CNN based approach operating on 3D heatmap volumes of the 2D joint skeletons. The skeleton sequences are uniformly sampled to a sequence with a length of 48, then each skeleton is converted to a 17x56x56 heatmap and stacked in the sequence. It was pre-trained on NTURGB+D-60-Xsub skeleton action recognition benchmark (Shahroudy et al., 2016; Duan et al., 2022a), the same dataset as the above ST-GCN++ implementation.

2.4. Experiments and validation setup

Two main scenarios were set up for the experiments. The first included the 4 classes of (a) large MOIs, in which setting all architectures were evaluated, while the combination of (a) large MOIs and (b) small hand MOIs, consisting of 7 classes, was only evaluated with image-based approaches, as the skeleton-based approaches only include the main joints of the body. Therefore, this latter setting aims to evaluate the limitations of image-based approaches to additionally distinguish these relatively small gestures.

All experiments were set up in a transfer learning setting, and pre-trained weights were utilized, with only the action classification heads reinitialized. The pre-training datasets of the skeleton-based approaches were the same (NTURGB+D-60-Xsub) and the image-based approaches had overlapping pre-training datasets (Kinetics 400 and 710), (Sec. 2.3), which provides a relatively fair comparison between them. The backbones' weights were frozen for I3D, Uniformerv2, and PoseC3D except for the last 1, 2, and 1 layer respectively. ST-GCN++ backbone pre-trained weights were utilized, however, the backbone was not frozen, as it has a relatively low number of parameters it still benefited from a fully unlocked training, without instantly overfitting, as opposed to the other architectures.

2.4.1. Training

Training of each architecture was following the original training setup, additionally utilizing weighted loss to address the remaining slight class imbalances and adjusted learning rate schedules and batch sizes for our dataset and compute resources (Tab. 3). The learning rate (lr) of each architecture was scaled proportionally to the original pre-training batch size and the actual training batch size. $lr_{train} = lr_{orig} * \frac{batchsize_{train}}{batchsize_{orig}}$

Table 3: Summary of model and training parameters (cls - class, cv - cross validation, e. - epochs, G - Giga, GB- Gigabyte, h - hours, M - Million, Params. - Parameters, seq. - sequence)

Model	Input type	batch	Params. (M)	FLOPs (G)	GPUs (VRAM)	GPU hours 1 cv run
I3D	Video (32x224x224x3)	3x16	32.22	127	4cls: 3xRTX2080ti (11GB)	3x2 h (100e.)
		3x16			7cls: 3xRTX2080ti (11GB)	3x3.5h (100e.)
Uniformerv2	Video (32x224x224x3)	3x3	354	2561	4cls: 3xRTX2080ti (11GB)	3x11 h (300e.)
		2x16			7cls: 2xRTX6000ada (48GB)	2x56.5h (300e.)
ST-GCN++	2D Skeleton seq. (100x17x3)	1x16	1.38	1.95	1xGTX1080ti (11GB)	1x10 min (100e.)
PoseC3D	2D Skeleton heatmaps (17x48x56x56)	1x16	2.00	20.6	1xGTX1080ti (11GB)	1x1.5h (100e.)

2.4.2. Validation

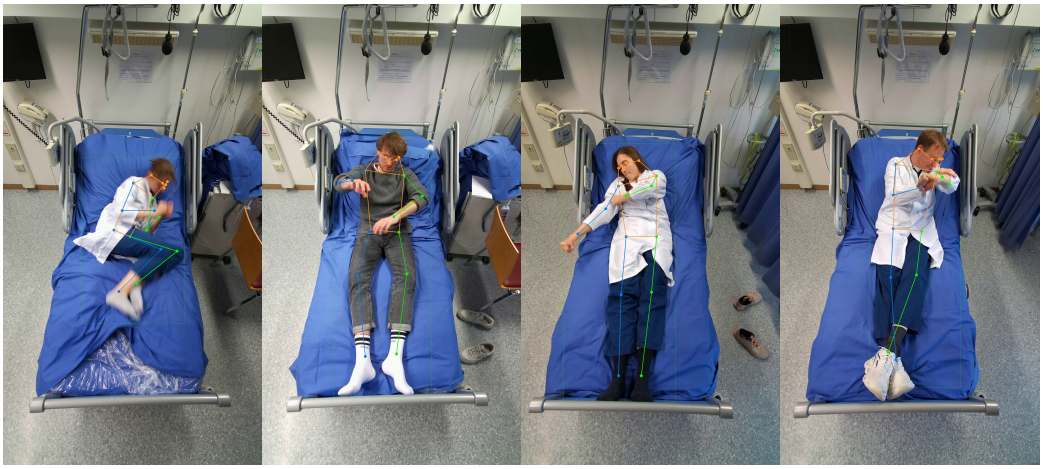
Validation of each experiment was carried out with an 8-fold, Leave One Subject Out (LOSO) cross-validation approach. Namely, the samples from 5 and 2 subjects were utilized for training and validation sets respectively during the training, early stopping was utilized based on the best f1 validation score. Then this architecture with the best f1 validation score was tested on the 1 hold-out test subject, to ensure the generalizability and cross-subject independence of the approaches validation. Then this was repeated for each subject, with the subjects organized in a circular list. In the results, the average of the test metrics from the 8-fold LOSO are provided for each experiment.

3. Results

3.1. 2D Human pose estimation

The 2D HPE approach successfully estimated the 2D poses even in challenging scenarios, with unusual poses, significant self-occlusion, and from an uncommon viewpoint. As no ground truth joint coordinates were available this was confirmed by visual inspection. Nevertheless, to provide insight

into the performance some of these more challenging scenarios and estimated poses are displayed in Figure 2.



(a) Hyperkinetic (b) Bilateral tonic (c) So4_rax (d) So4_lax

Figure 2: 2D pose estimation results for the following MOI: (a) Hyperkinetic, (b) Bilateral tonic, and Sign of four with the (c) right or (d) left arm extended classes (So4_rax or So4_lax)

3.2. Action recognition

The detailed results of the 8-fold LOSO cross-validation mean test metrics are presented in Tab. 4, including f1 score, accuracy, sensitivity, and specificity.

3.2.1. Image-based approaches

Image-based approaches mean test f1 score of the (a) 4 class setting was 0.84±0.21 and 0.91±0.13, for the I3D and Uniformerv2 architectures respectively, with Uniformerv2 achieving a 0.91 mean f1 test score, 0.07 larger than I3D architectures. However in the (b) 7 class setting this f1 score was 0.83±0.12 and 0.72±0.19 respectively, for I3D and Uniformerv2, with I3D having the performance advantage with a 0.11 higher mean f1 test score than the trained Uniformerv2 architectures. The performance drop was mainly introduced by the confusion introduced by the additional classes of the (b) hand MOIs group as represented on the confusion matrices on Fig. 3.

Table 4: Summary of quantitative results of the 8 fold LOSO cross-validation mean test metrics on the simulated semiology dataset. Architectures highlighted in bold represent the best-performing models under the given setup.

(a) 4 class results

Model	**Modality**	**f1 score**	**Accuracy**	**Sensitivity**	**Specificity**
I3D	Video	0.84±0.21	86.11±17.08%	85.83±16.91%	95.39±5.61%
Uniformerv2	Video	*0.91±0.13*	*92.96±9.07%*	*92.68±10.17%*	*97.65±3.29%*
ST-GCN++	Skeleton	0.91±0.1	91.23±9.81%	91.34±9.83%	97.08±3.29%
PoseC3D	Skeleton	**0.94±0.09**	**94.18±8.52%**	**93.97±8.52%**	**98.08±2.80%**

(b) 7 class results

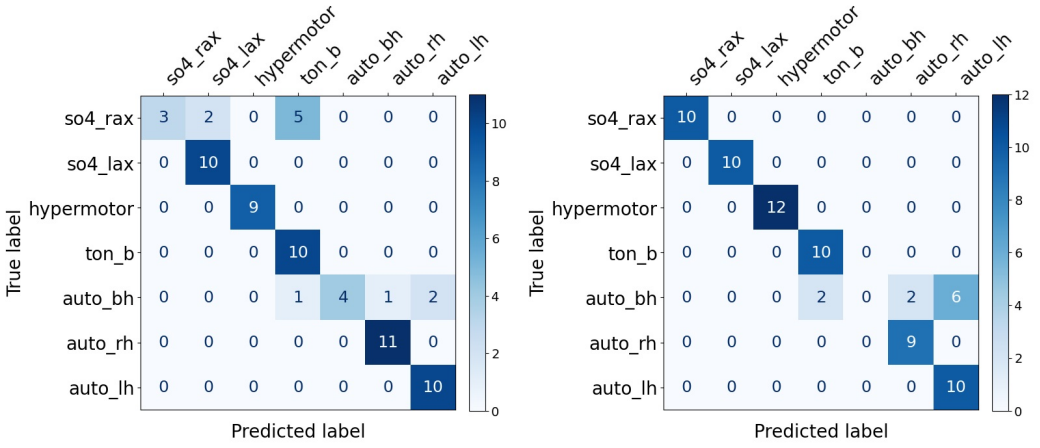
Model	**Modality**	**f1 score**	**Accuracy**	**Sensitivity**	**Specificity**
I3D	Video	**0.83±0.12**	**85.50±10.09%**	**84.78±10.61%**	**97.54±1.71%**
Uniformerv2	Video	0.72±0.19	74.99±17.37%	74.84±17.05%	95.82±2.91%

3.2.2. Skeleton-based approaches

Skeleton-based approaches both outperformed the image-based approaches in the 4 class setting, achieving mean test f1 scores of 0.91±0.1 and 0.94±0.09 for the ST-GCN++ and PoseC3D architectures respectively. Furthermore, they achieved a lower standard deviation of the f1 test scores across the 8 LOSO validation folds, than the image-based approaches.

4. Discussion

The recorded simulated dataset marks a pioneering step in enabling objective and quantitative assessment of epileptic seizure-related MOI classification within a clinical context. It provides the largest collection of MOI samples and unlike previous studies that lacked control over the distribution of MOI representations within the seizure videos and temporo-spatial features representation, this dataset allows for a direct comparison of classification strategies and the different snippet-level feature extraction strategies on the MOIs. When comparing the proposed MOI-based methods with seizure-level classification approaches, they have the significant advantage of enabling a more granular and interpretable classification, mirroring the cognitive processes used by epileptologists. Although, in end-to-end seizure level models these MOIs are implicitly learned from the full seizures during training, their distribution in the training dataset remains unknown and uncontrolled. As a result, these end-to-end models may overfit to certain MOIs while failing to capture others that are well documented in the clinical literature. In contrast, MOI-based approaches ensure that MOI distributions are explicitly



(a) I3D (b) Uniformerv2

Figure 3: Example confusion matrices of the end-to-end architectures, I3D and Uniformerv2 in the 7 class setting, a LOSO fold with an f1 score of 0.81 and 0.8 respectively (so4_rax,so4_lax - Sign of 4 right, left arm extended; ton_b - Bilateral tonic, auto_bh, auto_rh, auto_lh - Automotor both, right or left hands)

known and can be controlled during training. On the other hand, this approach requires more granular labeling work ahead of time, which makes it more challenging to acquire due to the limited availability of clinical staff.

In the proposed transfer learning setting, utilizing only a limited dataset, our evaluation of four action classification approaches revealed that skeleton-based methods excelled in a 4-class scenario. These methods are not only more lightweight compared to image-based approaches, making them ideal for smaller datasets, but also benefit from incorporating ViTs for HPE. This setup focuses on relevant extracted features, offsetting the demand for extensive training data to HPE, where large datasets can tackle most clinical challenges. Furthermore, skeleton-based approaches ensure accurate action classification by isolating patient movements from external influences like bystanders and clinicians.

While image-based approaches had the advantage of being able to address the 7 class setting, including the smaller MOIs, their performance was highly limited by the significantly lower relative resolution of the small hand MOI features compared to the full-body MOIs. Therefore, their practical performance might be misleading, as the class confusion distribution was much higher between the small movements. This could be problematic especially

if they would be utilized in seizure classification settings as a snippet-level feature extractor. This highlights the multiscale classification challenge of epileptic seizures. Uniformerv2 performed worse than I3D on the 7 class setting, likely because ViT-based architectures are significantly more data-hungry than 3D CNNs, thus it is worth considering that for limited-sized datasets with more classes, the more lightweight architectures might perform better.

This study acknowledges limitations, notably the variance in pre-training dataset sizes between image-based and skeleton-based architectures, with Kinetics datasets being substantially larger than those used for skeleton-based methods.

The skeleton-based approaches are particularly promising, including their ability to quantify movements. Furthermore, considering clinical in-bed scenarios, skeleton-based approaches enable preserving patient privacy, which is essential for international research collaborations, by providing an efficient way to share anonymized clinical data. This can be crucial for up-scaling the number of clinical collaborations, and acquiring large and diverse clinical datasets. Moreover, these methodologies have broad applicability in various in-bed patient activity monitoring scenarios, from intensive care and neurocritical care units to home sleep monitoring, which we discuss in (Karácsony et al., 2024).

In the future, we are expanding our dataset with more diverse simulated MOI classes and integrating hand gesture and facial expression recognition. Given the limited availability of clinical staff for labeling, we envision an iterative human-in-the-loop learning approach: first, training and evaluating models on an extended simulation dataset, then using these pre-trained models to pre-label patients' seizure data with MOI classes for the clinicians to validate and correct. Through ongoing retraining and validation, the model improves while accelerating the validation process. Eventually, this strategy may accelerate real-world clinical validation and enhance diversity of the dataset to enable future diagnosis support applications.

5. Conclusion

In conclusion, a novel simulated seizure semiology dataset was acquired, representing seven classes of MOIs commonly appearing in epileptic seizures. To the best of our knowledge this dataset is the first of its kind, the largest,

and controlled for environmental features by excluding occlusions, thus providing a fair platform to compare action recognition approaches in this domain. Utilizing this dataset, skeleton-based and image-based approaches were trained in a leave-one-subject-out (LOSO) manner. Among these approaches, the skeleton-based PoseC3D achieved an average F1 score of 0.94 in the 4-class, larger MOI setup. When small hand MOIs were included, extending to 7 classes, the image-based I3D approach achieved an average F1 score of 0.83. The acquired MOI simulation dataset enabled the objective and controlled comparison of DL action recognition approaches in the clinical setting of epileptic seizure classification applications. This opens opportunities for the future to train and evaluate these approaches on real-world clinical seizure MOIs, the building blocks of seizures, leading to more interpretable classification approaches, and it provides a controlled training source for real-world applications. After further validation, this could potentially lead to networks that are not only effective but also more interpretable.

Currently, the extracted skeleton dataset is only available for request from the corresponding authors.

CRediT authorship contribution statement

Tamás Karácsony: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Funding acquisition. **Nicholas Fearns:** Conceptualization, Methodology, Data Curation, Writing - Review & Editing. **Denise Birk:** Data Curation, Writing - Review & Editing. **Selina Denise Trapp:** Data Curation, Writing - Review & Editing. **Katharina Ernst:** Data Curation, Writing - Review & Editing. **Christian Vollmar:** Data Curation, Writing - Review & Editing. **Jan Rémi:** Resources, Data Curation, Writing - Review & Editing, Supervision, Funding acquisition. **László Attila Jeni:** Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Fernando De la Torre:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Funding acquisition. **João Paulo Silva Cunha:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Funding acquisition.

Acknowledgments

The 1st author was funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal program Ref PRT/BD/152202/2021.

DOI 10.54499/PRT/BD/152202/2021
(https://doi.org/10.54499/PRT/ BD/152202/2021).

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020
(https://doi.org/10.54499/LA/P/0063/2020)

References

Ahmedt-Aristizabal, D., Armin, M. A., Hayder, Z., Garcia-Cairasco, N., Petersson, L., Fookes, C., Denman, S., & McGonigal, A. (2024). Deep learning approaches for seizure video analysis: A review. *Epilepsy & Behavior*, *154*, 109735.

Begley, C., Wagner, R. G., Abraham, A., Beghi, E., Newton, C., Kwon, C.-S., Labiner, D., & Winkler, A. S. (2022). The global cost of epilepsy: A systematic review and extrapolation. *Epilepsia*, *63*, 892–903.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A Short Note about Kinetics-600. *arXiv:1808.01340v1*, .

Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 4724–4733). volume 2017-Janua.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., & Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, .

Cunha, J. P. S., Paula, L. M., Bento, V. F., Bilgin, C., Dias, E., & Noachtar, S. (2012). Movement quantification in epileptic seizures: A feasibility study for a new 3D approach. *Medical Engineering and Physics*, *34*, 938–945. doi:10.1016/j.medengphy.2011.10.013. Publisher: Elsevier.

Cunha, J. S., Vollmar, C., Fernandes, J., & Noachtar, S. (2010). Automated epileptic seizure type classification through quantitative movement analysis. In *World Congress on Medical Physics and Biomedical Engineering,*

September 7-12, 2009, Munich, Germany: Vol. 25/4 Image Processing, Biosignal Processing, Modelling and Simulation, Biomechanics (pp. 1435–1438). Springer.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE CVPR*. IEEE.

Duan, H., Wang, J., Chen, K., & Lin, D. (2022a). PYSKL: Towards Good Practices for Skeleton Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 7351–7354).

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022b). Revisiting Skeleton-based Action Recognition. *IEEE/CVF CVPR*, (pp. pp. 2969–2978).

Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C.-S., Dykeman, J., Pringsheim, T., Lorenzetti, D. L., & Jetté, N. (2017). Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology*, *88*, 296–303.

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, .

Hou, J. C., McGonigal, A., Bartolomei, F., & Thonnat, M. (2021). A Multi-Stream Approach for Seizure Classification with Knowledge Distillation. *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, .

Hou, J.-C., McGonigal, A., Bartolomei, F., & Thonnat, M. (2022). A self-supervised pre-training framework for vision-based seizure classification. In *IEEE ICASSP 2022* (pp. 1151–1155). IEEE.

Karácsony, T., Jeni, L. A., De la Torre, F., & Cunha, J. P. S. (2024). Deep learning methods for single camera based clinical in-bed movement action recognition. *Image and Vision Computing*, *143*, 104928.

Karacsony, T., Loesch-Biffar, A. M., Vollmar, C., Noachtar, S., & Cunha, J. P. S. (2020). A Deep Learning Architecture for Epileptic Seizure Classification Based on Object and Action Recognition. In *IEEE ICASSP 2020* (pp. 4117–4121).

Karácsony, T., Fearns, N., Vollmar, C., Birk, D., Rémi, J., Noachtar, S., & Silva Cunha, J. P. (2024). NeuroKinect4K: A Novel 4K RGB-D-IR Video System with 3D Scene Reconstruction for Enhanced Epileptic Seizure Semiology Monitoring. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1–5). doi:10.1109/EMBC53108.2024.10781546.

Karácsony, T., Loesch-Biffar, A. M., Vollmar, C., Rémi, J., Noachtar, S., & Cunha, J. P. S. (2022). Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification. *Scientific Reports 2022 12:1*, *12*, 1–13.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The Kinetics Human Action Video Dataset. *arXiv*, .

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., & Qiao, Y. (2022). Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *ArXiv*, *abs/2211.09552*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740–755). Springer.

Lüders, H., Acharya, J., Baumgartner, C., Benbadis, S., Bleasel, A., Burgess, R., Dinner, D., Ebner, A., Foldvary, N., Geller, E. et al. (1998). Semiological seizure classification. *Epilepsia*, *39*, 1006–1013.

MMAction2 Contributors (2020). Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`.

MMPose Contributors (2020). Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`.

Moro, M., Pastore, V. P., Marchesi, G., Proserpio, P., Tassi, L., Castelnovo, A., Manconi, M., Nobile, G., Cordani, R., Gibbs, S. A. et al. (2023).

Automatic video analysis and classification of sleep-related hypermotor seizures and disorders of arousal. *Epilepsia*, .

Noachtar, S., & Borggraefe, I. (2009). Epilepsy surgery: a critical review. *Epilepsy & Behavior*, *15*, 66–72.

Noachtar, S., & Peters, A. S. (2009). Semiology of epileptic seizures: a critical review. *Epilepsy & Behavior*, *15*, 2–9.

Pérez-García, F., Scott, C., Sparks, R., Diehl, B., & Ourselin, S. (2021). Transfer learning of deep spatiotemporal networks to model arbitrarily long videos of seizures. In *MICCAI* (pp. 334–344). Springer.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *IEEE CVPR* (pp. 1010–1019). volume 2016-Decem.

Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., & Zisserman, A. (2020). A Short Note on the Kinetics-700-2020 Human Action Dataset. *arXiv:1907.06987v1*, .

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–20).

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, *35*, 38571–38584.