Routledge
Taylor & Francis Group

Journal of Sports Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjsp20

Comparison of machine learning classifiers for differentiating level and sport using movement data

Gwyneth B. Ross, Allison L. Clouthier, Alistair Boyle, Steven L. Fischer & Ryan B. Graham

To cite this article: Gwyneth B. Ross, Allison L. Clouthier, Alistair Boyle, Steven L. Fischer & Ryan B. Graham (2022) Comparison of machine learning classifiers for differentiating level and sport using movement data, Journal of Sports Sciences, 40:19, 2166-2172, DOI: 10.1080/02640414.2022.2145430

To link to this article: https://doi.org/10.1080/02640414.2022.2145430

JOURNAL OF SPORTS SCIENCES
2022, VOL. 40, NO. 19, 2166–2172
https://doi.org/10.1080/02640414.2022.2145430

Routledge
Taylor & Francis Group

SPORTS MEDICINE AND BIOMECHANICS

Comparison of machine learning classifiers for differentiating level and sport using movement data

Gwyneth B. Ross[a], Allison L. Clouthier[a], Alistair Boyle[b], Steven L. Fischer[c] and Ryan B. Graham[a,c]

[a]School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, Ottawa, Ontario, Canada; [b]Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada; [c]Department of Kinesiology, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

The purposes of this study were to determine if 1) recurrent neural networks designed for multivariate, time-series analyses outperform traditional linear and non-linear machine learning classifiers when classifying athletes based on competition level and sport played, and 2) athletes of different sports move differently during non-sport-specific movement screens. Optical-based kinematic data from 542 athletes were used as input data for nine different machine learning algorithms to classify athletes based on competition level and sport played. For the traditional machine learning classifiers, principal component analysis and feature selection were used to reduce the data dimensionality and to determine the best principal components to retain. Across tasks, recurrent neural networks and linear machine learning classifiers tended to outperform the non-linear machine learning classifiers. For all tasks, reservoir computing took the least amount of time to train. Across tasks, reservoir computing had one of the highest classification rates and took the least amount of time to train; however, interpreting the results is more difficult compared to linear classifiers. In addition, athletes were successfully classified based on sport suggesting that athletes competing in different sports move differently during non-sport specific movements. Therefore, movement assessment screens should incorporate sport-specific scoring criteria.

ARTICLE HISTORY
Received 4 December 2021
Revised 29 July 2022
Accepted 25 October 2022

KEYWORDS
Recurrent neural networks; time-series; movement screens; reservoir computing; long short-term memory

1. Introduction

Movement screens are used to identify aberrant movement patterns believed to increase risk of injury and/or impede performance. Individuals' kinetic and/or kinematic data can be recorded as they perform standardised motions, where those data can be evaluated to determine the correctness and/or proficiency of movement (Cook et al., 2014; Donà et al., 2009; Kritz et al., 2009; McCall et al., 2014; McCunn et al., 2016; Padua et al., 2009). A criticism of movement screens is that they are not able to predict injury risk (McCunn et al., 2016), which is thought to be due to a lack of sensitivity within the scoring criteria (Frost et al., 2015). The scoring criteria do not account for natural variability between athletes (Frost et al., 2015), which may be due to skill level or sport played. Another limitation of movement screens is the poor inter- and intra-rater reliability due to scores being based on visual appraisal (Gulgin & Hoogenboom, 2014; McCunn et al., 2016; Onate et al., 2012; Smith et al., 2013).

To combat the current issues surrounding movement screens, our previous research has focused on the development of an objective movement screening tool that combines kinematic data and machine learning with the aim to create a tool that is able to better predict injury risk based on movement competency. For training machine learning algorithms for scoring movement competency, our previous research found that for differentiating elite and novice athletes, linear traditional machine learning classifiers outperformed other non-linear traditional machine learning classifiers both when using optical motion capture data and simulated inertial measurement unit (IMU) data (Ross et al., 2020). While past research highlights the utility of linear classifiers in the context of athlete movement screening, classification rates ranged from 78.1% to 84.7% depending on the movement task (Ross et al., 2020). Emerging methods that better consider the time-series nature of movement may permit better classification.

Traditionally, time-series data have been considered one of the most challenging problems for machine learning as they tend to be complex, high-dimensional, highly correlated (Esling & Agon, 2012), inherently noisy (Yang & Wu, 2006), and prone to overfitting (Esling & Agon, 2012). Previous research has focused on the development of techniques, such as principal component analysis (PCA) and recurrent neural networks (RNNs) to combat traditional issues with time-series data. In RNNs, the connections between nodes form a directed graph along a temporal sequence, which allows the network to learn dynamic, temporally-varying behaviour (Hochreiter & Schmidhuber, 1997; Lukoševicius & Jaeger, 2009; Troje, 2002a; Weng & Shen, 2008).

RNNs can then use past information along with current input to calculate an output, which works well when the gap between the current input and the relevant stored past information is small; however, if the gap is too large, the relevant information is lost (Hochreiter & Schmidhuber, 1997). Long short-term memory networks (LSTMs) are a type of RNN that

CONTACT Ryan B. Graham rgraham@uottawa.ca School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, 125 University Private, Ottawa ON K1N 6N5 Ontario, Canada
Supplemental data for this article can be accessed online https://doi.org/10.1080/02640414.2022.2145430

were developed to retain relevant information regardless of the sequential gap size (Hochreiter & Schmidhuber, 1997) by incorporating memory cells and gate units. Another limitation of traditional RNNs is that backpropagation is computationally expensive as it backpropagates through time requiring large amounts of computing power and time to train the networks (Lukoševicius & Jaeger, 2009). To address this problem, reservoir computing (RC) was developed. RC randomly generates the recurrent part (the reservoir) and then keeps it fixed with only the weights of the readout layer being trained, unlike the traditional RNN that trains the weights of each layer (Lukoševicius & Jaeger, 2009), decreasing the computational needs. However, it is unknown if RNNs will outperform the previously used traditional machine learning classifiers.

In addition to the inclusion of time-series specific algorithms, the creation of demographic-specific algorithms could improve classification rates. Currently, popular movement screens do not consider athlete-specific demographics such as competition level, sport, age, or sex, which has been noted as a possible explanation for why movement screens cannot predict injury (Frost et al., 2015). Previous research has found that during running, using a single accelerometer placed on the hip (Kobsar et al., 2014), runners can be differentiated from soccer players; however, it is unknown whether athletes competing in different sports move differently during non-sport specific movement screens. Before creating demographic-specific algorithms, it needs to be determined if machine learning algorithms can classify athletes performing a non-sport specific movement screen based on sport. If athletes are able to be separated by demographic, then demographic-specific algorithms could be useful in better predicting injury.

Therefore, the purpose of this study was two-fold: 1) to determine if RNN models designed for time-series analyses (i.e., LSTM and RC) can outperform the previously used linear and non-linear classifiers at classifying athlete skill level, and 2) to determine if athletes can be differentiated based on the sport played, using the same technique, and if so, to identify which machine learning algorithm(s) discussed previously perform(s) the best. To the best of the authors' knowledge, this is the first study to attempt to classify athletes based on sport played using movement patterns as input data.

2. Methods

2.1. Participants

Kinematic data were collected from 542 athletes ranging in competition level from youth to professional (e.g., NBA, NFL, FIFA, MLB) and competing in 11 different sports (baseball, basketball, soccer, golf, tennis, track and field, squash, cricket, lacrosse, football, or volleyball). However, due to collection error, obstructed data, and not all athletes performing all tasks, the number of athletes per task varied (see Supplemental Material, Tables S3 & S4). Before data collection, each athlete read and signed an informed consent form permitting the data to be used for secondary analyses. The Health Sciences Research Ethics Board at the University of Ottawa approved the secondary use of the data for research purposes (file no: H-08-18-1085). Athletes competing at the intercollegiate, semi-professional, or professional level were considered elite athletes, whereas athletes competing at all other levels were considered novice. For classifying sports, only elite athletes competing in football (n = 128), baseball (n = 96), basketball (n = 37), and soccer (n = 33) were analysed, as these were the sports with the greatest number of athletes. For the classification of sport, after constraining to elite athletes from these four sports, all retained athletes were male (see supplemental material, Table S4).

2.2. Protocol

Before data collection began, each athlete provided an injury history from the past 10 years, had their height (with shoes), and weight taken. For the data collection, the athletes were outfitted with 45 passive, reflective markers (B&L Engineering, Santa Ana, CA) to capture whole-body kinematics (McPherson et al., 2016; Ross et al., 2018, 2020). Once outfitted with the markers, the athlete completed both a static and dynamic calibration trial.

After the calibration trials, each athlete completed a movement screening battery consisting of 21 unique movements that were selected to test each athletes' range of motion at each joint and their global stability, power, and balance. For this study, only seven movements were used in the analysis. The seven tasks retained were the bird-dog, drop-jump, hop-down, L-hop, lunge, step-down, and T-balance, with each movement, except the drop-jump, being performed bilaterally, resulting in a total of 13 movement trials (Ross et al., 2018, 2020). For each bilateral movement, "left" or "right" is used in conjunction with the movement task name when discussing movement tasks to clarify between the two movements. Only one trial for each task was retained, with the athlete performing the task until they believed that they had performed it to the best of their ability. Kinematic data were captured using an 8-camera Raptor-E (Motion Analysis Corporation, Santa Rosa, CA) motion capture system at a rate of 120 Hz.

2.3. Data Analysis

2.3.1. Pre-processing

All motion capture data were collected, labelled, and gap-filled using Cortex (Motion Analysis Corporation, Santa Rosa, CA). The data were then exported to Visual3D (C-Motion, Inc., Germantown, MD) and a 3D model was developed using the static calibration trial and applied to the motion capture data to calculate joint centres bilaterally for the wrist, elbow, shoulder, foot, ankle, knee, and hip; centres of gravity for the trunk, head, and pelvis; and marker positional data for the left and right heel, T2, T8, sternum, and the back, front, and sides of the head. Data were then exported to Python and all trials were trimmed to specific start- and end-point criteria (Ross et al., 2018), and filtered using a zero-phase, fourth-order, low-pass Butterworth filter with a cut-off frequency of 15 Hz. Due to there being significant differences in height between groups (level: $F = 138.25$, $p < 0.001$; sport: $F = 12.32$, $p < 0.001$), the data for each movement were normalised to each athlete's height. The 3D positional data were then rotated and translated so that the local coordinate system of the trunk was aligned with the global coordinate system with the origin of the coordinate system aligning with the midpoint between the left and right hip of the first frame of the data. The rotated data were then time-normalised to 500 frames.

2.3.2 Traditional Machine Learning

The time-normalised data were then placed in an [n × 39,000] matrix for each movement, where n was the number of subjects and 39,000 was the time-normalised data for the x, y, and z-axes for each joint centre, centre of gravity, and retained markers (500 timepoints × 3 axes × 26 positions). PCA was then applied to each matrix, resulting in a unique model per task per classifier. Using the principal components (PC) scores as features to classify skill level and sport, ensemble feature selection using the Python scikit-learn library (Pedregosa et al., 2011) with Pearson Correlation, chi-squared, recursive feature elimination, lasso, random forest, and LightGBM was used to rank the PCs based on contribution for each movement task and classifier (i.e., level and sport). The PCs that ranked in the top 25 for at least 50% of the techniques (i.e., 3) were retained for that task and classifier (Ross et al., 2020). To minimize overfitting, the maximum number of PCs retained was less than or equal to the square root of the number of subjects for that task and classifier.

To classify the data, scores for the PCs retained from the ensemble feature selection were used as inputs and either sport (basketball/football/baseball/soccer) or level (elite/novice) was used as the class. Seven classifiers were used: binary (level) or multinomial (sport) logistic regression (MLR), decision tree classifier (DT), *k*-Nearest Neighbours (KNN), naïve bayes (NB), linear discriminant analysis (LDA), and support vector machines with a linear (SVM) and radial basis function (RBF) kernel. All classifiers were programmed using the scikit-learn library (Pedregosa et al., 2011).

2.3.3. RNNs

For the time-series-specific classifiers (RC and LSTM), for each task, a 3D matrix was constructed that was [n × 500 × 78], where n was the number of subjects, 500 was the number of frames, and 78 was the number of positions times the number of axes (26 × 3). The class labels were then encoded into one-hot arrays. For both RC (Bianchi et al., 2021) and LSTM (Clouthier et al., 2021), the hyperparameters were tuned using grid search (see Supplemental Material, Tables S5 & S6) and a train-test split of 80–20.

2.3.4. Validation

All classifiers were validated using 10-fold cross-validation with 80% of the data being used for training and 20% for testing. Due to unequal group sizes for both level and sport, the training and testing sets were assigned the same ratio of elite and novice and basketball, football, baseball, and soccer players as the overall sample. For the traditional machine learning algorithms, feature selection and PCA were fit on the training data only and were refit with each cross-validation. Lastly, a naïve algorithm was used that predicted all athletes in the testing set as the majority group for that training set. For example, if the training set for predicting level was 40% novice and 60% elite, the naïve classifier would predict that 100% of the testing set were elite. For all classifiers, accuracy and time to complete each cross-fold were recorded and the mean and standard deviation across all tasks were calculated for each classifier. In addition, the maximum random access memory (RAM) required across all tasks for each classifier was recorded.

2.3.5. Statistics

In SPSS 20.0 (IBM, Armonk, New York, USA), Shapiro-Wilks tests were run to test for normality ($p > 0.05$). With the data being normally distributed, one-way ANOVAs with Tukey post-hoc tests were run in SPSS for each class and movement task between the 10 different classifiers to test for significant differences ($p < 0.05$) in accuracy.

3. Results

3.1. Time

For both level and sport, RC took the least amount of time to train with an average time of 0.16 ± 0.08 seconds and 0.12 ± 0.06 seconds, respectively (Table 1). The traditional classifiers all had similar training times with average times of 25.6 ± 11.7 seconds (level) and 81.6 ± 13.7 seconds (sport; Table 1). Due to the LSTM being computationally expensive and in order to reduce training times, all LSTM models were trained using an NVIDIA Titan RTX GPU and had average times of 76.6 ± 52.0 seconds and 23.4 ± 18.9 seconds for level and sport, respectively, where 69.1 GB (level) and 88.4 GB (sport) of RAM were required (Table 1). All other classifiers were less computationally expensive and were able to be trained in a timely fashion using an Intel® Xeon® Gold 6248 CPU and 384GB of ECC RAM. In terms of memory, the traditional classifiers all required the same amount (level: 2.4 GB, sport: 15.8 GB).

Table 1. The computer architecture, the average training time in seconds per cross-fold validation and standard deviation, and the maximum memory needed across all tasks for level and sport.

		Level		Sport	
ML Algorithm	Computing Architecture	Time (s)	Memory (GB)	Time (s)	Memory (GB)
RC	CPU	0.16 (0.08)	12.6	0.12 (0.06)	48.0
LSTM	GPU	76.7 (52.0)	69.1	23.4 (18.9)	88.4
LDA	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8
MLR	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8
SVM	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8
KNN	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8
DT	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8
NB	CPU	25.6 (11.7)	2.4	81.5 (13.7)	15.8
RBF	CPU	25.6 (11.7)	2.4	81.6 (13.7)	15.8

RC required 12.6 GB and 48.0 GB for level and sport, respectively (Table 1).

3.2. Level

Overall, the RNNs (LSTM and RC) and the linear classifiers (LDA, MLR, and SVM) outperformed the non-linear classifiers (KNN, DT, NB, and RBF) and the naïve classifier (Figure 1). Within the non-linear classifiers, RBF and NB outperformed KNN and DT. There was a significant main effect between classifiers for all tasks ($p < 0.001$; Table 2). Since similar results were found between tasks, only the results of the post-hoc tests for the lunge right will be presented in text (see Supplemental Material; Tables S9 for the results of all other post-hoc tests). Looking at just the average classification rates for the lunge right, RC (77.9% ± 4.7%), LSTM (79.0% ± 3.8%), LDA (76.9% ± 3.9), MLR (77.0% ± 3.8%), and SVM (76.3% ± 3.1%) all had significantly greater classification rates than KNN (66.0% ± 4.3%; $p < 0.001$), DT (62.6% ± 4.6%; $p < 0.001$), and naïve (65.4%; $p < 0.001$); however, there were no significant differences between RC, LSTM, LDA, MLR, or SVM ($p > 0.05$; Table 3). RBF (73.6% ± 4.1%) had significantly greater classification rates than KNN ($p = 0.001$), NB (72.6% ± 3.9%; $p < 0.001$), naïve ($p < 0.001$; Table 2) and NB had significantly greater classification rates than DT ($p < 0.001$; Table 3).

3.3. Sport

For classifying sport, the results closely resembled classification of athlete's level. Overall, the RNNs and linear classifiers outperformed the non-linear classifiers and the naïve classifier, and within the non-linear classifiers, RBF and NB outperformed DT and KNN (Figure 2). There was a significant main effect between the different classifiers for all tasks ($p < 0.001$; Table 2). Similar to level, similar results were found between tasks, therefore only the results of the post-hoc test for the lunge right will be

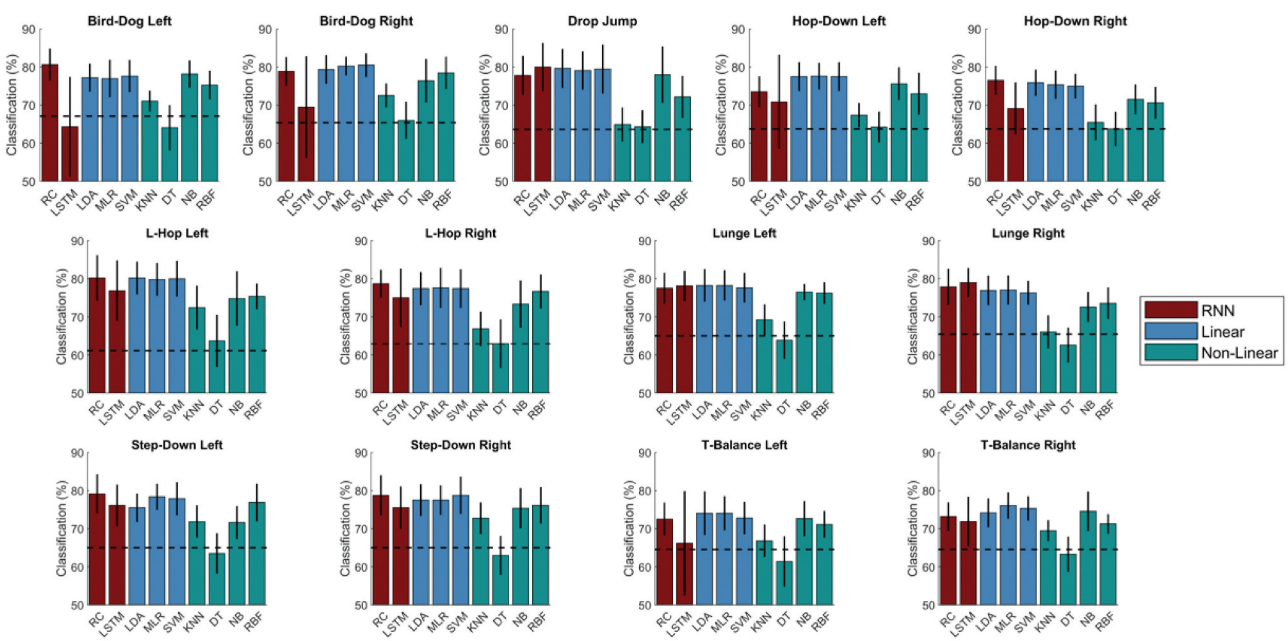

Figure 1. The mean classification rates and standard deviations when classifying for level for all movement tasks and all machine learning classifiers. Maroon bars represent recurrent neural network classifiers, blue bars represent linear machine learning classifiers and green bars represent non-linear machine learning classifiers. The dotted black line denotes the naïve classification rates.

Table 2. The overall results of the one-way ANOVAs for each movement task for level and sport. Main effect is denoted by F and significance level is denoted as p.

	Level		Sport	
	F	p	F	p
Bird-Dog Left	12.01	<0.001	11.851	<0.001
Bird-Dog Right	11.427	<0.001	22.948	<0.001
Drop-Jump	17.352	<0.001	15.131	<0.001
Hop-Down Left	10.123	<0.001	16.96	<0.001
Hop-Down Right	14.733	<0.001	14.821	<0.001
L-Hop Left	15.921	<0.001	11.282	<0.001
L-Hop Right	14.74	<0.001	8.811	<0.001
Lunge Left	24.558	<0.001	21.324	<0.001
Lunge Right	23.761	<0.001	20.994	<0.001
Step-Down Left	16.063	<0.001	15.94	<0.001
Step-Down Right	14.964	<0.001	20.939	<0.001
T-Balance Left	5.453	<0.001	26.053	<0.001
T-Balance Right	12.561	<0.001	20.599	<0.001

Table 3. The results of the Tukey post-hoc test for the lunge for level. The orange cells denote lunge left, whereas the blue cells denote lunge right.

	Lunge									
	RC	*LSTM*	*LDA*	*MLR*	*SVM*	*KNN*	*DT*	*NB*	*RBF*	*Naïve*
RC		1	1	1	0.995	<0.001	<0.001	0.076	0.276	<0.001
LSTM	1		0.967	0.978	0.855	<0.001	<0.001	0.012	0.063	<0.001
LDA	1	1		1	1	<0.001	<0.001	0.276	0.644	<0.001
MLR	1	1	1		1	<0.001	<0.001	0.241	0.595	<0.001
SVM	1	1	1	1		<0.001	<0.001	0.496	0.855	<0.001
KNN	<0.001	<0.001	<0.001	<0.001	<0.001		0.595	0.009	0.001	1
DT	<0.001	<0.001	<0.001	<0.001	<0.001	0.045		<0.001	<0.001	0.819
NB	1	0.992	0.986	0.986	1	0.001	<0.001		1	0.03
RBF	0.999	0.978	0.967	0.967	0.998	0.002	<0.001	1		<0.001
Naïve	<0.001	<0.001	<0.001	<0.001	<0.001	0.234	1	<0.001	<0.001	

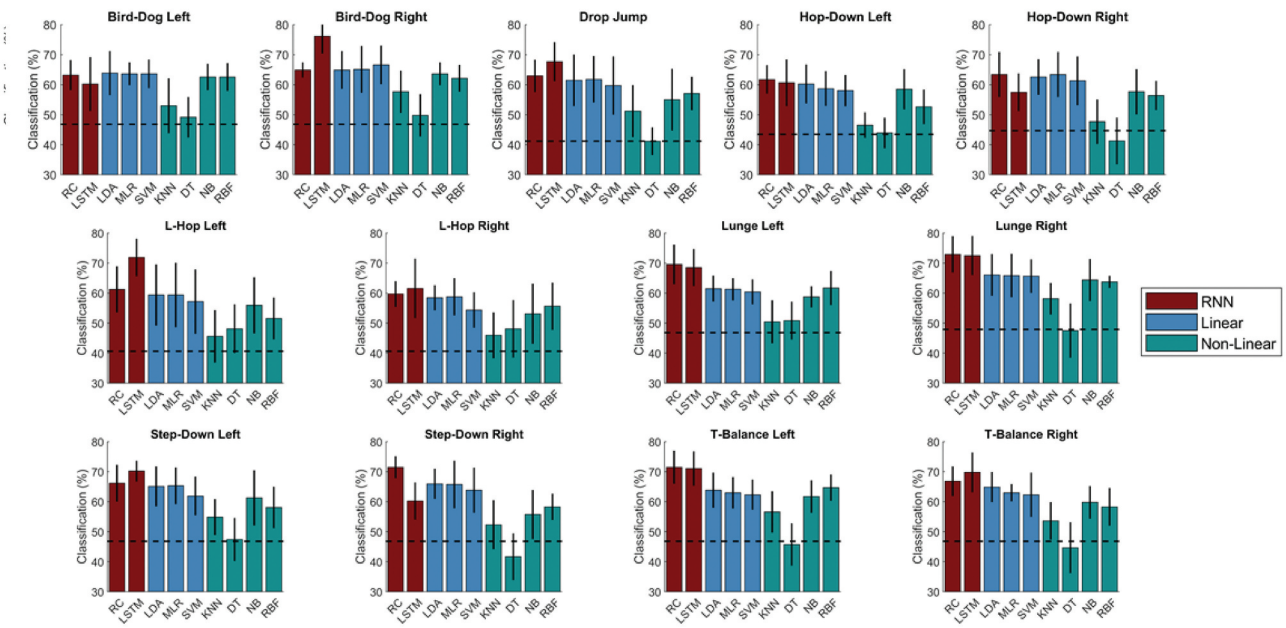

Figure 2. The mean classification rates and standard deviations when classifying for sport for all movement tasks and all machine learning classifiers. Maroon bars represent recurrent neural network classifiers, blue bars represent linear machine learning classifiers and green bars represent non-linear machine learning classifiers. The dotted black line denotes the naïve classification rates.

presented in text (see Supplemental Material, Tables S10 for the results of all other post-hoc tests). Looking at just mean accuracy rates for the lunge right, RC (72.9% ± 6.1%) and LSTM (72.5% ± 6.5%) had a significantly better classification rates than KNN (58.1% ± 5.2%; $p < 0.001$), DT (47.5% ± 9.0%; $p < 0.001$), and naïve (47.9%; $p < 0.001$); however, there were no significant differences between RC and LSTM ($p = 1.00$; Table 7.4.4). LDA (66.0% ± 6.9%), MLR (65.8% ± 7.2%), SVM (65.6% ± 5.6%), NB (64.4% ± 7.0%), and RBF (63.8% ± 2.0%) had significantly greater classification rates than DT ($p < 0.001$) and naïve ($p < 0.001$); however, there were no significant differences between LDA, MLR, SVM, NB and RBF ($p > 0.05$; Table 4). KNN ($p = 0.011$) and DT ($p = 1.00$) were not significantly different than naïve (Table 4).

Table 4. The results of the Tukey post-hoc test for the lunge for sport. The orange cells denote lunge left, whereas the blue cells denote lunge right.

	Lunge									
	RC	*LSTM*	*LDA*	*MLR*	*SVM*	*KNN*	*DT*	*NB*	*RBF*	*Naïve*
RC		1	0.273	0.235	0.201	<0.001	<0.001	0.068	0.036	<0.001
LSTM	1		0.358	0.314	0.273	<0.001	<0.001	0.1	0.056	<0.001
LDA	0.024	0.086		1	1	0.121	<0.001	1	0.998	<0.001
MLR	0.018	0.067	1		1	0.144	<0.001	1	0.999	<0.001
SVM	0.006	0.024	1	1		0.171	<0.001	1	1	<0.001
KNN	<0.001	<0.001	<0.001	<0.001	0.002		0.171	0.406	0.559	0.011
DT	<0.001	<0.001	0.001	0.001	0.003	1		<0.001	<0.001	1
NB	<0.001	0.002	0.971	0.983	0.999	0.018	0.031		1	<0.001
RBF	0.031	0.108	1	1	1	<0.001	<0.001	0.953		<0.001
Naïve	<0.001	<0.001	<0.001	<.001	<.001	0.86	0.763	<0.001	<0.001	

4. Discussion

The purpose of this study was two-fold: 1) to determine if RC and LSTM were better at predicting athlete level than the machine learning classifiers previously used, and 2) to determine if athletes could be classified based on sport. Overall, for both level and sport, the RNNs performed better than non-linear classifiers and similarly to the linear classifiers. Both the RNNs and linear classifiers performed better than the naïve classifier.

For level, RC and LSTM performed similarly for most tasks except for the bird-dog left and right and the hop-down right, where RC performed significantly better than LSTM. In addition, for the bird-dog left and right, hop-down left and right, L-hop left and right, and the T-balance left and right, LSTM had higher variability than RC. This was expected due to the nature of the classifiers. RNNs in general are known for being unstable due to the exploding/vanishing gradient effect due to back propagation through all layers (Bengio et al., 1994; Hochreiter, 1998). RC is designed to be more stable due to only manipulating the weights of the read-out layer (Lukoševicius & Jaeger, 2009). For the linear and non-linear classifiers, similar results were found to previous research (Ross et al., 2020). The linear classifiers within themselves all performed similarly. For the non-linear classifiers, KNN and DT tended to perform significantly worse than NB and RBF. When looking at the confusion matrices for RC, novice athletes were more likely to be classified as an elite than elite athletes as a novice (see Supplemental Material, Figure S1). This is most likely due to the fact that a large portion of the novice athletes were attending an elite youth sports academy, which boasts a high percentage of students continuing to compete at the collegiate and professional levels. Therefore, some of the novice athletes were on track to become elite athletes at the time of testing.

For sport, the RNNs performed similarly to each other, except for the bird-dog left where LSTM had significantly greater classification rates ($p = 0.001$). The instability of the LSTM seen with the level classifier was not as present for the sport classifier, which may be due to the smaller sample sizes and to larger differences in group sizes rather than the models being more stable. Again, similar to previous research (Ross et al., 2020) and the level classifier, the linear classifiers within themselves performed similarly and for the non-linear classifiers, KNN and DT tended to perform significantly worse than NB and RBF. When looking at the confusion matrices for RC, the percentage of athletes correctly classified for baseball and basketball were similar (See Supplemental Material, Figure S2). However, soccer and football players were more likely to be classified as basketball and baseball players. This is likely due to the uneven group sizes, with baseball and basketball having 3 and 4 times as many athletes than football and soccer.

All classifiers were trained on an Intel® Xeon® Gold 6248 CPU and 384GB of ECC RAM, except for LSTM which was trained on an NVIDIA Titan RTX GPU. For both sport and level, RC took the least amount of time. These results agree with previous research that has found that the LSTM and RC models perform similarly, but that the LSTM takes a significantly longer time to train due to backpropagating through each layer (Jirak et al., 2020). The traditional classifiers all performed similarly, with the feature selection portion requiring the most amount of time, and took longer to train than RC.

For all tasks except for the T-balance left for level, the RC and the linear classifiers had significantly better classification rates than the naïve classifier, suggesting that the classifiers were classifying based on actual differences between classes and not noise or by chance. Due to having one of the highest classification rates and taking the least amount of time to train for all tasks, going forward, it is suggested to use RC for these types of analyses. In addition, it was possible to classify athletes based on the sport, which suggests that athletes move differently based on the sport they play. Currently, popular movement screens do not consider athlete-specific demographics such as competition level, sport, age, or sex. Based on athletes moving differently depending on their level of play and sport, it argues that there should be sport- and level-specific scoring criteria for movement competency assessments, which may increase their ability to predict injury risk.

The dataset for this study contained unequal group sizes especially when classifying sport which may limit the generalisability of these results. To help combat this limitation, the ratio of athletes within groups for the entire dataset was maintained within each cross-fold. A further limitation for classifying sport was the absence of female athletes due to the selection criteria being restricted to elite athletes in baseball, basketball, football, and soccer.

Movement screens are often used for identifying athletes at a higher risk of injury or for talent identification (Cook et al., 2014; Donà et al., 2009; Kritz et al., 2009; McCall et al., 2014; McCunn et al., 2016; Padua et al., 2009). Approximately, 20–40% of athletes were misclassified based on the task being examined and/or classifier being used. Future research should perform a longitudinal study to look further into the misclassified athletes and whether novice athletes misclassified as elite athletes have a higher likelihood of making it to an elite level of play or if elite athletes misclassified as novice athletes have a higher risk of injury.

In conclusion, across all tasks, RC had one of the highest classification rates and took the least amount of time to train, therefore going forward, for these types of analyses we recommend using RC. Both the RNNs and linear classifiers performed better than the non-linear and naïve classifiers for level of play and sport played, suggesting the classifiers were classifying based on differences rather than chance. In addition, athletes were successfully classified based on sport suggesting that athletes competing in different sports move differently during non-sport-specific movements. Therefore, movement assessment screens should incorporate sport-specific scoring criteria.

Acknowledgments

The authors would like to thank the athletes who partook in the study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada (RGPIN-2020-04748, Ryan Graham; PGSD3-504132-2017, Gwyneth Ross) and the Ontario Graduate Scholarship (Gwyneth Ross).

ORCID

Ryan B. Graham http://orcid.org/0000-0001-7502-8065

References

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bianchi, F. M., Scardapane, S., Lokse, S., & Jenssen, R. (2021). Reservoir computing approaches for representation and classification of multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(5), 2169–2179. https://doi.org/10.1109/TNNLS.2020.3001377

Clouthier, A. L., Ross, G. B., Mavor, M. P., Coll, I., Boyle, A., & Graham, R. B. (2021). Development and validation of a deep learning algorithm and open-source platform for the automatic labelling of motion capture markers. *IEEE Access*, *9*, 36444–36454. https://doi.org/10.1109/ACCESS.2021.3062748

Cook, G., Burton, L., & Hoogenboom, B. J. (2014). Functional movement screening : The use of fundamental movements as an assessment of function- Part 2. *International Journal of Sports Physical Therapy*, *9*(4), 549–563. https://doi.org/10.1111/j.1600-0838.2010.01267.x

Donà, G., Preatoni, E., Cobelli, C., Rodano, R., & Harrison, A. J. (2009). Application of functional principal component analysis in race walking: An emerging methodology. *Sports Biomechanics*, *8*(4), 284–301. https://doi.org/10.1080/14763140903414425

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1),1–34. Article 12. https://doi.org/10.1145/2379776.2379788

Frost, D. M., Beach, T. A. C., Campbell, T. L., Callaghan, J. P., & Mcgill, S. M. (2015). An appraisal of the Functional Movement Screen™ grading criteria - Is the composite score sensitive to risky movement behavior? *Physical Therapy in Sport*, *16*(4), 324–330. https://doi.org/10.1016/j.ptsp.2015.02.001

Gulgin, H., & Hoogenboom, B. (2014). The Functional Movement Screening (FMS)™: An inter-rater reliability study Between raters of varied experience. *International Journal of Sports Physical Therapy*, *9*(1), 14–20. https://pubmed.ncbi.nlm.nih.gov/24567851/

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(2), 107–116. https://doi.org/10.1142/S0218488598000094

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jirak, D., Tietz, S., Ali, H., & Wermter, S. (2020). Echo state networks and long short-term memory for continuous gesture recognition: A comparative study. *Cognitive Computation*, 1–13. https://doi.org/10.1007/s12559-020-09754-0

Kobsar, D., Osis, S. T., Hettinga, B. A., & Ferber, R. (2014). Classification accuracy of a single tri-axial accelerometer for training background and experience level in runners. *Journal of Biomechanics*, *47*(10), 2508–2511. https://doi.org/10.1016/j.jbiomech.2014.04.017

Kritz, M., Cronin, J., & Hume, P. (2009). The bodyweight squat: A movement screen for the squat pattern. *Strength and Conditioning Journal*, *31*(1), 76–85. https://doi.org/10.1519/SSC.0b013e318195eb2f

Lukoševicius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149. https://doi.org/10.1016/j.cosrev.2009.03.005

McCall, A., Carling, C., Nedelec, M., Davison, M., Le Gall, F., Berthoin, S., & Dupont, G. (2014). Risk factors, testing and preventative strategies for non-contact injuries in professional football: Current perceptions and practices of 44 teams from various premier leagues. *British Journal of Sports Medicine*, *48*(18), 1352–1357. https://doi.org/10.1136/bjsports-2014-093439

McCunn, R., Aus der Fünten, K., Fullagar, H. H. K., McKeown, I., & Meyer, T. (2016). Reliability and association with injury of movement screens: A critical review. *Sports Medicine*, *46*(6), 763–781. https://doi.org/10.1007/s40279-015-0453-1

McPherson, A. L., Dowling, B., Tubbs, T. G., & Paci, J. M. (2016). Sagittal plane kinematic differences between dominant and non-dominant legs in unilateral and bilateral jump landings. *Physical Therapy in Sport*, *22*, 54–60. https://doi.org/10.1016/j.ptsp.2016.04.001

Onate, J. A., Dewey, T., Kollock, R. O., Thomas, K. S., Van Lunen, B. L., DeMaio, M., & Ringleb, S. I. (2012). Real-time intersession and interrater reliability of the Functional Movement Screen. *Journal of Strength and Conditioning Research*, *26*(2), 408–415. https://doi.org/10.1519/JSC.0b013e318220e6fa

Padua, D. A., Marshall, S. W., Boling, M. C., Thigpen, C. A., Garrett, W. E., & Beutler, A. I. (2009). The Landing Error Scoring System (LESS) is a valid and reliable clinical assessment tool of jump-landing biomechanics: The JUMP-ACL study. *American Journal of Sports Medicine*, *37*(10), 1996–2002. https://doi.org/10.1177/0363546509343200

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Ross, G. B., Dowling, B., Troje, N. F., Fischer, S. L., & Graham, R. B. (2018). Objectively differentiating movement patterns between elite and novice athletes. *Medicine & Science in Sports & Exercise*, *50*(7), 1457–1464. https://doi.org/10.1249/MSS.0000000000001571

Ross, G. B., Dowling, B., Troje, N. F., Fischer, S. L., Graham, R. B., & Graham, R. B. (2020). Classifying elite From novice athletes using simulated wearable sensor data. *Frontiers in Bioengineering and Biotechnology*, *8*, 1–10. https://doi.org/10.3389/fbioe.2020.00814

Smith, C. A., Chimera, N. J., Wright, N. J., & Warren, M. (2013). Interrater and intrarater reliability of the Functional Movement Screen. *Journal of Strength and Conditioning Research*, *27*(4), 982–987. https://doi.org/10.1519/JSC.0b013e3182606df2

Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Visualized Experiments*, *2*(5), 371–387. https://doi.org/10.1167/2.5.2

Weng, X., & Shen, J. (2008). Knowledge-based systems classification of multivariate time series using two-dimensional singular value decomposition. *Knowledge-based Systems*, *21*(7), 535–539. https://doi.org/10.1016/j.knosys.2008.03.014

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*(4), 597–604. https://doi.org/10.1142/S0219622006002258