

# Online Deep Squat Evaluation: Leveraging Subject-Specific Adaptation and Information Retention

Sara Sardari<sup>1,2</sup>, Bahareh Nakisa<sup>2</sup>, Sara Sharifzadeh<sup>3</sup>, Alireza Daneshkhah<sup>4</sup>, Seng W. Loke<sup>2</sup>, Michael J. Duncan<sup>5</sup>, Matteo Crotti<sup>6</sup>, Vasile Palade<sup>1</sup>

<sup>1</sup> Research Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK

<sup>2</sup> School of Information Technology, Faculty of Science Engineering and Built Environment, Deakin University, Geelong, Vic, Australia

<sup>3</sup> Department of Computer Science, Swansea University, Swansea, UK

<sup>4</sup> School of Mathematics and Data Science, Emirates Aviation University, Dubai, UAE

<sup>5</sup> Centre for Sport, Exercise and Life Sciences, Coventry University, Coventry, UK

<sup>6</sup> Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy

## ABSTRACT

Evaluating deep squats accurately during automatic physical rehabilitation monitoring across different subjects remains challenging due to inter-subject variability and limited labelled data. The challenges include: 1) conventional methods presuppose that a “one-model-fits-all” approach works for activity evaluation, ignoring that subject-specific differences can lead to suboptimal results if these differences are not considered. 2) Previous studies focus on offline learning, where models are trained on the entire dataset, which can be updated later through retraining. This approach neglects the need for continual learning, where models adapt sequentially to new subjects while retaining past knowledge to prevent catastrophic forgetting. This study addresses these challenges by proposing a novel continual meta-learning approach and a memory buffer to provide personalized deep squat evaluations. Using Azure Kinect sensors, we collected RGB-D videos and 3D skeletal data from 33 participants performing deep squats, annotated with Functional Movement Screen (FMS) scores. Our model dynamically adapts to new participants while retaining knowledge from previous ones, preventing performance degradation over time. Experimental results demonstrate that our approach outperforms a model without a buffer memory technique by retaining learned knowledge across participants and adapting to new individuals with minimal data.

## KEYWORDS

Action quality assessment, Skeleton data, Meta-learning, Continual Learning, Few-shot learning

## 1 Introduction

Physical rehabilitation involves carefully designing and formulating exercises to restore functional ability. Deep squatting is one crucial exercise often recommended in the rehabilitation process of patients to improve functional strength and flexibility and achieve complete recovery of the overall limb [1]. However, even healthy individuals have challenges while performing this activity since it requires strength and balance. Therefore, it is vital to evaluate the accuracy of this movement to obtain adequate therapeutic results without injury [2].

Experts use the FMS assessment tool to evaluate the quality of deep squat action [3]. FMS is a standardized tool for evaluating movement patterns and identifying mobility limitations. It assists

experts in determining the patient's functional problem and adjusting the rehabilitation program accordingly. These scores range from zero to three, where three is the best possible score. However, assessing the squats, even using a powerful tool such as FMS, can be time-consuming and require the expertise of a trained professional. Therefore, several studies focused on developing an automatic assessment system for deep squats [3-5]. These automatic models can assist the expert's judgment on the movement while providing fewer inconsistencies in the real-time feedback. These approaches include using data from motion-capturing sensors and applying an AI-driven model to assess movement.

In computer vision, action evaluators can support experts by providing objective FMS scores using vision-based data and deep learning classifiers [6], [7]. However, the complexity of the movement—due to various body limbs, joints, and high degrees of freedom—leads to significant variations in participant's actions and imbalanced performance classes. In addition, the fluctuations in human movement, flexibility, and reactions, particularly across different subjects, pose a significant challenge for developing a “one-model-fits-all” approach for FMS scoring. Previous studies assume that training a model on generalized deep squat data will yield good performance for unseen episodes from new subjects [3-5]. However, due to the diverse movement patterns among individuals, models often struggle to generalize, leading to confusion with new signals. This highlights the need for subject-specific adaptation techniques, where feedback is personalized for each individual to improve scoring accuracy and relevance.

In this work, we focus on meta-learning as the most common framework for few-shot learning, where the model learns to adapt to new tasks quickly (in this case, subjects) with only a few labelled samples [8], [9]. Unlike the Transfer Learning techniques, which need large data samples while pretraining the model and fail to generalize on new tasks with few samples, the meta-learning techniques learn rich knowledge from similar events and reuse the past knowledge to adapt to unseen samples incrementally [10]. A few studies have utilized different versions of meta-learning in Human Activity Recognition (HAR) (not activity evaluation). In a study by Wijekoon and Wiratunga [11], two personalized Model-Agnostic Meta-Learning (MAML) and Personalized Relation Networks models are optimized for activity recognition. In another study by Nafi and Hsu [12], meta-learning was utilized for video-

based HAR (MetaVHAR). Li et al. [13] proposed a federated meta-learning technique for HAR deployed on two datasets, illustrating better performance than baseline classifiers.

In addition to the previous challenge, former studies primarily rely on offline learning, where models are trained on the entire dataset and updated only by retraining on the new dataset. This approach overlooks the need for continual learning, where models adapt sequentially to new subjects while reusing past knowledge to prevent forgetting. Continual learning models [14-16] aim to retain past information during sequential learning of dynamically changing data. As a knowledge retention technique, the memory replay approach stores a subset of previous data in a memory buffer to consider along with the current data. For example, Duan et al. [14] utilized a memory buffer and formulated an optimization problem on adaptive hyperparameters for forgetting mitigation.

In this study, we collected deep squat movement data from 33 participants to explore the aforementioned challenges. We proposed combining continual learning with meta-learning to achieve personalized adaptation while preserving knowledge from previous subjects. Using affordable and accurate Azure Kinect sensors [17], [18], we captured RGB-D videos of participants performing deep squat actions and recorded their 3D positions of joints through time. This setup allows for non-intrusive, cost-effective, and privacy-preserving data collection, making it suitable for home-based monitoring and commercial applications [7]. After the preprocessing stage, two sports science experts annotated each video with an FMS score of the movement. The combined proposed method dynamically adapts to each subject by leveraging a meta-learning framework that learns how to learn from previous subjects and employs a memory buffer to retain knowledge from past subjects, mitigating the risk of forgetting. This approach allows rapid adaptation to new individuals with minimal data and ensures the model retains performance on previously learned participants.

During the adaptation phase in a meta loop, we incorporate mutual information loss between original support data and augmented support data. This base-level mutual information loss, calculated using Information Noise-Contrastive Estimation (InfoNCE), encourages the model to learn robust features invariant to small perturbations in the input data. By augmenting the support data with controlled noise and minimizing the distance between the original and augmented representations, the model learns to capture essential movement characteristics while ignoring irrelevant variations. This regularization aids in the model generalization of new movements and minimizes the dependency on the limited support samples. The model is then evaluated on the query set of the current participant. To ensure that the model retains knowledge from previous participants, we use a memory buffer that stores representations of support data from past participants. The meta-level mutual information loss is then calculated between the current query representations and the stored representations in the memory buffer to maintain consistency with previously learned knowledge. This helps the model remember what it has learned about past participants while adapting to new ones. The results have shown that our proposed technique is outperforming a model with similar architecture (missing the buffering technique), which illustrates the importance of preventing forgetting prevention. The summary of contributions are as follows:

1. **Personalized Adaptation:** In this study we developed a continual meta-learning framework that dynamically adapts to each individual subject, considering the inter-subject variability in deep squat performance.
2. **Knowledge Retention:** We integrated a memory buffer to prevent catastrophic forgetting to ensure consistent performance across sequentially learned subjects.
3. **Robust Feature Learning:** Leveraged InfoNCE loss and controlled data augmentation to enhance feature robustness and generalization.

The remainder of this paper delves into the details of the data collection and methodology, as well as the results and discussion.

## 2 Data collection

In this section, we provide information about the data collection strategy for capturing the skeleton information of deep squat actions.

We developed a dataset of RGB-D videos and related FMS scores for the deep squat action quality assessment. The data collection process has human ethics approval from both Coventry University (reference number: P131561) and Deakin University Human Ethics Advisory Group (reference number: SEBE-2022-12). The RGB-D video is captured using the latest Kinect technology, MS Azure Kinects, as illustrated in Figure 1. However, the data utilized in this paper include only the front view illustrated in this figure. Thirty-three healthy participants were asked to perform the deep squat in front of the sensor. Due to ethical considerations and limited patient access, we recruited healthy individuals to perform the same action, categorized as correct or incorrect, with five repetitions for each category. We provided videos of an expert performing perfect and imperfect versions of the action to guide the participants. They were asked to mimic both versions based on the reference. The description of the correct action is to bend the knees to descend the body toward the floor with the heels on the floor while the knees are aligned over the feet, and the upper body remains aligned in the vertical plane. For imperfect actions, the subjects will lean forward a bit to maintain balance, which prevents the heel bone from touching the ground.

We preprocessed the captured videos in two stages. First, two sports science experts carefully examined each video and labelled them based on the FMS scoring tools. To simplify the scoring process, the experts placed greater emphasis on evaluating the lower limbs of the participants. They labelled each activity into four classes of 0-3, where scores three and zero are assigned to the action with the best and lowest quality, respectively. Next, we captured the three-dimensional skeleton data of the spatial positions of 32 body joints using MS SDK and Python programming. The raw skeleton data needed spatial and temporal alignments.

The spatial alignment included considering the left foot joint in the origin with coordinates of (0, 0, 0) and rescaling and relocating the body skeletons. This is due to aligning the subjects with the exact video and body scaling coordinates. In addition, we considered temporal alignment (with interpolation) to have temporal data with the same number of frames. After preprocessing the 3D skeleton data, we had signals with 96 channels (32 joints with three

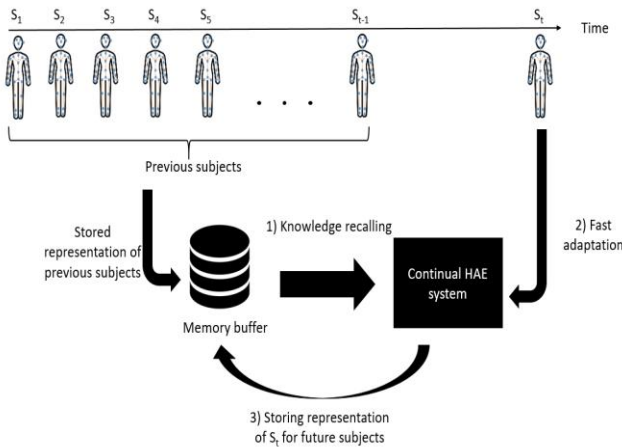
dimensions) with an equal number of frames. However, due to a limited number of samples and to avoid the possibility of confusion in model learning, we diminished the number of joints to be processed by the model. For each subject, we have only kept a subset of the 3D locations, including 11 major joints in the lower limb (knees, hips, and ankles) for several frames in addition to the corresponding FMS scores (0, 1, 2, and 3 classes to represent the quality of movement).



**Figure 1: Overview of the entire setup. Participant performing deep squat while three MS Azure Kinects are capturing the motion.**

### 3 Proposed continual meta-learning method

The continual meta-learning model (illustrated in Figure 2) that have been developed for FMS scoring operates by continuously adapting to new subjects while retaining information on previously learned subjects. This model leverages a meta-learning framework with a memory buffer to effectively handle the challenges of forgetting and subject-specific variations.



**Figure 2: The proposed framework of the continual meta-learning Human Action Evaluation (HAE) system with the knowledge retention and adaptation procedure.**

**Base Model architecture:** The architecture of the base model for training (often mentioned as the inner model or Meta model) discussed in the algorithm is LSTM-based to capture the temporal dynamics of human motion. The model consists of one LSTM layer with 64 units and L2 regularization of  $1e-4$ . Following that, two fully connected dense layers with 32 and 4 neurons with ReLU and Softmax activation functions were used, respectively, for multi-class classification.

**Continual Meta-Learning Framework:** In the proposed framework, the process involves two main phases of meta-training and testing. During the meta-training phase, a memory buffer (MBuffer) is utilized to store the adapted representations from previously seen subjects, allowing the model to balance learning new subjects and retaining previously acquired knowledge.

**Meta-training phase:** The meta-training phase begins by initializing the memory buffer and selecting the available subjects, excluding the test subjects. For each epoch, a meta loss is initialized to zero ( $l_{meta} = 0$ ). The training loop iterates through each available subject ( $S_i$ ), and within each subject, the subject dataset is split into support ( $S_{i_{sup}}$ ), and query sets ( $S_{i_{qu}}$ ). For each step in the inner loop, the inner model adapts to the current subject's support set. The adaptation involves calculating the cross-entropy (CE) loss between the model predictions and true labels, and the InfoNCE loss between the representations of the support set and its augmented versions. It should be noted that for creating new samples, controlled noise (adding only 1% noise to the actual value of the 3D axis) is applied to the 3D positions of randomly chosen joints. The total inner loss, defined as the sum of the CE loss and the InfoNCE loss, is used to update the inner model's parameters.

**Formulation of Loss Functions:** The total inner loss is computed as the sum of the cross-entropy loss and the InfoNCE loss:

$$\mathcal{L}_{Inner} = \mathcal{L}_{CE} + \mathcal{L}_{InfoNCE} \quad (1)$$

The cross-entropy loss measures the discrepancy between the predicted class probabilities and the true labels of the support data and is given by:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log(\hat{y}_{i,j}) \quad (2)$$

where,  $y_{i,j}$  represents the true label for the  $i^{th}$  sample and  $j^{th}$  class,  $\hat{y}_{i,j}$  is the predicted output for the  $j^{th}$  class for the  $i^{th}$  sample,  $N$  is the number of samples in the batch, and  $C$  is the number of classes.

The mutual information loss (InfoNCE) encourages the model to learn robust and distinct representations by maximizing the similarity between representations of the original and augmented data while minimizing similarity to negative samples. This loss is defined as:

$$\mathcal{L}_{InfoNCE} = -\sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau) + \sum_{j=1, j \neq i}^N \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j^+)/\tau)} \quad (3)$$

where,  $\mathbf{z}_i$  and  $\mathbf{z}_i^+$  are representations of the  $i^{th}$  original (anchor) and augmented ( $i^{th}$  anchor's positive pair) samples, respectively.  $N$  is the batch size,  $\mathbf{z}_j^+$  are negative samples that are not the anchors positive pair,  $\text{sim}(\cdot)$  denotes cosine similarity, and  $\tau$  is a temperature parameter. This loss function promotes the learning of discriminative features that are resilient to small perturbations in the input data.

After inner loop adaptation, the adapted support representations are stored in the memory buffer. The meta loop then evaluates the inner model on the query set of the current subject ( $S_{i_{Qu}}$ ). The query set loss is calculated, along with the InfoNCE loss between the query representations and those stored in the memory buffer. The total loss ( $l_{tot}$ ), which combines both losses, is added to the overall meta loss. After processing all subjects, the meta model's parameters are updated using the meta optimizer based on the computed meta loss gradients.

**Testing phase:** During testing, the trained meta model is adapted to the test subject's support data using inner steps similar to the training phase. After adaptation, the model is evaluated on the query data of the test subject to assess performance. The results of this accuracy as evaluation reflect the model's ability to generalize to new subjects while retaining its performance on previously learned tasks.

It is worth mentioning that the Adam weight decay, and SGD optimizers are utilized as optimizers in the inner and the meta loops, respectively.

#### Algorithm 1: Continual Meta-Learning with Mutual Information

##### Training phase:

Initialize MBuffer, available subjects excluding the test subject

For each epochs in range of epochs do:

$$l_{meta} = 0$$

For subjects in available subjects do:

For each batch in support set do

Split data to support and query sets

For each step in range of steps do:

Calculate CE loss of support set as  $l_{CE_s}$

Calculate InfoNCE loss of aug\_support and support

$$l_{inner_{total}} = l_{CE_s} + l_{InfoNCE_{s,a}}$$

Update Inner model parameters using gradients

Store adapted support representations in MBuffer

For each batch in query data do:

Calculate CE loss of query set as  $l_{CE_q}$

Calculate InfoNCE (MBuffer data and query) as  $l_{InfoNCE_{q,m}}$

$$l_{tot} = l_{CE_q} + l_{InfoNCE_{q,m}}$$

$$l_{meta} += l_{tot}$$

Compute gradients of Meta loss

Update Meta model using meta optimizer

Output: Trained Meta model ready for adaptation

##### Testing phase:

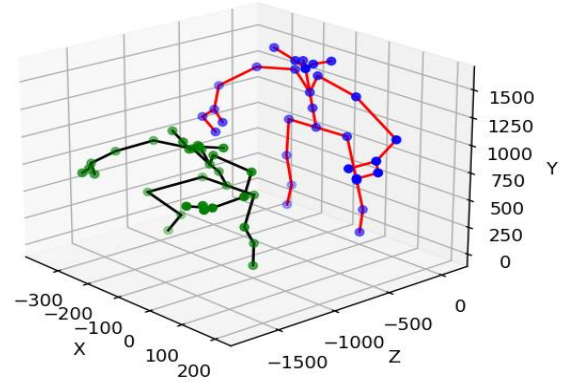
Split Test data into support data and query data

Adapt Meta model on support data using inner steps

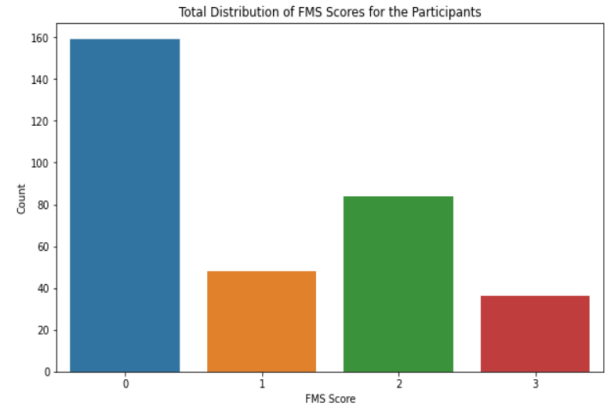
Evaluate on query data and report accuracy

## 4 Results and discussion

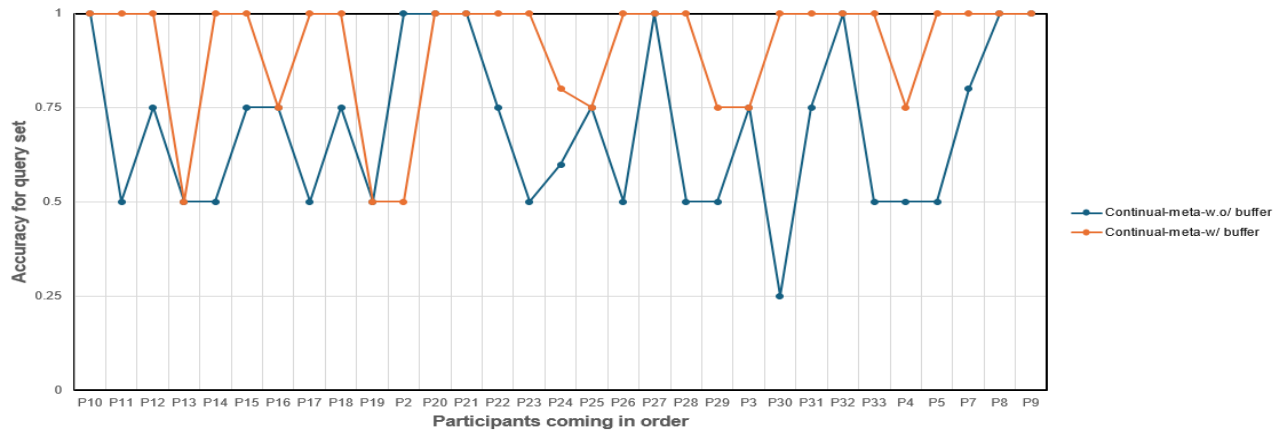
In this section, we discuss the tools utilized for developing the model and discuss the preliminary results. For this study, we utilized a combination of C++ programming and the Graphical User Interface (GUI) provided by Xing et al. [3] for capturing the RGB-D videos from participants. Later, Python (3.8) programming (using Microsoft Software Development Kits) assisted us in capturing the 3D positions of 32 joints. Finally, we utilized Python programming and the Tensorflow package for building the pipeline of the proposed model, including preprocessing, model training, and testing. We captured skeleton data from each repetition of the action and preprocessed them for spatial and temporal alignments. Figure 3 illustrates the skeleton data of the same participant in two separate frames. This figure shows the first (red body) and 40th (black) frames of a video.



**Figure 3:** This figure illustrates the participant performing the deep squat in two separate frames of same video. The red and black bodies illustrate the first and 40<sup>th</sup> frames, respectively.



**Figure 4:** The distribution of FMS scores for actions performed by participants.



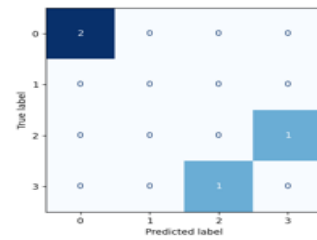
**Figure 5: The plot of performance comparison of two continual meta-learning techniques with and without buffer. The plots illustrates a good performance of model with buffer in FMS scoring and forgetting prevention.**

Figure 4 illustrates the distribution of the FMS scores for all of the deep squat movements. This figure depicts that most of the actions are scored as zero and very few of them tend to be a perfect performance of deep squat. This creates an imbalanced classification problem, which we handled it before further analysis. For solving this issue and limited data availability, we introduced a targeted data augmentation strategy. To mitigate this issue, particularly the underrepresentation of minority classes, we generated additional samples by adding controlled slight noises to the original joint position data. Therefore, we preserved the inherent patterns and temporal dependencies of action, while introducing slight variations to enhance model generalization.

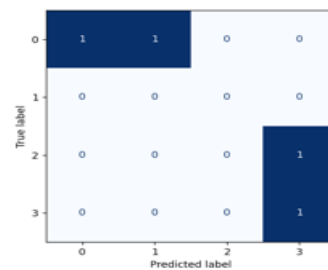
As mentioned before to the best of our knowledge this is the first time a continual meta-learning with buffer technique is utilized to provide subject-specific FMS feedback for the deep squat action performed by different subjects. Therefore, for comparison we have utilized the proposed architecture without the memory buffer technique to investigate the importance of the buffer in recalling the information of previous subjects learned by the model.

Figure 5 shows the accuracy of the two models being trained on different subjects in a sequential manner. For the first subject (P10) the models were trained on P1. Then for P11 the model was trained on the information of P1 and P10 and so on. The results depicted by this plot show that interestingly the buffer technique elevated the model’s performance in predicting the scores for future participants. It is shown that the model without the buffer technique especially in later subjects (P22 to P30) fails to recall the information of the previous subjects in the sequence, which it was being trained on. In addition, the proposed model with buffer outperforms the other technique on most of the participants even if the information of an action for a participant is confusing. The overall accuracies for the meta-continual learning with buffer and without buffer are 0.9 and 0.7, respectively. This illustrates the

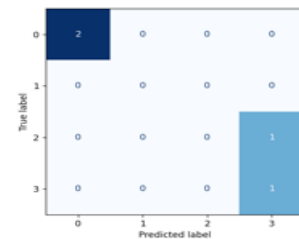
importance of the memory buffer technique in avoiding forgetting from past information.



A) Simple meta learning



B) Continual meta learning without buffer



C) Continual meta learning with buffer

**Figure 6: Confusion matrices of three simple meta-learning, continual meta-learning without buffer, and continual meta-learning with buffer for FMS scoring of queries in P4.**

For comparison of the continual meta-learning model with/without buffer with a non-continual technique we compared the proposed meta-learning model which trains the model on all of the participants and adapts for one test participant (P4) and gets evaluated on the query set. This technique misses the buffer and mutual information technique. Figure 6 illustrates the confusion matrices of three techniques (considering that this participant only has samples from class 0, 2, and 3 in query set). This figure shows that proposed model with buffer outperforms the non-continual technique even with not seeing all of the participants in the sequence. Considering that non-continual model learns from all of the participants, it is still failing on class 2 and 3. Our comparison with this non-continual meta-learning approach (trained on all participants at once) depicts that even with access to the entire dataset, the non-continual model struggles to handle unseen data effectively. Its confusion matrix highlights poor performance on higher class of FMS scores. This shows that the absence of a continual learning mechanism and mutual information loss hinders the model's ability to generalize to new tasks. This underscores the advantage of continual meta-learning in adapting dynamically to new subjects without sacrificing previously learned knowledge. In addition, based on Figure 6 the continual model without buffer is probably forgetting some of the information previous subjects.

For future studies, we aim to provide more ablation study on different versions of activities and architectures to illustrate the effectiveness of the continual meta-learning for human activity evaluation. In addition, we will explore the reason behind the confusion for simple meta-learning. In addition, we will explore other augmentation techniques such as GAN-based architectures to explore their effectiveness in improving the model's performance.

## REFERENCES

- [1] Hoogenboom, B.J., May, C.J., Alderink, G.J., Thompson, B.S. and Gilmore, L.A., 2023. Three-Dimensional Kinematics and Kinetics of the Overhead Deep Squat in Healthy Adults: A Descriptive Study. *Applied Sciences*, 13(12), p.7285.
- [2] Yoshiko, A. and Watanabe, K., 2021. Impact of home-based squat training with two-depths on lower limb muscle parameters and physical functional tests in older adults. *Scientific reports*, 11(1), p.6855.
- [3] Xing, Q.J., Shen, Y.Y., Cao, R., Zong, S.X., Zhao, S.X. and Shen, Y.F., 2022. Functional movement screen dataset collected with two azure kinect depth sensors. *Scientific Data*, 9(1), p.104.
- [4] Lee, J., Joo, H., Lee, J. and Chee, Y., 2020. Automatic classification of squat posture using inertial sensors: Deep learning approach. *Sensors*, 20(2), p.361.
- [5] Luna, A., Casertano, L., Timmerberg, J., O'Neil, M., Machowsky, J., Leu, C.S., Lin, J., Fang, Z., Douglas, W. and Agrawal, S., 2021. Artificial intelligence application versus physical therapist for squat evaluation: a randomized controlled trial. *Scientific Reports*, 11(1), p.18109.
- [6] Lei, Q., Du, J.X., Zhang, H.B., Ye, S. and Chen, D.S., 2019. A survey of vision-based human action evaluation methods. *Sensors*, 19(19), p.4129.
- [7] Sardari, S., Sharifzadeh, S., Daneshkhah, A., Nakisa, B., Loke, S.W., Palade, V. and Duncan, M.J., 2023. Artificial Intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review. *Computers in Biology and Medicine*, p.106835.
- [8] Hospedales, T., Antoniou, A., Micaelli, P. and Storkey, A., 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9), pp.5149-5169.
- [9] Chen, Y., Liu, Z., Xu, H., Darrell, T. and Wang, X., 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9062-9071).
- [10] Gharoun, H., Momenifar, F., Chen, F. and Gandomi, A., 2024. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12), pp.1-41.
- [11] Wijekoon, A. and Wiratunga, N., 2020. Learning-to-learn personalised human activity recognition models. *arXiv preprint arXiv:2006.07472*.
- [12] Nafi, N.M. and Hsu, W., 2023. MetaVHAR: Meta-Learning for Video-Based Human Activity Recognition.
- [13] Li, C., Niu, D., Jiang, B., Zuo, X. and Yang, J., 2021, April. Meta-har: Federated representation learning for human activity recognition. In *Proceedings of the web conference 2021* (pp. 912-922).
- [14] Duan, T., Wang, Z., Shen, L., Doretto, G., Adjeroh, D.A., Li, F. and Tao, C., 2024. Retain and Adapt: Online Sequential EEG Classification with Subject Shift. *IEEE Transactions on Artificial Intelligence*.
- [15] Aljundi, R., Belilovsky, E., Tuytelaars, T., Charlin, L., Caccia, M., Lin, M. and Page-Caccia, L., 2019. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32.
- [16] Lopez-Paz, D. and Ranzato, M.A., 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.
- [17] Kurillo, G., Hemingway, E., Cheng, M.L. and Cheng, L., 2022. Evaluating the accuracy of the azure kinect and kinect v2. *Sensors*, 22(7), p.2469.
- [18] Tölggyessy, M., Dekan, M. and Chovanec, L., 2021. Skeleton tracking accuracy and precision evaluation of Kinect V1, Kinect V2, and the azure kinect. *Applied Sciences*, 11(12), p.5756.