Assessment of Machine Learning Models for Classification of Movement Patterns During a Weight-Shifting Exergame

Elise Klæbo Vonstad, Beatrix Vereijken, Kerstin Bach, Xiaomeng Su, and Jan Harald Nilsen

***Abstract*—In exercise gaming (exergaming), reward systems are typically based on rules/templates from joint movement patterns. These rules or templates need broad ranges in definitions of correct movement patterns to accommodate varying body shapes and sizes. This can lead to inaccurate rewards and, thus, inefficient exercise, which can be detrimental to progress. If exergames are to be used in serious settings like rehabilitation, accurate rewards for correctly performed movements are crucial. This article aims to investigate the level of accuracy machine learning/deep learning models can achieve in classification of correct repetitions naturally elicited from a weight-shifting exergame. Twelve healthy elderly (10F, age 70.4 SD 11.4) are recruited. Movements are captured using a marker-based 3-D motion-capture system. Random forest (RF), support vector machine, k-nearest neighbors, and multilayer perceptron (MLP) are the employed models, trained and tested on whole body movement patterns and on subsets of joints. MLP and RF reached the highest recall and F1-score, respectively, when using combined data from joint subsets. MLP recall range are 91% to 94%, and RF F1-score range 79% to 80%. MLP and RF also reached the highest recall and F1-score in each joint subset, respectively. Here, MLP ranged from 93% to 97% recall, while RF ranged from 73% to 80% F1-score. Recall results, show that >9 out of 10 repetitions are classified correctly, indicating that MLP/RF can be used to identify correctly performed repetitions of a weight-shifting exercise when using full-body data and when using joint subset data.**

***Index Terms*—Classification, exergaming, machine learning, movement patterns, movement quality, reward systems, weight-shifting.**

I. Introduction

With the overall rise in gamification in recent years, serious games have been employed in a wide variety of fields, including education [1], professional training [2], [3], cognitive training [4], and physical exercise (e.g., [5]). Gamification refers to the introduction of elements from gaming, such as goals, reward systems, and challenges, into ordinary tasks to make them more fun and thereby increase motivation and adherence [6]. An essential element when designing serious games is how to determine whether the player's answer or action is correct and thus should be rewarded in the games. Typically, serious games predefine correct answers or actions, and track the performance of players directly using controllers, keyboards, or smartphones, allowing for relatively straight-forward checks of correctness. In serious games for exercise ("exergaming"), the player is interacting with the game using bodily movements that are captured by cameras or other devices [5]. Movements are subsequently assessed against predefined decision rules or thresholds, as seen in e.g., [7], [8], and rewards are given if these body parts performed as predefined, regardless of the correctness of the movements of other parts of the body.

As commercial exergames aim at being entertaining and easy to use, broad ranges and definitions of what is considered "correct" by the game are necessary to accommodate different body shapes and sizes. Because of these broad definitions, players often figure out quickly what the minimum required behavior is for receiving rewards [9]. When the game rewards the player even when performing the movements in this manner, players can easily cheat, or worse, not even know whether they were performing the movements correctly or incorrectly. For entertainment purposes, this may well be irrelevant. However, in the context of regaining or maintaining physical function, performing the correct movements is essential for effectiveness and progress [10]. Effective exercise depends on performing the necessary movements correctly, thus supervised exercise programs typically report better results than nonsupervised exercise programs [11].

For older adults, exergaming is regarded as a promising tool to deliver guided exercise without the presence of therapists or clinicians. Furthermore, exercise delivered through exergames has been shown to be more fun and motivating than traditional exercise [5], [12]. This could help increase adherence and motivation for exercising in older adults, which is a prerequisite for mediating the strain the ongoing demographic change will place on our health care systems [13]. Older adults often have different requirements for movements during exercise compared to healthy people, as they might have physical constraints due to ageing [14]. ColorRules and settings in exergames therefore need to be adapted to individual constraints and goals, but still allow for proper form and tempo to progress in training [15]. If

Manuscript received May 29, 2020; revised November 12, 2020 and January 19, 2021; accepted January 28, 2021. Date of publication March 18, 2021; date of current version May 19, 2021. This article was recommended by Associate Editor M. S. Neubert. *(Corresponding author: Elise Klæbo Vonstad.)*

Elise Klæbo Vonstad, Kerstin Bach, Xiaomeng Su, and Jan Harald Nilsen are with the Department of Computer Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway (e-mail: elise.k.vonstad@ntnu.no; kerstin.bach@ntnu.no; xiaomeng.su@ntnu.no; jan.h.nilsen@ntnu.no).

Beatrix Vereijken is with the Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway (e-mail: beatrix.vereijken@ntnu.no).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3059716.

Digital Object Identifier 10.1109/THMS.2021.3059716

exergames are to be effective in serious exercise settings such as rehabilitation, we need game systems that accurately identify and reward correctly performed movements to ensure efficiency and progress [16], [17].

One alternative to using broadly defined rules and thresholds to determine the correctness of a movement is to study the occurrences of movement patterns with good or poor quality and build models that embody the features of each of these. These models can then be used to assess movement quality, potentially with high accuracy, as the model is trained to recognize features of a correctly performed movement pattern without being fed predefined rules or thresholds. Recent developments in machine learning (ML)/deep learning (DL) have made it possible to efficiently analyze large amounts of data, which is promising for using high-volume data from whole-body movement patterns. Such models have been used successfully to recognize different everyday activities like walking, sitting, and lying down [18], [19], and movement patterns during traditional exercise (e.g., [20]). However, to the best of our knowledge, it has yet to be applied to assessment of movement pattern quality during exergaming.

A. Pilot Study

To study the suitability and potential of applying ML for our objective, we conducted a pilot study first to investigate whether ML models can distinguish between similar full-body movement patterns where some are performed correctly and others incorrectly [21]. In this pilot study, participants (N = 11, 6 F, mean age 69.3 years, SD 4.0) performed repetitions of weight-shifting movements where half of the movements were performed with clear incomplete weight shifts (i.e., incorrectly performed repetitions), and the other half with clear complete weight-shifts (i.e., correctly performed repetitions). Participants were instructed on how to perform the movements to ensure that the right movement patterns for incorrect and correct repetitions were recorded. A marker-based 3-D motion capture system (3DMoCap) was used to track participants' movements, and statistical features were calculated for each repetition. Three different ML models [Random forest (RF), support vector machine (SVM), and K-nearest neighbor (K-NN)] were trained and evaluated for classification performance using leave-one-group-out (LOGO) cross-validation. All three models achieved good performance (>90% accuracy, [18]). These results encouraged us to investigate whether ML models can accurately classify movements that are *naturally* elicited (i.e., not instructed) from a balancing exergame. As naturally elicited movements are more varied, both within and across participants, classification can be more challenging.

B. Aim of This Article

The present aticle investigates what level of F1-score and recall four different ML/DL models can achieve in classifying correctly performed whole-body and joint-subset movement patterns naturally performed during a balance exergame.

C. Article Organization

This article is organized as follows. Related work is outlined in Section II. The experimental set-up and data analysis procedures are described in Section III. Section IV presents results comparing four different ML models in the classification of movement correctness. Discussion of the results and limitations of the study are presented in Section V. Conclusion and future work are presented in Section VI.

II. Related Work

In general, exergaming for older adults is considered a promising tool for facilitating unsupervised exercise at home or in an elderly care center (e.g., [4], [22], [23]). Research has shown that exergames are effective in delivering exercise for several physical and mental functions, such as balance and postural control [24], gait [25], upper body movements [12], cognitive function [26], problem solving [27], and memory [28]. Exergames are also found to be more motivating and fun than traditional exercise [9], [29], which is an essential feature that could facilitate adherence and motivation for exercise [12]. In addition, the technologies that exergames are based on make it possible to tailor games to individual needs and goals [30], which is a major advantage that could make exergaming even more effective than traditional exercise. Furthermore, to ensure that exergames are appropriate for older adults, extensive research has been conducted into the design and usability requirements for this population, resulting in guidelines and design principles that apply to exergames for older adults [16], [31].

In recent years, there has been a proliferation of work implementing the (semi)automatic classification and recognition of actions and activities based on multimodal data recorded from human movement [18]. Although research on movement classification, as shown in [18], is an adjacent field of research, these models only focus on identifying *what* movement has been performed, not the *quality* of the movement (e.g. how well the movement was performed). We are particularly interested in assessing the quality of movement and will therefore focus primarily on related work that sheds light on evaluating movement quality.

High-quality research has been conducted with the aim of identifying errors in movement patterns compared to predefined movement templates [32]–[35], and rules/thresholds [7], [8], [36]. Movement performance compared to the predefined goal is used to provide feedback on how to improve movement patterns. Comparison of movements to thresholds and/or rules is also done in comprehensive work on modelling and evaluation of human movement, as seen in [15], [14], and [37].

Using template movements and decision rules can be appropriate for players that do not have physical constraints or do not need individual adaptation of movement patterns during exercise and are aiming to perform the exercises similarly to a healthy person. As mentioned, participants need to have goals that are adjusted to their needs and constraints, so comparing their movements to a healthy person or a template movement can be detrimental to motivation or might push them to perform the exercise outside their safe limits.

One earlier study also aimed to classify movement quality in a more naturally elicited, less instructed, fashion [38]. Here, exercise repetitions near exhaustion were used as examples of incorrectly performed movements and classified as correct or incorrect using ML models. This study was conducted on healthy children, using a smartphone (i.e, an inertial measurement unit) to capture movements.

In conclusion, we find that there is a wide variety of settings and contexts where automatic identification of movement errors during exercise is receiving attention, including technique analysis in general fitness and elite sports, as well as exercising for elderly at home or in rehabilitation centers. However, research into classification of movement quality specifically during exergaming is scarce, especially regarding identification of correctly and incorrectly performed movements.

Further, a large body of the related work demonstrated that errors in movement patterns can be identified during exercise by comparing performed movements to rules and template movements or expert scoring. Conversely, our study aims to build ML/DL models that can classify correctly performed movements that are naturally elicited, without comparing to a template movement or a set of rules or thresholds. Then, we assess the accuracy with which these models can identify correctly performed movements in unseen samples of the movement patterns.

III. Experimental Setup and Analysis: Assessing Movement Patterns Using ML

1) Participants: Participants were healthy older adults recruited from local exercise groups in the municipality. All participants gave their written, informed consent. There were 12 participants in total (10F); average age was 70.4 (SD 11.4) years (range 54–92). Average height and weight were 172.3 (SD 11.4) cm and 70.4 (SD 12.1) kg, respectively. Exclusion criteria were physical or cognitive injuries/impairments that affected their balance and gait ability, and age <50 or age >80 years. The project was approved by the Norwegian Regional Ethics Committee and the Norwegian Centre for Research Data (REK case number: 2017/2078-1).

2) Experimental Protocol: The experiment was conducted at the Motion Capture and Visualization Laboratory ("Vislab") at NTNU Trondheim in June 2019. A marker-based 3-D motion capture (3-DMoCap) system was used to measure participants' movements for use in analysis and classification. Four cameras (MX400, 90 Hz, Qualisys AB) were used. Thirty-six reflective markers were placed following the Plugin-Gait (PiG) marker placement protocol [39], excluding head and fingers.Two digital video cameras (Hero 3+ Black, 25 Hz, 1080p, GoPro Inc) captured movements in the sagittal and frontal planes of the player. Two 3-axial force plates (1000 Hz, 600x400x35 mm, Kistler Nordic AB) were located under the participants' feet to measure the ground reaction forces while playing. A platform matching the force plates' height was placed laterally of each force plate. The experimental setup can be seen in Fig. 1.

3) Game System: The game was built in Unity (v. 5, Unity Technologies, Denmark). As time-of-flight camera technology is commonly used in exergaming [5], [40], we used the Kinect

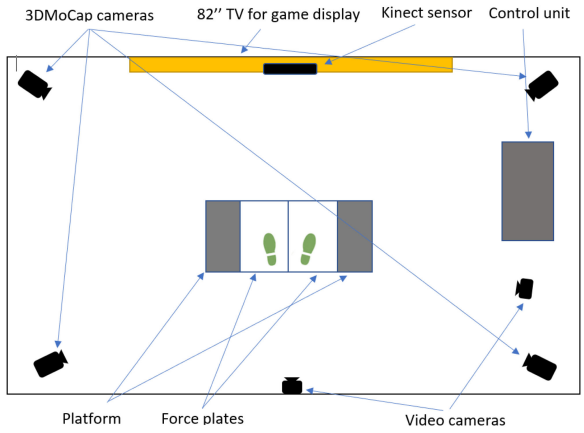

Fig. 1. Experimental setup.

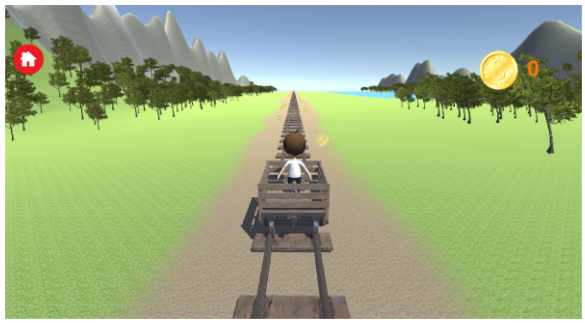

Fig. 2. Game interface.

v2 (30 Hz, Microsoft Inc), set up in front of the participants, to enable gameplay. The participants played three rounds of the two parts of the game, totaling six trials for each participant. If the movement tracking from the Kinect was not satisfactory, for example when the avatar did not follow the participants' movements, avatar movements were jittery, or if the sensor failed to identify the player at all, the trial was stopped and started again until smooth, continuous movement tracking from the Kinect was achieved.

The two parts of the game were designed to elicit different movement patterns from the players: the first aimed at having the player perform a complete, and thus correct, weight shift by moving their upper body over their weight-bearing foot. The second part was designed to make the player perform movements without moving their upper body over the weight-bearing foot, i.e., incompletely performed weight shifts. The game interface consisted of a rail cart with an avatar in it, representing the player, as shown in Fig. 2. On each side of the rail were coins which the player would try to hit with the cart as they moved along the rail. The cart tilted from side to side, following the medio-lateral leaning movements of the player. There were never more than two coins successively, and the coins appeared in random places for each participant. There were a total of approximately 100 coins in each game part, with approximately 50% of the coins on each side of the rail. The player was rewarded with points if they hit a coin with the cart, and the position of their upper body decided the amount of points rewarded in each of the game parts. There was a bar above the avatar. In part 1 the bar was grey, in part 2 the bar was multicolored as seen in Fig. 3. The grey

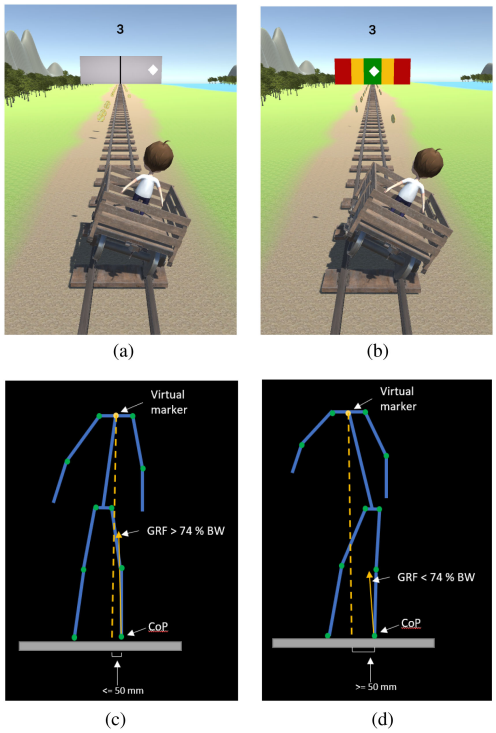

Fig. 3. (a) and (b) Two versions of the exergame. (c) and (d) Typical body postures when playing the two different exergame versions. (a) Part 1: Two-split grey bar, shown at the end of the track, with the star to the right of the dividing line, rewarding 3 points. (b) Part 2: Three-split color bar, shown at the end of the track, with the star in the middle 33%, rewarding 3 points. (c) Typical body posture when being rewarded 3 points in part 1 of the game. Here, the player is leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being <50 mm, and the GRF Z-component being >74% of body weight. BW = body weight. GRF = ground reaction force. CoP = center pressure. (d) Typical body posture when being rewarded 3 points in part 2 of the game. Here, the player is not leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being >50 mm, and the GRF z-component being <74% of body weight. GRF = ground reaction force. CoP = center of pressure.

bar was divided in the middle: if the star was on the line when a coin was hit, the player was rewarded 1 point. Three points (max score) were awarded if the star was as far away from the dividing line as possible, i.e., at any of the lateral parts of the grey bar as seen in Fig. 3(a). The multicolored bar was divided into three equally sized color fields: green in the middle 33%, yellow in the next 33% on each side, and red at the 33% most lateral fields. The red field rewarded 1 point, the yellow two points and the green three points, as seen in Fig. 3(b). Fig. 3(c) shows a typical posture form playing version 1, and Fig. 3(d) shows a typical posture from playing version 2 of the game.

4) Preprocessing: Joint center locations of shoulders (SHO), hips (HIP), knees (KNE) and ankles (ANK), as well as center of pressure (CoP), were extracted from the standard PiG biomechanical model from each of the six game trials for all participants. Game trials were then segmented into single medio-lateral movement repetitions using the peak-finding algorithm peakutils (v 1.3.3 for Python) on the *y*-axis of the right SHO joint in the Qualisys coordinate system. One repetition was defined as a continuous movement starting at the most lateral point of a medio-lateral movement, ending at the most lateral point on the opposite side. Python for Windows (v. 3.8.2) was used for all analyzes. An overview of the data analysis pipeline can be seen in Fig. 4.

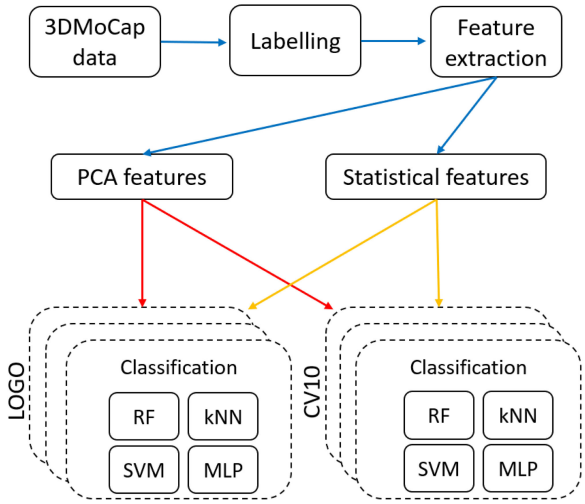

Fig. 4. Data analysis pipeline. The process, from "Feature extraction," was repeated for all joint data combined, and for each joint subset separately. PCA = Principal component analysis, RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation.

5) Labeling: The repetitions were subsequently assessed for the weight shift being correctly (i.e., a complete weight shift) or incorrectly (i.e., an incomplete weight-shift) performed. A physical therapist experienced in rehabilitation was consulted to determine the features of a correctly performed weight shift. The following criteria had to be met for a repetition to be deemed a correct weight shift. 1) The majority of the persons' body weight (over 74%, as 50% on each foot means that the person is standing with equal amount of weight on their feet) must be shifted to the weight-bearing foot. 2) The upper body must be moved over the weight-bearing foot as the weight is shifted. To evaluate whether condition 2 was met, a virtual marker was calculated as the 3-D midpoint between the left and right SHO, and the distance between the y-position of this virtual marker and the y-position of the CoP was calculated. Mean distance of <50 mm was required for the repetition to be deemed correctly performed. Sample videos form all participants were consulted to ensure that these criteria captured actual incorrectly and correctly performed movement patterns. All repetitions were assessed according to these criteria and assigned a target variable for incorrect (0) or for correct performance (1). This resulted in 2821 repetitions, where 1803 were labeled 1 (correct) and 1018 0 (incorrect).

6) Feature Extraction: After the target labels were assigned, statistical features were extracted from each repetition using the TSfresh library [41] (v. 12.0) for Python. See Appendix 1 for an exhaustive list of features. Furthermore, the feature dimensions were reduced using principal component analysis (PCA). Principal components that combined explained 95% of variance in the data were retained for further analysis.

7) Classification Models and Hyperparameter Tuning: Four models were employed in this study: RF, SVM, kNN, and an artificial neural network [multilayer perceptron (MLP)]. SciKit-Learn (version 22.1) for Python was used for analysis. RF is an

TABLE I
HYPERPARAMETER VALUES FOUND TO ACHIEVE THE BEST ACCURACY FROM GRIDSEARCHCV. RF = RANDOM FOREST, SVM = SUPPORT VECTOR MACHINE, KNN = K-NEAREST NEIGHBOR, MLP = MULTILAYER PERCEPTRON, LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD CROSS-VALIDATION

	Hyperparameter 1	Hyperparameter 2	Hyperparameter 3
	Name = Value	Name = Value	Name = Value
RF	Criterion = Entropy	Min. leaf = 4	Min.samples at split = 10
SVM	C = 0.01	Gamma = 0.01	
KNN	Leaf Size = 15	Metric = Manhattan	No. neighbors = 35
MLP	#Hidden layers = 50	Alpha = 0.01	

ensemble classifier that employs a set of decision trees to predict class labels, where each tree sees a random subset of features, and uses the majority class predicted by each tree's leaf nodes to classify a sample. Ensemble classifiers have been used successfully in similar work on movement quality (e.g, [42]) and in adjacent fields such as action classification [18], [19]. SVM is a linear model that finds the optimal line (or hyperplane) to separate classes, using the line/hyperplane that yields the largest support vectors (i.e., decision boundaries) between classes. SVM is often used in action recognition, as it is a powerful classifier [18], [19]. The kNN model evaluates the *(k)* nearest data points' class for each feature and classifies the sample based on the majority of these neighbors' class. kNN is a fast and simple, yet powerful classifier that has been used in adjacent work [15], [43]. MLP is a layered network of nodes that classifies samples based on activation of nodes in the "hidden" layers between the input and the output layer, using backpropagation to adjust weights and biases in the hidden layer nodes for each iteration of training. MLP requires more training data and processing power than ML methods, but often outperforms ML methods in action classification when provided with sufficient training data [18].

The optimal combination of hyperparameter tunings for each model [44], with regard to classification accuracy, was found using grid search (threefold CV) from the SciKit-Learn-pckage. Table I shows the hyperparameter tunings (that are not default for the models in the current SciKit-Learn version) that achieved the highest accuracy for each model. These hyperparameter tunings were used in subsequent analyzes.

8) Cross-Validation and Classification Procedures: The models were trained and tested using cross-validation (CV) by LOGO, and tenfold CV (CV10). LOGO entails training the model on all the data except one participant and using this participants' data as the testing set. CV10 creates ten random subsets of the data from all participants and holds one subset out for testing in each iteration. To simulate a situation where only subsets of joints are reliably tracked, each model was also trained and tested in the same manner by using only subsets of joint data, i.e., only ankle data, knee data, hip data, or shoulder data. Thus, all models were trained and tested on 20 different versions of the data set as seen in the last step of Fig. 4.

9) Evaluation: Model performance was evaluated using the F1-score and the recall. F1-score is an accuracy measure (the harmonic mean between precision and recall), which gives more useful insight into model performance in an imbalanced dataset than standard accuracy [45]. Recall, or sensitivity, is the true positive rate and describes the ratio of correctly identified positive samples out of all sampled classified as positive by the model. This is a useful measure as it says how many of the correctly performed repetitions were actually labeled as correct, i.e., how many of the correct repetitions a model identified as a correct repetition.

TABLE II
PERCENT F1-SCORE ACHIEVED ON JOINT SUBSETS [SHOULDER (SHO), HIP (HIP), KNEE (KNE), AND ANKLE (ANK) JOINTS]. MODELS ARE RANDOM FOREST (RF), SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBOR (KNN), AND ARTIFICIAL NEURAL NETWORK (MLP). THE FEATURE REPRESENTATIONS (FEATS) ARE STATISTICAL (STAT) AND PCA (PRINCIPAL COMPONENTS). CV = CROSS-VALIDATION: LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD. SD = STANDARD DEVIATION. M = MEAN. THE HIGHEST AVERAGE RECALL ACHIEVED BETWEEN JOINT SUBSETS (COLUMNS), AND HIGHEST AVERAGE BETWEEN THE MODELS (ROWS) ARE HIGHLIGHTED IN BOLD FONT. THE HIGHEST RECALL ACHIEVED WITHIN JOINT SUBSETS IS HIGHLIGHTED IN GREEN

	FEATS	CV	SHO (SD)	HIP (SD)	KNE (SD)	ANK (SD)	M(SD)
RF	Stat	LOGO	79.7 (10.8)	78.9 (11.4)	69.9 (18.5)	73.1 (10.0)	75.4 (12.7)
		CV10	79.2 (10.5)	77.1 (13.0)	75.9 (12.3)	74.8 (11.0)	77.4 (11.7)
	PCA	LOGO	77.3 (9.3)	76.7 (10.3)	76.2 (9.3)	75.3 (10.8)	76.8 (9.9)
		CV10	77.0 (10.2)	76.5 (9.9)	77.6 (9.9)	75.9 (9.2)	77.1 (9.8)
SVM	Stat	LOGO	76.3 (11.1)	68.0 (17.2)	64.7 (17.2)	64.7 (17.2)	70.0 (15.7)
		CV10	77.9 10.7)	72.8 (14.0)	72.5 (14.9)	72.5 (14.8)	74.7 (13.6)
	PCA	LOGO	76.0 (10.7)	67.5 (18.4)	64.5 (17.0)	61.9 (20.4)	69.2 (16.6)
		CV10	77.4 (10.7)	72.4 (13.4)	72.6 (14.5)	69.1 (11.7)	73.8 (12.6)
KNN	Stat	LOGO	79.8 (10.3)	76.9 (12.5)	75.9 (9.1)	75.9 (9.1)	77.7 (10.2)
		CV10	79.2 (8.5)	75.9 (9.8)	75.2 (11.0)	75.2 (11.9)	77.0 (10.1)
	PCA	LOGO	78.9 (10.0)	77.5 (11.6)	77.1 (8.7)	74.2 (9.2)	77.3 (9.9)
		CV10	78.0 (9.0)	77.0 (9.6)	76.2 (10.1)	75.7 (9.5)	77.0 (9.6)
MLP	Stat	LOGO	79.6 (10.0)	78.1 (10.1)	74.7 (12.7)	77.5 (10.8)	77.8 (10.9)
		CV10	79.9 (8.7)	77.7 (8.7)	76.2 (9.7)	76.6 (9.0)	**78.2 (9.0)**
	PCA	LOGO	79.7 (10.5)	77.9 (9.9)	76.3 (9.3)	77.5 (10.3)	78.1 (10.0)
		CV10	79.3 (8.2)	77.3 (9.2)	77.4 (9.5)	77.2 (8.9)	78.0 (8.9)
M(SD)			**78.4 (1.3)**	75.5 (3.4)	73.9 (4.0)	73.4 (4.4)	

TABLE III
PERCENT RECALL ACHIEVED ON JOINT SUBSETS [SHOULDER (SHO), HIP (HIP), KNEE (KNE), AND ANKLE (ANK) JOINTS]. MODELS ARE RANDOM FOREST (RF), SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBOR (KNN) AND ARTIFICIAL NEURAL NETWORK (MLP). THE FEATURE REPRESENTATIONS (FEATS) ARE STATISTICAL (STAT) AND PCA (PRINCIPAL COMPONENTS). CV = CROSS-VALIDATION: LOGO = LEAVE-ONE-GROUP-OUT, CV10 = 10-FOLD. SD = STANDARD DEVIATION. M = MEAN. THE HIGHEST AVERAGE RECALL ACHIEVED BETWEEN JOINT SUBSETS (COLUMNS), AND HIGHEST AVERAGE BETWEEN THE MODELS (ROWS) ARE HIGHLIGHTED IN BOLD FONT. THE HIGHEST RECALL ACHIEVED WITHIN JOINT SUBSETS IS HIGHLIGHTED IN GREEN

	FEATS	CV	SHO (SD)	HIP (SD)	KNE (SD)	ANK (SD)	M (SD)
RF	Stat	LOGO	87.5 (11.2)	88.3 (10.3)	79.1 (26.0)	82.9 (14.7)	84.4 (15.6)
		CV10	84.8 (8.9)	82.2 (14.1)	82.8 (13.9)	82.0 (12.3)	82.9 (12.3)
	PCA	LOGO	89.4 (7.3)	89.0 (8.4)	87.9 (14.8)	88.4 (10.4)	88.7 (10.2)
		CV10	87.7 (7.2)	87.5 (8.1)	89.2 (7.9)	88.5 (9.7)	88.2 (8.2)
SVM	Stat	LOGO	78.2 (14.4)	66.5 (21.9)	29.2 (64.7)	67.4 (29.2)	60.3 (32.5)
		CV10	79.0 (11.0)	70.6 (16.7)	73.5 (18.1)	73.5 (18.1)	74.2 (16.0)
	PCA	LOGO	77.8 (14.0)	66.0 (22.6)	66.6 (28.9)	61.5 (25.5)	68.0 (22.8)
		CV10	78.3 (11.2)	70.0 (15.9)	73.6 (18.0)	66.4 (12.7)	72.1 (14.4)
KNN	Stat	LOGO	87.3 (8.6)	83.9 (10.9)	84.5 (10.6)	84.5 (10.6)	85.1 (10.2)
		CV10	86.9 (6.4)	82.0 (8.6)	84.3 (15.1)	84.3 (15.1)	84.4 (11.3)
	PCA	LOGO	87.2 (8.6)	86.8 (8.2)	88.1 (10.0)	84.3 (10.3)	86.6 (9.3)
		CV10	85.9 (6.9)	85.4 (7.8)	86.5 (12.4)	88.1 (8.8)	86.5 (9.0)
MLP	Stat	LOGO	92.1 (7.9)	95.7 (5.1)	87.0 (19.4)	93.3 (8.6)	92.0 (10.3)
		CV10	90.8 (6.0)	94.9 (4.3)	89.0 (10.9)	92.7 (8.7)	91.8 (7.5)
	PCA	LOGO	94.5 (5.8)	96.3 (3.8)	90.9 (11.7)	96.4 (5.5)	94.5 (6.7)
		CV10	94.1 (4.4)	96.5 (3.0)	93.4 (6.6)	96.4 (3.2)	**95.1 (4.3)**
M(SD)			**86.3 (5.3)**	83.9 (10.1)	80.3 (14.9)	83.2 (10.4)	

IV. RESULTS

Results for *each joint subset* are presented with F1-score in Table II and recall in Table III. Results for *joint subsets combined* are shown with F1-score in Fig. 5 and recall in Fig. 6.

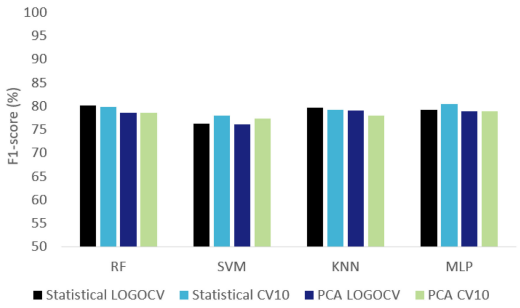

Fig. 5. F1-score achieved using different feature representations and CV methods on all joint subsets combined. RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation. PCA = principal components.

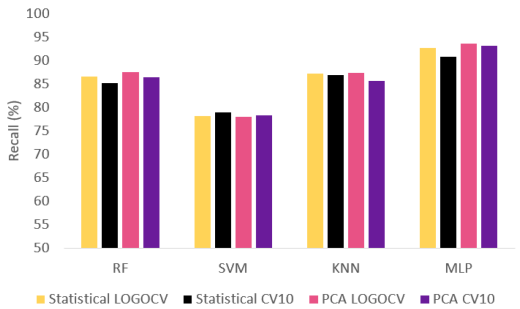

Fig. 6. Recall achieved using different feature representations and CV methods on all joint subsets combined. RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation. PCA = principal components.

A. *F1-Score*

The four models reached different levels of F1-score on different joint subsets of the data. Table II shows the F1-scores for each subset of joints in classifying correct repetitions, as well as the average performance of each joint subset. All models achieved similarly good results, with a mean of 75.3% (SD 11.3) for the F1-score. RF slightly outperformed other models on hip and knee joint subsets, while MLP performed best on shoulder and ankle joint subsets. Overall, the performance variation in using different feature representations or cross-validation methods was small. Somewhat surprisingly, the SVM achieved the overall lowest performance in terms of F1-score. All joint subsets also had high average F1-scores, with over 70%, but the SHO subset achieved the highest average with 78.4% (SD 1.3).

Fig. 5 shows the F1-score achieved by using all joint subsets combined, using different feature representations and cross-validation methods. These are results from all joint data only F1-score on joint subsets can be seen Appendix 2. Results show good performance from all models, with 78.5% (1.3 SD) F1-score on average. Different feature representations and cross-validation models are not affecting performance to any noteworthy degree.

B. *Recall*

Table III shows the recall achieved by the models on joint subsets of the data using different feature representations and CV methods. On average, the models achieved 83.3% (SD 17.6)

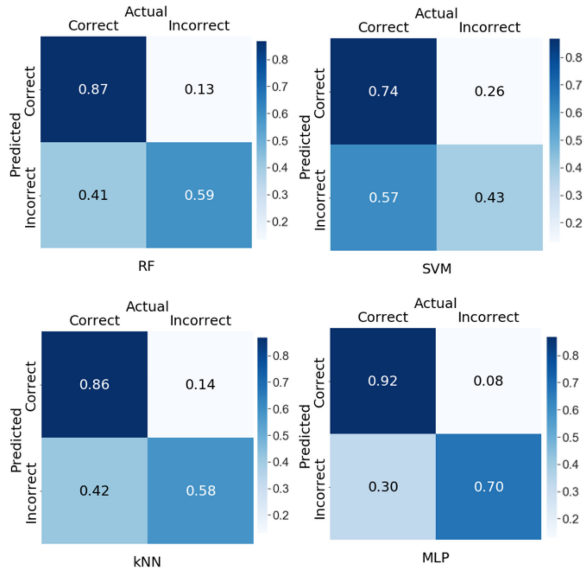

Fig. 7. Confusion Matrices for all models, with ratios of (clockwise from top left) true positive, false positive, true negative, and false negative predictions. Darker blue = higher ratio of samples predicted to belong in quadrant. Going clockwise from top left, the panels are for random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), and k-nearest neighbor (kNN).

recall (see Table III). The MLP outperformed the SVM and kNN models by 10%–25%, and was around 10% better than the RF model. Lowest recall was by SVM on the knee joint subset with statistical features and LOGO CV, with only 29.2%. On average, the SHO joint subset achieved the highest recall with 86.3% (SD 8.7) but other joint subsets also achieved high recall with >80%.

Fig. 6 shows the recall achieved by different feature representations and cross-validation methods. These are results from all joint subsets combined joint subset recall results can be seen Appendix 2. The MLP slightly outperformed the other models, with an excellent average of 92.6% (SD 1.1) recall. RF and KNN achieved comparable results, with an average of 86.5% (SD 0.8) recall and 86.9% (SD 0.7) recall, respectively. SVM was the overall lowest performing model in recall of correct repetitions, with an average of 78.4% (SD 0.3). Feature representation and CV methods showed only small differences, but PCA with LOGO was the marginally best configuration in three out of four models.

C. *Classification of Incorrect Repetitions*

Even though classification of correctly performed weight-shift repetitions may be sufficient for many applications, being able to accurately identify incorrect repetitions is important in a feedback perspective. An exergame system often needs to be able to identify e.g. an incomplete weight shift, and provide feedback to the player on how the movement pattern can be adjusted to achieve a complete weight shift. We analyzed the current models' ability to identify samples labeled as incorrect. This is not captured in metrics such as F1-score and recall, as they attenuate the influence of true negative samples. As seen in Fig. 7, incorrect samples were not classified with as high accuracy as correct samples, although the MLP achieved 70%

TABLE IV
PERFORMANCE OF EACH MODEL AND CROSS-VALIDATION METHOD IN MEAN TIME CONSUMPTION FOR TRAINING AND PREDICTION. RF = RANDOM FOREST, SVM = SUPPORT VECTOR MACHINE, KNN = K-NEAREST NEIGHBOR, MLP = MULTILAYER PERCEPTRON, LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD CROSS-VALIDATION

		Training time (s, SD)	Prediction time (ms, SD)
kNN	LOGO	0.8 (0.1)	1738 (476)
	CV10	0.1 (0.1)	153 (15)
SVM	LOGO	9.9 (0.7)	843 (220)
	CV10	0.8 (0.1)	72 (5)
RF	LOGO	8.3 (0.3)	15 (4)
	CV10	2.8 (0.2)	15 (1)
MLP	LOGO	7.0 (1.2)	2 (0.5)
	CV10	3.3 (0.1)	1 (0.5)

accuracy. Overall, models were able to classify about half of the incomplete weight shifts correctly.

V. DISCUSSION AND LIMITATIONS

In this article, we investigated the level of recall and F1-score the employed ML/DL models achieved in classification of correctly performed weight-shifting exercise repetitions, naturally elicited while playing a balancing exergame.

A. Correct Weight Shifts

Classification of correctly performed whole-body movement patterns is found to be feasible for all models used in this study, arriving at results ranging between 70%–80% F1-score (Table II) and 75%–95% recall (Table III). The best performing models in our study achieve over 90% recall and around 80% F1-score, which demonstrates that these models could be used in real-world applications for medio-lateral balance exercises. Although there are few directly comparable studies, our results show that using MLP or RF for classification of correct repetitions is in line with the state-of-the-art activity classification systems as reported in [18], [19], and [46]. Even though some of the models did not perform at a satisfactory level, we showed that the best performing models are promising in settings where it is useful to be able to receive feedback on movement pattern quality without having a clinician present, such as in home exercise.

The recall achieved by all models show that 90%–95% (see Fig. 6) of the correctly performed repetitions were, in fact, identified as such, which in an exergame situation would imply rewarding the player for close to all correctly performed repetitions. In other words, only a rather low number of correct repetitions were missed by the models. This is an indication that the models accurately captured and represented the movement features of a correct weight shift, without using manually designed rules or thresholds. This work echoes the results in [20] and [46].

The different classification models performed with slightly different results, as seen in Figs. 5 and 6. When it comes to computational performance, the models performing best on average, RF and MLP, were also the most efficient in training and prediction in terms of time usage (see Table IV). kNN was very fast in training, but slowest in prediction with >1.5 s used for each LOGO iteration, which is likely due to kNN having to build the model for each datapoint. As expected, SVM was the slowest in terms of training time, as well as being slow in prediction time. The distance-based models (kNN and SVM) often perform worse in terms of classification accuracy when the number of features is large compared to the number of samples [47], as a complex feature space makes it difficult to define decision boundaries that separate classes. The high MLP performance is likely due to the manner MLP models adjust the weights and biases in an iterative manner for a given classification problem by using gradient descent [48]. As such, MLP models also intuitively assess importance of different features during training. This is similar to what RF models do: features with high importance for the given classification task are used in early splits. Furthermore, features are used in a random fashion in the different decision trees, which contributes to high performance despite a complex feature space. This is also possibly the situation with the current dataset. The overall high recall can be attributed to the high quality of the data; low levels of noise have been shown to improve model performance [18], [49]. These results suggest that RF is likely the model that should be considered in similar applications for the following reasons: 1) RF achieves high recall; 2) RF is considered a "white box", e.g., it is possible to extract the decision making process in situations where transparency in the decision process is required; 3) the computational cost of prediction in RF is low, especially compared to MLP. These three features are likely of importance for a ML/DL system to be usable in e.g., a clinical or rehabilitation exercise setting. However, as the No Free Lunch theorem suggest, and as is shown in these results, there is no one model that is universally "best" for all problems (e.g., joint subsets). The model that performs best on average might not always be the best performing model in all problem subsets [50]. This indicates that it is necessary to evaluate the specific problem at hand, and how different models perform with the given data types, available computational power and noise level.

Results from the two cross-validation experiments are promising with respect to classification of previously unseen movement patterns. The models' performance did not worsen when classifying movement patterns from a participant that the models were not trained on. This is evident as the LOGO method performs similarly to the CV10 method, which holds out random subsets of all participants' data. Such similarity might be explained in two ways. 1) Participants performed the correct movement patterns similarly. 2) The models were indeed not overfitting, but truly and accurately captured and represented the features for correctly performed movement patterns to a good enough degree to identify unseen data with high accuracy. The practical implication of such models is that people who have not been playing a game using these assessment models before, will receive rewards when performing weight-shifting movements correctly. This is in line with the findings in [18]. Authors of [35] similarly found that using different neural network configurations with LOGO cross-validation produced good results. This further supports our findings that a person can use such a game system even though

the employed model for assessing movement pattern quality has not seen his/her movement patterns before.

When looking at results from separate joint subsets, shoulder movement patterns produced the best results in both F1-score and recall. This suggests that the shoulder movement pattern is the most relevant in assessment of weight shifting, and should be included to ensure high classification accuracy. Overall, using joint subsets, our models also achieved a level of performance (about 75% F1-score and 83% recall) comparable to other classification models using joint subsets [18]. One might argue that using any of these joint subsets could provide accurate rewards in weight-shifting exergames. Whole-body movement patterns still achieve slightly better results than joint subsets, both in terms of F1-score and recall, indicating that whole body movement patterns might still be a preferred setup if the primary goal is to achieve the best quality assessment possible. However, if the available tracking method only allows for accurate tracking of subsets of joints, using subsets is nonetheless a worthy alternative (even a preferred one if and when any cost benefit consideration renders the whole body tracking setup unsuitable) as it still achieves a very good classification accuracy of correct movement patterns using those subsets.

Regarding feature representation, there is no clear indication of any of the methods producing superior classification results. This suggests that statistical features are representing the exercise repetitions well, and that the principal components explaining 95% of the variance in the feature data sufficiently represent the latent information in the statistical features. PCA might be preferable over statistical features in future use, as they are lower dimensional and thus more computationally efficient.

B. Incorrect Weight Shifts

Being able to identify and provide feedback on erroneous movement patterns is useful in serious exergaming situations like rehabilitation, as exergames could be used to guide rehabilitation exercises without the presence of a clinician. The player would then need feedback on how to improve their movement pattern (such as having a larger range of motion, or moving faster) in order to perform the exercise in a efficient manner. In earlier work, where samples were labeled by error class, error types were classified with 85%–95% accuracy [42], [51]. The results from classification of incorrect repetitions in the current study support this notion that classification models needs to be trained on erroneous movement patterns that are labeled by error type, in order to construct representations of the error types in the features. Hence, actively classifying incorrect samples should be the goal of classification systems aiming for use in feedback during exergaming in rehabilitation settings. The current dataset does not contain enough samples of different error types, and is therefore not suited for such analyzes. Furthermore, the movement patterns in the erroneous repetitions probably vary significantly between participants, making it challenging to find robust representations of incorrect repetitions in the features. This also indicates that the features in the current study might not capture the information required for the models to represent an incorrect repetition, as some incorrect repetitions might have very similar movement patterns to correct repetitions. Still, the MLP is able to classify incorrect samples with 70% accuracy, as seen in Fig. 7, indicating that DL models might be usable for such tasks in future work.

C. Limitations

There are some limitations to this study that are necessary to keep in mind. Because this study included 12 participants only moving in a single plane, it is important to keep limitations of applicability of our results in mind. The movement performed is restricted to a medio-lateral weight shifting exercise, which is (ideally) confined to movement in the frontal plane of the body, so movements in other planes or in combinations of planes might be more difficult to classify correctly. Even though our results are promising, further research should be conducted to investigate the performance of these ML/DL models in more complex and challenging settings. Furthermore, data from other motion capture tools that are commonly available should be evaluated as this might impact classification performance.

VI. Conclusion

In conclusion, this study shows that RF and MLP are able to identify correctly performed weight-shifting repetitions with high recall and F1-score. In the development of exergame systems we should consider using the best models presented here for evaluating movement patterns, especially when aiming to reward players for correctly performing exercise repetitions in weight-shifting exercises. We showed that training ML/DL models using labeled training data is a feasible option for identifying correctly performed movement patterns, which can subsequently be used to reward players in an accurate manner during exergaming. This is an important improvement of many existing exergame systems that are based on comparisons to templates, or assessments using coarse rules and thresholds. Moreover, implementing a self-learning approach based on our work can allow a system to learn new movements without requiring a priori explicit identification of their templates. Trusting that the game system is actually rewarding the correct movements is a prerequisite for using exergames in serious settings like physical rehabilitation or independent exercise for older adults. If the game system is trusted, the threshold for using exergame systems might be lower for both users and clinicians, making it possible to benefit from higher motivation and adherence in the rehabilitation process. In future work, the implementation of the present classification models into game systems would be an interesting next step, possibly testing differences in rewards and/or feedback compared to rule-based or template-based systems. Exploring features is also a natural next step. The results of this study also warrant further investigation into how well these models perform in patient populations with more variable movement patterns, and in classification of error types. Furthermore, other movement patterns are also interesting to examine for classification accuracy, especially more complex movements that combine movements in various anatomical planes.

APPENDIX A
FEATURES

TABLE V
FEATURES CALCULATED FROM TSFRESH

Variable	Parameters/Units
Variance	
Standard deviation	
Mean	
Maximum	
Minimum	
Sum of values	
Count below mean	
Count above mean	
Sum reoccurring values	
Longest strike above mean	
Has duplicate values	True, False
Kurtosis	
Skewness	
Complexity invariance distance	True, False
Absolute sum of changes	
Change quantiles	Var,mean
Max Langevin Fixed Point	
Fourier Transform Coefficient	Abs,angle,real,imag
Fourier Transform Aggregated	Skew,centroid, kurtosis,variance
Mean absolute change	
Quantile	Q 0.1-0.9
Spektral Welch Density	Coeff 2,5,8
Large sd	R 0.01,0.05,0.25
Variance larger than sd	True,False
Binned entropy	Max bins 10
Number crossing m	-1,0,1
Range count	Max 1, min-1
Value count	0
Ratio beyond r sigma	0.5, 1.5, 5
Linear trend	P-value,intercept,slope
Aggregate linear trend	Max, min, mean
Quantile	
Has duplicate minimum	
Has duplicate maximum	
First location of minimum	
Last location maximum	
Last location minimum	
Has duplicate maximum	
Has duplicate minimum	
First location minimum	
Quantile	0.1-0.9
Autocorrelation	Lag 1-9
Agg autocorrelation	Mean,median,var
Partial autocorrelation	Lag 1-9
Absolute energy	
Continous wavelet transform	Width, peaks
Autoregressive AR(k)	2,3,4
Count above mean	
Augmenter dickey fuller	p-value, teststat
Energy ratio by chunks	
Friedrich Coefficients	
% of reoccurring values	
Value to time series length	Ratio
Number of peaks	
Mean second derivative central	
Index mass quantile	

APPENDIX B
JOINT SUBSET CLASSIFICATION RESULTS

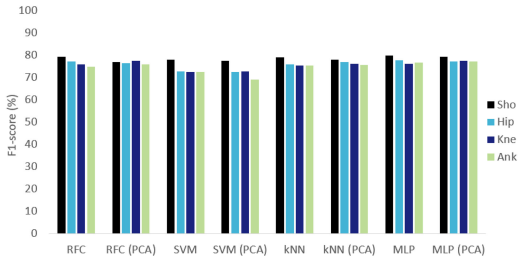

Fig. 8. F1-score achieved using different feature representations with CV10 on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

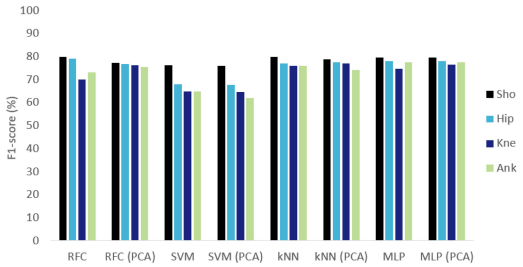

Fig. 9. F1-score achieved using different feature representations with LOGO on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

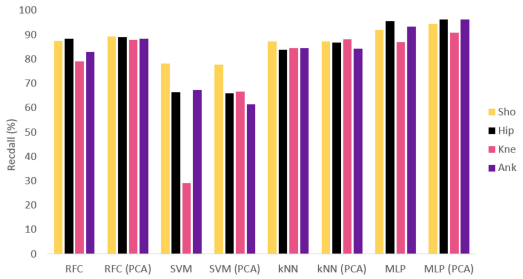

Fig. 10. Recall achieved using different feature representations with LOGO on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

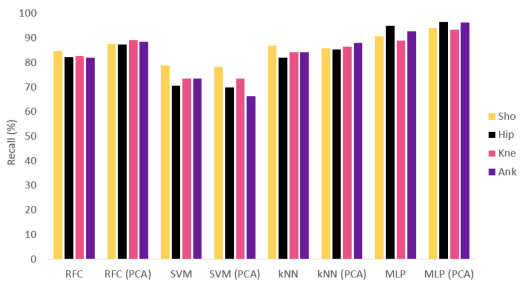

Fig. 11. Recall achieved using different feature representations with CV10 on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

ACKNOWLEDGMENT

The authors would like to thank A. Øvreness and M. A. Berge for assisting in data collection, C. Velvin for game development, and R. Stock for valuable discussions on weight shifting in physical rehabilitation.

REFERENCES

[1] C. Girard, J. Ecalle, and A. Magnan, "Serious games as new educational tools: How effective are they? A meta-analysis of recent studies," *J. Comput. Assist. Learn.*, vol. 29, no. 3, pp. 207–219, 2013.

[2] A. Ahmed and M. J. Sutton, "Gamification, serious games, simulations, and immersive learning environments in knowledge management initiatives," *World J. Sci., Technol. Sustain. Develop.*, vol. 14, no. 2/3, pp. 78–83, 2017.

[3] M. Graafland, J. M. Schraagen, and M. P. Schijven, "Systematic review of serious games for medical education and surgical skills training," *Brit. J. Surgery*, vol. 99, no. 10, pp. 1322–1330, 2012.

[4] T. M. Fleming *et al.*, "Serious games and gamification for mental health: Current status and promising directions," *Front. Psychiatry*, vol. 7, Jan. 2017, Art. no. 215.

[5] N. Skjæret, A. Nawaz, T. Morat, D. Schoene, J. Lægdheim, and B. Vereijken, "Exercise and rehabilitation delivered through exergames in older adults: An integrative review of technologies, safety and efficacy," *Int. J. Med. Inform.*, vol. 85, no. 1, pp. 1–16, 2016.

[6] S. Deterding, "Gamification: Designing for motivation," *Interactions*, vol. 19, no. 4, 2012, Art. no. 14. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2212877.2212883

[7] N. Gal, D. Andrei, D. I. Nemeş, E. Ndşan, and V. Stoicu-Tivadar, "A Kinect based intelligent e-rehabilitation system in physical therapy," *Studies Health Technol. Inform.*, vol. 210, pp. 489–493, 2015.

[8] W. Zhao, A. M. Reinthal, D. D. Espy, and X. Luo, "Rule-based human motion tracking for rehabilitation exercises: Realtime assessment, feedback, and guidance," *IEEE Access*, vol. 5, pp. 21382–21394, 2017.

[9] M. Pasch, N. Bianchi-Berthouze, B. van Dijk, and A. Nijholt, "Movement-based sports video games: Investigating motivation and gaming experience," *Entertainment Comput.*, vol. 1, no. 2, pp. 49–61, 2009.

[10] L. H. Skjaerven, K. Kristoffersen, and G. Gard, "An eye for movement quality: A phenomenological study of movement quality reflecting a group of physiotherapists' understanding of the phenomenon," *Physiotherapy Theory Pract.*, vol. 24, no. 1, pp. 13–27, 2008.

[11] A. Lacroix, T. Hortobágyi, R. Beurskens, and U. Granacher, "Effects of supervised vs. unsupervised training programs on balance and muscle strength in older adults: A systematic review and meta-analysis," *Sports Med.*, vol. 47, no. 11, pp. 2341–2361, 2017.

[12] J. D. Smeddinck, M. Herrlich, and R. Malaka, "Exergames for physiotherapy and rehabilitation: A medium-term situated study of motivational aspects and impact on functional reach," in *Proc. ACM CHI'15 Conf. Human Factors Comput. Syst.*, 2015, vol. 1, pp. 4143–4146. [Online]. Available: https://dx.doi.org/10.1145/2702123.2702598

[13] J. R. Beard and D. E. Bloom, "Towards a comprehensive public health response to population ageing," *Lancet*, vol. 385, no. 9968, pp. 658–661, 2015.

[14] F. Ofli, G. Kurillo, Š. Obdržálek, R. Bajcsy, H. B. Jimison, and M. Pavel, "Design and evaluation of an interactive exercise coaching system for older adults: Lessons learned," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 201–212, Jan. 2016.

[15] A. W. Lam, D. Varona-Marin, Y. Li, M. Fergenbaum, and D. Kulić, "Automated rehabilitation system: Movement measurement and feedback for patients and physiotherapists in the rehabilitation clinic," *Human-Comput. Interact.*, vol. 31, no. 3/4, pp. 294–334, 2016.

[16] M. Pirovano, E. Surer, R. Mainetti, P. L. Lanzi, and N. Alberto Borghese, "Exergaming and rehabilitation: A methodology for the design of effective and safe therapeutic exergames," *Entertainment Comput.*, vol. 14, pp. 55–65, 2016.

[17] E. J. Lyons, "Cultivating engagement and enjoyment in exergames using feedback, challenge, and rewards," *Games Health J.*, vol. 4, no. 1, pp. 12–18, 2015.

[18] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, May 2016.

[19] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 1192–1209, 2013.

[20] A. Depari, P. Ferrari, A. Flammini, S. Rinaldi, and E. Sisinni, "Lightweight machine learning-based approach for supervision of fitness workout," in *Proc. IEEE Sensors Appl. Symp., Conf.* 2019, vol. 5, pp. 1–6.

[21] E. K. Vonstad, X. Su, B. Vereijken, J. H. Nilsen, and K. Bach, "Classification of movement quality in a weight-shifting exercise," *in Proc. CEUR Workshop.*, 2018, vol. 2148, pp. 27–32.

[22] J. Wiemeyer and A. Kliem, "Serious games in prevention and rehabilitation—A new panacea for elderly people?," *Eur. Rev. Aging Physical Activity*, vol. 9, no. 1, pp. 41–50, 2012.

[23] E. Flores, G. Tobon, E. Cavallaro, F. I. Cavallaro, J. C. Perry, and T. Keller, "Improving patient motivation in game development for motor deficit rehabilitation," in *Proc. Int. Conf. Adv. Comput. Entertainment Technol.*, Jan. 2008, pp. 381–384. [Online]. Available: https://portal.acm.org/citation.cfm?doid=1501750.1501839

[24] M. van Diest, C. C. Lamoth, J. Stegenga, G. J. Verkerke, and K. Postema, "Exergaming for balance training of elderly: State of the art and future developments," *J. Nanoeng. Rehabil.*, vol. 10, no. 1, 2013, Art. no. 101.

[25] I. J. M. de Rooij, I. G. L. van de Port, and J.-W. G. Meijer, "Effect of virtual reality training on balance and gait ability in patients with stroke: Systematic review and meta-analysis," *Physical Therapy*, vol. 96, no. 12, pp. 1905–1918, 2016.

[26] E. F. Ogawa, T. You, and S. G. Leveille, "Potential benefits of exergaming for cognition and dual-task function in older adults: A systematic review," *J. Aging Physical Activity*, vol. 24, no. 2, pp. 332–336, 2016.

[27] Y. Gao and R. L. Mandryk, "The acute cognitive benefits of casual exergame play," in *Proc. Conf. Human Factors Comput. Syst.*, 2012, pp. 1863–1872.

[28] M. Adcock, F. Sonder, A. Schättin, F. Gennaro, and E. D. De Bruin, "A usability study of a multicomponent video game-based training for older adults," *Eur. Rev. Aging Physical Activity*, vol. 17, no. 1, pp. 1–15, 2020.

[29] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis, "Towards understanding the effects of individual gamification elements on intrinsic motivation and performance," *Comput. Human Behav.*, vol. 71, pp. 525–534, 2017. [Online]. Available: https://dx.doi.org/10.1016/j.chb.2015.08.048

[30] S. Göbel, S. Hardy, V. Wendel, F. Mehm, and R. Steinmetz, "Serious games for health - personalized exergames," in *Proc. ACM Multimedia Int. Conf.*, 2010, pp. 1663–1666. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1873951.1874316

[31] K. Gerling, I. Livingston, L. Nacke, and R. Mandryk, "Full-body motion-based game interaction for older adults," in *Proc. ACM Annu. Conf. Human Factors Comput. Syst.*, 2012, pp. 1873–1882. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2207676.2208324

[32] M. Antunes, R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Visual and human-interpretable feedback for assisting physical activity," in *Proc. Comput. Vision Workshops*, 2016, pp. 115–129.

[33] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 556–571.

[34] X. Yu and S. Xiong, "A dynamic time warping based algorithm to evaluate kinect-enabled home-based physical rehabilitation exercises for older people," *Sensors (Switzerland)*, vol. 19, no. 13, 2019, Art. no. 2882.

[35] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 468–477, Feb. 2020.

[36] N. A. Borghese, M. Pirovano, P. L. Lanzi, S. Wüest, and E. D. de Bruin, "Computational intelligence and game design for effective at-home stroke rehabilitation," *Games Health J.*, vol. 2, no. 2, pp. 81–88, 2013.

[37] L. Tao *et al.*, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Comput. Vision Image Understanding*, vol. 148, pp. 136–152, 2016.

[38] L. D. M. Carvalho and V. Furtado, "Using machine learning for evaluating the quality of exercises in a mobile exergame for tackling obesity in children," in *Proc. SAI Intell. Syst. Conf.*, 2016, vol. 16, pp. 373–390. [Online]. Available: https://link.springer.com/10.1007/978-3-319-56994-9

[39] *Plug-in Gait Reference Guide*, Vicon Motion Systems Ltd., Oxford, U.K., 2016.

[40] A. Da Gama, P. Fallavollita, V. Teichrieb, and N. Navab, "Motor rehabilitation using kinect: A systematic review," *Games Health J.*, vol. 4, no. 2, pp. 123–135, 2015.

[41] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," in *Proc. ACML Workshop Learn. Big Data*, Nov. 2016, pp. 1–17. [Online]. Available: https://arxiv.org/abs/1610.07717

[42] P. E. Taylor, G. J. Almeida, J. K. Hodgins, and T. Kanade, "Multi-label classification for the analysis of human motion quality," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 2214–2218.

[43] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-D posture data," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 5, pp. 586–597, Oct. 2015.

[44] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated Machine Learning*. Cham, Switzerland: Springer, 2019, pp. 3–33.

[45] C. J. Van Rijsbergen, *Information Retrieval*. 2nd ed. London, U.K.: Butterworth, 1979.

[46] E. Zdravevski *et al.*, "Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering," *IEEE Access*, vol. 5, pp. 5262–5280, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7883880/

[47] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[48] D. P. Kingma and J. Lei Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR 2015)*, 2015, *arXiv:1412.6980*.

[49] A. Atla, R. Tada, V. Sheng, and N. Singireddy, "Sensitivity of different machine learning algorithms to noise," *J. Comput. Sci. Colleges*, vol. 26, no. 5, pp. 96–103, 2013.

[50] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 161–168.

[51] A. Yurtman and B. Barshan, "Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals," *Comput. Methods Programs Biomed.*, vol. 117, no. 2, pp. 189–207, 2014.