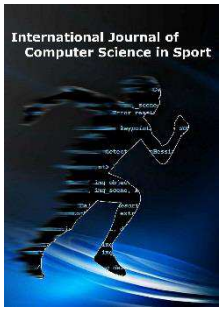
International Journal of Computer Science in Sport

Volume 24, Issue 1, 2025

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2025-0007



Can machine learning distinguish between elite and non-elite rowers?

Kristine Fjellkårstad Orten[1], Sander Elias Magnussen Helgesen[1], Bihui Chen[1], Adel Baselizadeh[1,2], Jim Torresen[1,2], Henrik Herrebrøden[3]

[1] Department of Informatics, University of Oslo, Oslo, Norway

[2] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Oslo, Norway

[3] Department of Psychology, Pedagogy and Law, School of Health Sciences, Kristiania University College, Oslo, Norway

Abstract

A major challenge for sports coaches and analysts is to identify critical elements of athletes' movement patterns. A potentially relevant tool is machine learning, useful because of its ability to extract patterns from data. In the current study, we employed various deep learning frameworks, including Gated Recurrent Unit networks (GRUs), Convolutional Neural Networks (CNNs), and Multi-Layer Perceptrons (MLPs), to search for differences between elite and non-elite rowers using a rowing ergometer. The MLP model achieved an accuracy of 100% when using all input features, indicating that the problem is suitable as a machine learning task. Our research focused on using a limited amount of the data. Despite using fewer input features, the models managed to classify skill levels with reasonable precision, reaching a best performance of 77% accuracy for the model combining GRU and CNN architectures, 78% for the GRU model, and 94% for the MLP model. From a rowing perspective, the results suggest that movement coordination between upper and lower body limbs, as represented by different feature combinations, is informative in distinguishing between elites and non-elites. The current work suggests that machine learning may supplement human experts in sports coaching, analytics, and talent identification.

KEYWORDS: SPORTS ANALYSIS, ROWING, MACHINE LEARNING, NEURAL NETWORKS, DEEP LEARNING, GATED RECURRENT UNIT, CONVOLUTIONAL NEURAL NETWORK, MULIT-LAYER PERCEPTRON .

Introduction

A main target in sports, at all levels, is to improve and continuously develop one's skills. To do so, athletes need feedback regarding current movement patterns and, preferably, information regarding alternative movement solutions such as those employed by skilled performers at higher levels. Traditionally, athletes have depended heavily on sports coaches in providing this information. Indeed, experienced coaches can have the ability to visually detect critical motion features and predict relevant outcomes of sporting actions (e.g., Wood et al., 2023). However, not all athletes have consistent access to coaching. Further, coaches' visual efficacy may depend on their experience, their viewing angle, and the age of the athletes they observe (Wood et al., 2023). Fast sporting actions may be particularly challenging to track visually (Knudson, 1999; Wood et al., 2023). This is where various technologies can assist coaches, and machine learning (ML) seems to hold promise as a supplement to sports analytics and coaching. ML is a type of artificial intelligence where the goal is to make machines learn in a similar way to humans. That is, ML algorithms can be fed with data for training and, subsequently, use this "experience" to make decisions or predictions.

ML can serve a range of purposes in sports. This includes recognizing movement patterns as well as predicting important outcomes related to success and injuries (for reviews, see Bunker & Susnjak, 2022; Cust et al., 2019; Horvat & Job, 2020; Rico-González et al., 2023; Van Eetvelde et al., 2021). For example, Rindal et al. (2017) gathered motion data from the chests and arms of skiers. They subsequently used a neural network algorithm for training, validation, and testing, in order to classify sub-techniques in classical cross-country skiing. Overall, the test data reached 93.9% accuracy, although some techniques were misclassified more frequently than others. In a rowing study, Chen et al. (2023) used a graph-matching network model to investigate postures in rowers using an ergometer. Specifically, videos of high school rowers using the ergometer were used to extract features related to postures (represented by graph structures), and then posture differences between rowers were computed. The results suggested that the rowers varied in terms of posture, but some had similar postures which, assumedly, may indicate that they are more suited to pair up for competitions. Overall, ML seems suited to classify complex movement and extract important information regarding athletes' differences and similarities.

In general, it is easier to classify a given type of movement or action, than it is to determine the quality of actions in sports (Wood et al. 2023). One approach to infer critical components in effective sporting actions, however, is to compare athletes of different performance standards. Bosch et al. (2015) used the K-nearest neighbor algorithm to explore the difference between four novices and three experienced rowers. Based on inertial sensor modules placed on the rowers' bodies, they found that experienced rowers demonstrated greater consistency in postural angles as well as more consistent stroke timing. This was, however, based on a very small sample and the experienced rowers had merely a minimum of three years of rowing experience. Ross et al. (2018) conducted a study using a larger sample, wherein 542 athletes were categorized as either novice or elite. The elite category consisted of high-level athletes, from the inter-collegiate to the professional levels. All athletes completed a battery of general physical tests (e.g., lunges, drop jumps), while a motion capture system gathered 3D movement data. Subsequently, various ML algorithms were used to classify elites and novices, respectively, based on the tests. The different tests' classification accuracies ranged from 75.1% to 84.7%. This suggests that ML could distinguish between elites and novices at a level far above chance, albeit based on tests that are non-sport specific. Similar work during sport-specific action should help illuminate crucial aspects of athletes' techniques.

Identifying important characteristics in high-level athletes, in their given performance environment, would be in line with the expert performance approach (Williams & Ericsson,

2005). This involves identifying the essence of skilled performance (i.e., what aspects give experts their advantages?) and the underpinning mechanisms. Such information can guide practice and be fruitful for performers at all levels. However, identifying the key ingredients of elite performance can be a severe challenge (e.g., Lorenz et al., 2013). With respect to technical execution, athletes typically demonstrate notable movement variability (Bartlett et al., 2007). This includes inter-individual variability, as different athletes have different body proportions and technical preferences. It also includes intra-individual variability, as individual athletes can vary their movement patterns (and still achieve similar outcomes). Movement variability suggests that there is no universal template for athletic actions, and this speaks to the challenge of trying to isolate the most crucial information in sports. This is where motion capture technology and ML can assist sports coaches, analysts, and the athletes themselves.
The current study aimed to apply ML to motion capture data to classify athletes based on complex and sport-specific actions. Data features were based on rowing ergometer trials involving elite and non-elite rowers. As ergometer rowing involves complex movements (i.e., involving multiple joints and degrees of freedom; Cordo & Gurfinkel, 2004), our goal was to investigate which parts of the rowers' bodies and the rowing ergometer components that are most informative when distinguishing between elite and non-elite rowing trials. This ML approach research occurred in parallel with more targeted and conventional analyses using the same motion capture data set (see Herrebrøden et al., 2023).

Methods

Participants

Current data came from recordings of 18 male rowers – nine elite and nine non-elite.

All the elite rowers were part of Team Norway, the highest-level training group in the Norwegian national rowing team. When the study was conducted, most of the elites were qualified for the 2020 Olympic games in Tokyo, and the rest were high-level rowers that practiced with Team Norway. Their mean rowing experience was 14.89 (SD = 5.84) years, and their mean age was 29.67 (SD = 6.06) years. A head coach of Team Norway estimated that the elite rowers on average spent approximately 17 hours a week on physical training in 2020. All elites had competed in the World Cup, along with the rest of the world's elite rowers.

The non-elite group consisted of recreational rowers. Their mean rowing experience was 4.33 (SD = 4.95) years, and their mean age was 29.89 (SD = 11.70) years. All the non-elite participants estimated their weekly physical training average over the last year to be a minimum of three sessions or a minimum of seven hours. The criteria for including the non-elite rowers were that they had never won an individual medal in any major national or international competition and that they had never represented Norway's official national team as a rower.

The participants were offered no reward or reimbursement. All the rowers voluntarily agreed to participate at a time that suited their training schedule.

Measurements

The data used in the current work were based on an experiment with rowing trials in six different conditions. For our current purposes, data from only one condition was included. In this particular condition, participants used a dynamic rowing ergometer (Row Perfect 3 Model S; www.rp3rowing.com) for three minutes, with approximately 85% of their 2-kilometer race speed. Hence, each rower adhered to a specific speed level tailored to their individual capacity, resembling a moderate physical intensity level. For further experiment details, see Herrebrøden et al. (2023).

The data were collected using a Qualisys (www.qualisys.com) motion capture system with 11 infrared cameras recording at a frequency of 240 Hz (images per second). A total of 12 reflective markers were placed on locations of interest, that is, on different parts of the rower's body and on the ergometer (see Figure 1). Four markers were attached to the rowers themselves, on each of their shoulders and hips. On the ergometer, two markers were attached to the seat (on both sides), three markers were placed on the handle, and three markers were attached to the ergometer front. It is worth noting that these ergometer parts move as a function of the rower's movements. For instance, for this particular ergometer, the seat marker moved vertically based on the rower's balance (or instability), while the ergometer front moved as a function of how extended the rower's legs were. Hence, all markers could provide spatial information related to rowing technique.

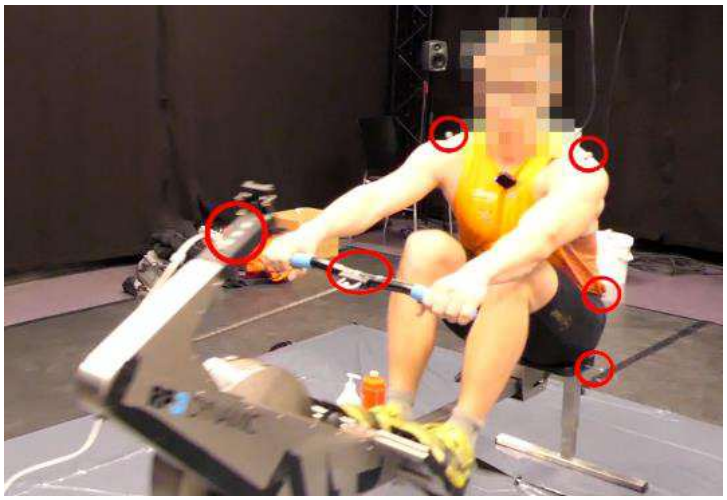


Figure 1: A rower using the rowing ergometer with reflective markers (marked with red).

Data preprocessing

The data set consisted of 3D coordinates indicating the location of the 12 reflective markers during trials. We removed the first and last 10 seconds of each trial.

The motion data were extracted via Qualisys Track Manager (version 2021.2) where the reflective markers were labelled and manually processed. One of the current authors identified and labeled the markers that the motion capture system failed to detect during parts of the recordings. This process continued until at least one marker at each location of interest had less than 10% missing data. Specifically, either the left or right markers on the rower's shoulders and hips, the ergometer's seat and handle, and at least one marker on the front of the ergometer had to be captured during at least 90% of the samples used for analysis.

The data were further normalized to remove differences that occur in the data because the rowers have different body proportions (e.g., different heights). Normalization is often used in ML to enhance the performance and training stability of the model by transforming features to a similar scale. We used min-max normalization to scale the data between 0 and 1.

Originally, the data consisted of one time series of coordinates for each rower. Because of the limited number of rowers in our dataset, we chose to segment the data into shorter, overlapping sequences (visualized in Figure 2). During this process, we decided to (1) include more than a complete stroke per sequence and (2) segment with overlapping windows. Including more than a complete stroke both mitigates the need for highly precise stroke detection and preserves potentially important transition patterns between strokes, while segmenting with overlapping windows maximize the amount of training data. Increasing the number of available samples

helps reduce overfitting and improves model generalization.

The length of a rowing sequence was determined by the rower's median rowing stroke duration, calculated by detecting the motion turning point of each rowing stroke for each rower. We detected this point for each stroke by calculating the motion direction of the handle every 0.25 seconds of trial data. The calculation was done with an error margin of 1 mm in the movement direction. We wanted to evaluate movement direction often enough to not miss the turning point, but we were not overly concerned with the exact accuracy as the goal of the sequence splitting was to create more data points, not calculate exact rowing stroke duration. To avoid too much difference between rowing segments, the max length was set to 2.75 seconds. We added a margin of 0.25 seconds to each rowing segment to ensure that each sequence contained at least one complete stroke. A sequence overlap of 1 second was introduced to generate a larger number of sequences for analysis and to relate the sequences to each other, which is beneficial for ML analysis. In total, the data was split into 1741 strokes.

The preprocessed data was randomly divided into data sets based on rowers' skill level, ensuring a balanced representation of both elite and non-elite participants in each set. However, due to variations in the length of rowing sequences among rowers, the number of strokes for each skill level was not necessarily balanced within data sets.

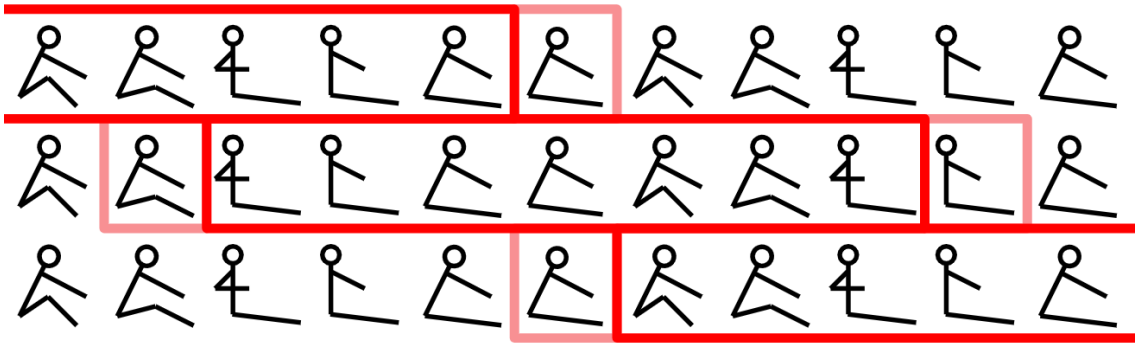


Figure 2: Sequence splitting. Example visualization of rowing sequence input data. The sequence of rowing strokes was divided into multiple parts of single strokes (dark red box) with a 0.25-sec error margin (light red) to ensure a complete stroke was included. Each new sequence had a 1-second overlap with the previous sequence.

ML models

We implemented three ML models using Python and TensorFlow Keras. Our first model was a Gated Recurrent Unit (GRU) Network (Cho et al., 2014), which is a type of Recurrent Neural Network (RNN) that allows adequate training for small data sets (Chung et al., 2014). RNNs were chosen because of the sequential properties of the rowing data. Our model used two GRU layers with 64 units each, a dropout rate of 0.2 and a recurrent dropout rate of 0.2.

Our second model was a GRU Convolutional Neural Network (GRU-CNN) model. We wanted to explore whether incorporating other deep learning architectures could enhance our results. This model extended the initial GRU model with a CNN and additional layers (including max pooling and two dense layers). This model is detailed in Figure 3.

Lastly, we implemented a Multi-Layer Perceptron (MLP) model with two layers, the first containing 128 neurons and the second 64 neurons. The dropout rate was set to 0.2. This model was employed to see how a simpler model would perform on the data. Unlike GRU or GRU-CNN models, which capture temporal dependencies and spatial features, MLP models processes inputs independently, making it suitable as a baseline for comparison against more complex architectures.

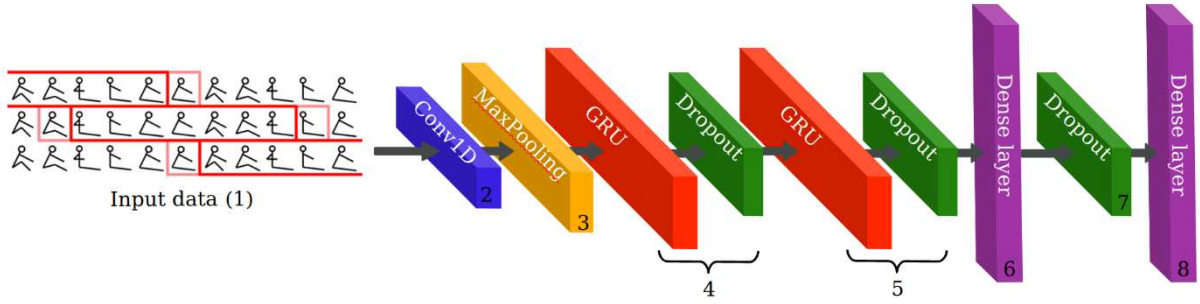


Figure 3: The GRU-CNN model extended the initial GRU model with a CNN and more layers. The input (1) was sent through a convolutional layer (2) with 128 filters, a kernel size of 3 and ReLU activation, then a max pooling layer (3) with pool size 2, then two GRU layers (4, 5) with the same parameters as the original GRU model. At last, the data is sent through a dense layer (6) with 128 units and ReLU activation, then a dropout layer (7) with a dropout rate of 0.3 and then a final one-unit dense layer (8) with sigmoid activation.

Input features

We constructed two different types of input features used for the models: input represented as coordinates and input represented as angles. We tested multiple types of features to see how different input complexity would impact model performance.

For the first input feature type, we used time series of 3D coordinates to represent the position of the reflective markers during rowing. We evaluated the performance of the model for the following combinations of features:

- ShoHip: Shoulders (left, right) and hips (left, right)
- ShoSea: Shoulders (left, right) and seat (left, right)
- ShoFro: Shoulders (left, right) and ergometer front
- HanFro: Ergometer handle (left, right) and ergometer front
- HipFro: Hips (left, right) and ergometer front

In addition, we introduced a sixth coordinate input type, using all reflective marker positions (AllJoints). This was only employed for the MLP model because it provided a useful baseline for comparisons while remaining computationally efficient. For the more complex GRU and GRU-CNN models, using all joints significantly increased computational costs and did not align with our goal of identifying the most informative input feature combinations rather than simply using all available data to maximize accuracy.

For the second input feature type, we used time series of angles between the different 3D points (shown in Figure 4). The angles were calculated directly from the stroke time series data. We evaluated the performance of the following combinations of features:

- Shoulder-angle α: The angle between the line from the right hip to the right shoulder and the line from the right shoulder to the right ergometer handle marker.
- Hip-angle ß: The angle between the line from the right shoulder to the right hip and the line from the right hip to the right seat.
- Seat-angle γ: The angle between the line from the left seat to the right seat and the line from the right seat to the right hip.

These particular coordinate combinations and angles have potential relevance for rowing technique (see, for example, Smith & Loschner, 2002; Soper & Hume, 2004). Based on results

from our initial model, we decided to test the GRU-CNN and the MLP model solely with coordinates as input, not angles. This choice allowed us to focus on the most effective features while still showing the results of the angle-based input, as it might offer useful insight for future research in the field.

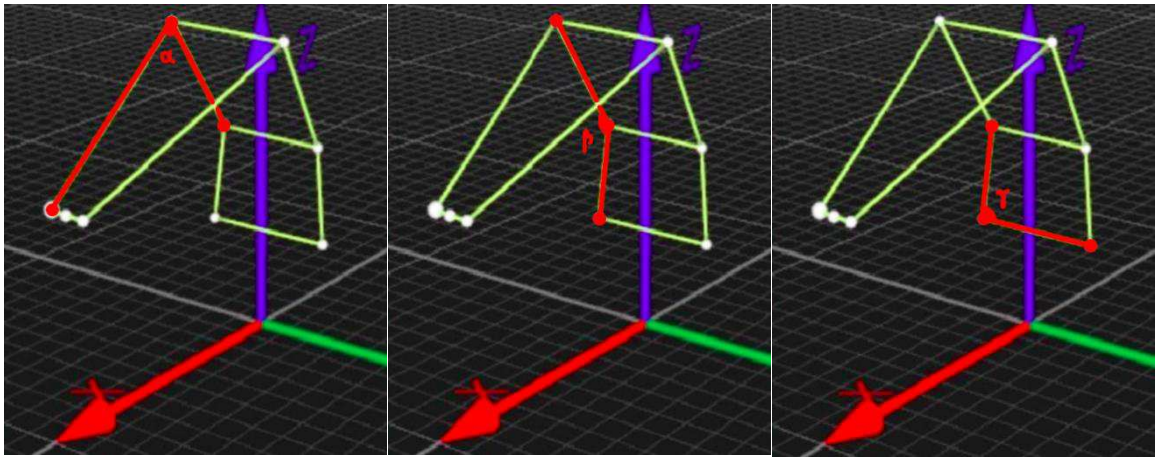


Figure 4: Rowing angles. Example visualization of the second type of input features: Angles between reflective markers. From the left: α, ß and γ.

Target variables

Each model was trained with one target variable: the skill level of the rower, binary encoded as "elite" or "non-elite".

Training and evaluation

For the GRU model, the data set was divided into training, validation, and test sets with a 60-20-20 split. This resulted in 10 rowers (five elite and five non-elite) for the training set and four rowers (two elite and two non-elite) for each of the validation and test sets. To evaluate different combinations of input features, we created eight instances of the model, with each instance corresponding to a unique combination of input features. Each instance of the model was trained 10 times, with each training run lasting for a maximum of 15 epochs. During each epoch, the model's performance was evaluated on the validation set, and the version of the model that achieved the highest accuracy on the validation set was saved. We used the Adam optimizer and a batch size of 4. For the loss function, we used binary cross-entropy loss. Due to time limitations, hyperparameter tuning for the first model was performed based on only two epochs, not the full 15 epochs used for training. The focus of the tuning was on the model employing coordinate features. The same hyperparameter values were used for all input types. For the eight chosen coordinate- and angle-feature combinations, we got a total of 80 test set accuracy values. These scores ranged from 0 to 1. For binary problems (such as categorizing elite and non-elite athletes) a score of .5 indicates an accuracy level that is equivalent to guessing (50%), while a score of 1 indicates 100% accuracy.

Initially, our study only included the GRU model. Based on the results from this model, we made adjustments to the development process for the subsequent GRU-CNN and MLP models. These adjustments included changing the evaluation method to 10-fold cross validation to get more robust results, only using coordinate feature input as the coordinate features showed the most promise in the GRU model, and calculating two additional performance metrics: F1 score and AUC (Area Under the Curve). These new performance metrics enabled us to provide a more comprehensive evaluation of the model. Beside these differences, the training setup for the

GRU-CNN model was kept as similar as possible to the GRU model. For instance, we used 15 epochs for each fold of the cross-validation. We ran exploratory tests for the GRU and GRU-CNN models with up to 50 epochs, but as this did not cause notable changes to the accuracy scores we decided to stick with 15 epochs per fold. We trained the MLP model similarly to the previous models, using five different input feature combinations, and additionally evaluated the model with all reflective marker data (12 markers) as input. For training, we used 10-fold cross-validation, with each fold consisting of 100 epochs. Other training parameters were consistent with those used in the earlier models.

The GRU and GRU-CNN models were trained on coordinate time-series data. The coordinate input features used for training had the shape (batch_size, 720, 12), with 720 being the number of time steps and 12 the result of using four reflective markers in 3D (4x3). The MLP model was tested with both normal coordinate input and average coordinate input. As these methods both produced quite similar results, we decided on average input as that is less computationally intensive. The average input was created by calculating the average of the 720 time steps for each feature. Consequently, the input data for the MLP model had a shape of (batch_size, 1, 12). For the GRU model using angle input, the shape was (batch_size, 720, 1), where the one denotes a single angle calculated based on three provided marker positions.

Results

GRU model

Figure 5 provides an overview of our results from the validation and testing phases. With respect to testing data with coordinate features, the greatest mean accuracy was found for HipFro (M = .7807, SD = .1226), followed by ShoHip (M = .7631, SD = .1636), ShoSea (M = .6010, SD = .1502), ShoFro (M = .5822, SD = .1440), and HanFro (M = .4396, SD = .0988).

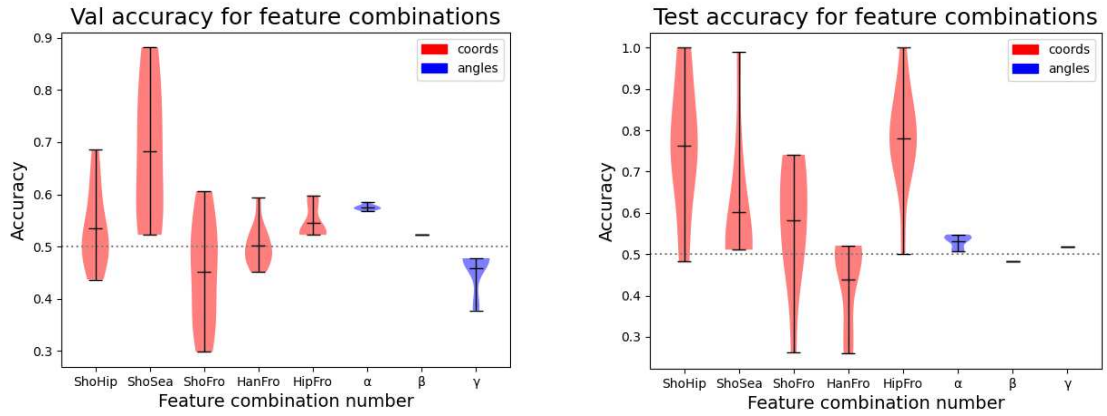


Figure 5: Violin plot showing validation and test accuracies for different feature combinations in the GRU model. The mean value and range are marked in the plot. The mean was computed across 10 instances of training the model. The left-most plot shows the accuracies achieved for different input features when using the validation data set, while the right-most plot shows the accuracies for the test data set.

The angle feature testing data suggested that the most accurate angle feature was the shoulder angle (α; M = .5315, SD = .0151), followed by the seat angle (γ; M = .5183, SD = .0000) and the hip angle (β; M = .4817, SD = .0000).

GRU-CNN model

Figure 6 provides an overview of our results from the validation phase of the GRU-CNN model with coordinate features. Table 1 provides testing phase metrics for the model.

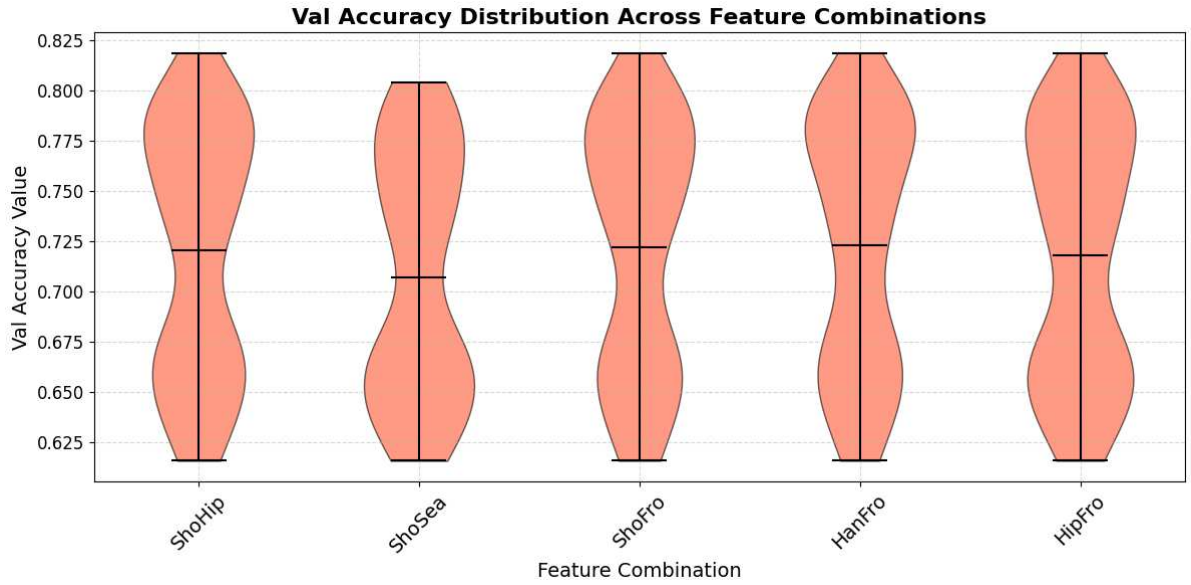


Figure 6: 10-fold cross-validation results for the GRU-CNN model. The mean value and range are marked in the plot.

Table 1: Performance metrics for the best-trained GRU-CNN model when evaluated on the test dataset.

	Input features	**F1 score**	**AUC**	**Accuracy**
Coords	Shoulders and hips	0.4226	0.4026	0.5181
	Shoulders and seat	0.4150	0.4460	0.5181
	Shoulders and ergometer front	0.6989	0.5376	0.7744
	Ergometer handle and front	0.6792	0.5368	0.7632
	Hips and ergometer front	0.5222	0.5117	0.5209

MLP model

Training and validation accuracy for the MLP model is shown in Figure 7. Table 2 provides testing phase metrics for the model.

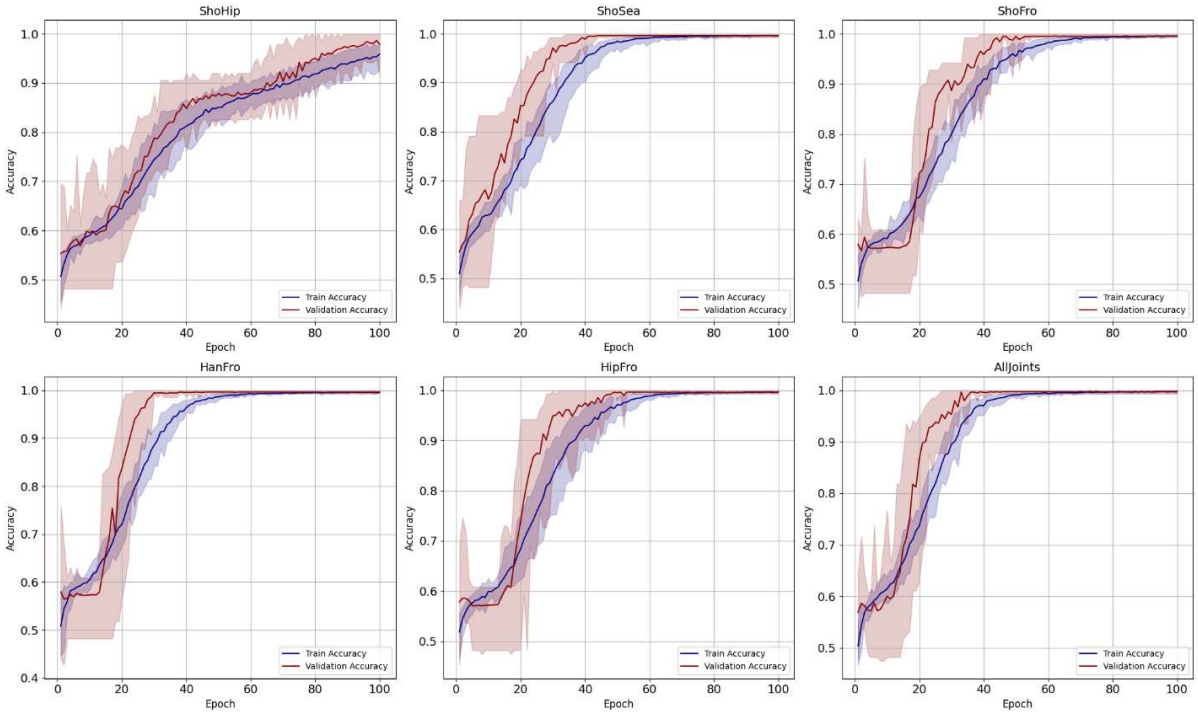


Figure 7: Training and validation accuracies for the MLP model for coordinate input feature combinations.

Table 2: Performance metrics for the best-trained MLP model when evaluated on the test dataset.

	Input features	**F1 score**	**AUC**	**Accuracy**
Coords	Shoulders and hips	0.2836	0.9962	0.6178
	Shoulders and seat	0.3557	0.9949	0.6591
	Shoulders and ergometer front	0.9328	0.9996	0.9482
	Ergometer handle and front	0.9190	0.9999	0.9387
	Hips and ergometer front	0.6812	0.9948	0.8365
	All joints	1	1	1

Discussion

In the present study, we used a GRU model, a GRU-CNN model, and an MLP model with various movement-related features based on ergometer trials to see if elite and non-elite rowers could be distinguished. Our GRU and GRU-CNN results suggest that one can achieve classification accuracies in the 76-78% range. This is far above chance level (i.e., 50%) and these findings echo Ross et al. (2018) who used general physical tests and ML to classify elites and non-elites with approximately 75% to 85% accuracy. When using an MLP model, with its simpler architecture, we achieved accuracy scores in the 94-95% range. These accuracy scores match those obtained by Rindal et al. (2017), achieving approximately 94% accuracy in classifying different sub-techniques in cross-country skiing. We obtained these results by using feature combinations based on markers placed on two parts on the rower's body and/or the ergometer, suggesting that limited motion data can provide sufficient data to distinguish between skill levels in rowing. When adding all features in one MLP model, we classified skill levels with 100% accuracy. Overall, our study confirms the potential of using ML and motion capture data to distinguish between performers at different levels in sports.

Regarding different input feature types, the performance of the GRU model with angle input features was worse than the performance with coordinate features. This is reasonable because

the number of features is reduced during transformation from coordinates to angles. The coordinate features used four input data points in 3D where two and two of them were pairs of the same reflective marker (left and right). In comparison, the angle features used three different data points in combination, but only in two spatial dimensions. Because of this, it can be expected that angle representations achieve lower accuracies than the coordinate features. Based on the differences observed between using coordinates versus angles in the GRU model, we decided to test the additional models, GRU-CNN and MLP, solely with coordinates as input, and did not train or test any more models based on angles.

Our findings were heterogenous in determining which two features, in combination, were more effective in order to classify skill levels. Various combinations achieved different test accuracy scores in different models. Our GRU results suggested that shoulders and hips as well as shoulders and seat, were the best combinations in distinguishing between elites and non-elites. These two combinations were, however, the worst-performing when using GRU-CNN and MLP. In the latter models, the combinations with shoulders and ergometer front as well as ergometer handle and front were the most accurate. These diverse findings could be due to several factors. First, ML training based on a relatively small dataset makes variable accuracy scores likely. Second, all feature combinations provided information regarding movement coordination between the rower's upper and lower body, which is a key aspect of rowing success (see Herrebrøden et al., 2023; Soper & Hume, 2004). Overall, it seems that various feature combinations can be used to distinguish elites and non-elites in rowing.

While the GRU-CNN model did not outperform the GRU model in terms of accuracy, it appeared more robust. Employing k-fold cross-validation for the GRU-CNN model means the model results are more reliable than those from the GRU-model. In addition, the model shows similar performances across folds, indicating better generalization. The more reliable and stable results of the GRU-CNN model come at the cost of increased computational demand. As the more complex GRU-CNN model did not enhance the results from the GRU model, we also explored a simpler ML architecture to see how it would perform. The MLP model architecture was simpler both in terms of structure and computational cost. From the MLP results, it is evident that the simpler model outperforms the previous GRU and GRU-CNN models. The MLP model likely outperforms the GRU and GRU-CNN models due to its simpler architecture. One potential reason for this is overfitting in the more complex models. Overfitting occurs when a model learns noise or irrelevant patterns from the training data, reducing its ability to generalize to new data. GRU and CNN models, with their larger number of parameters, are prone to overfitting, particularly when the dataset is small or the problem does not require as much complexity. An MLP, being simpler, may generalize better. From the results, we observed that using all reflective marker data significantly improved the performance of the MLP model. This indicates that more comprehensive input data allows the model to capture more features and nuances, leading to higher accuracy.

Limitations

A main limitation was our limited sample size and a relatively small data set. This is quite common in research on experts or elite athletes at the highest performance levels, given their distinguished status. A small data set can affect model generalizability and the likelihood of variability in model performance. The variable results shown by the GRU model was mitigated by using k-fold cross-validation, but we still see that the more complex models suffer compared to the simpler MLP model, potentially due to overfitting because of the small dataset. This is observed despite splitting the data with overlap to generate more samples.

In addition, the rowing sequences might have been too long for the GRU and GRU-CNN models.

For the MLP model, we calculated the average of the data sequence in each direction for each reflective marker and fed this simplified data into the model. This simplification did not negatively impact the results. We also tested using the raw data sequence without averaging, and the results remained unchanged. However, for the GRU and GRU-CNN models, using long sequences of 720 data points as input may have hindered proper training. Shortening the sequence length could potentially lead to better results, but optimizing this hyperparameter requires extensive testing.

Hyperparameter tuning is another important limitation of the current research. GRU and CNN models have more hyperparameters that need to be properly tuned, which can be time-consuming. While we trained the GRU and GRU-CNN models for 15 epochs, it is possible that increasing the number of epochs might lead to better results. However, after exploratory testing with up to 50 epochs, we did not observe convergence to higher accuracies, suggesting that 15 epochs might be close to the optimal training time for these models. It is possible that with even more epochs (e.g., 200), these models could reach higher accuracies, but due to the time-consuming nature of training, we chose to limit the epochs. In addition, our research aimed to show which features are important for distinguishing between elite and non-elite rowers, rather than train an optimal classifier. Due to this and time constraints, we chose to limit our hyperparameter tuning. We ran the model for only a few epochs while tuning the hyperparameters, instead of the full 15 epochs used when training the model.

Another limitation of this study is the absence of time normalization for the rowing strokes. Time normalization could have provided significant benefits by aligning the duration of each stroke to a common timescale, facilitating direct comparisons across different participants, regardless of their individual stroke rates. This approach would have allowed for a more precise analysis of the biomechanical patterns within each stroke phase, enabling the identification of subtle differences between elite and non-elite rowers. While median-based segmentation helped mitigate outlier effects, time normalization could have enhanced the robustness and interpretability of our findings.

Finally, it is worth noting the difference between ergometer rowing and rowing on water. Ergometer rowing is biomechanically different from on-water rowing (see Kleshnev, 2005). For example, the ergometer handle invites a limited number of pulling trajectories in ergometer rowing. When rowing on water, on the other hand, careful maneuvering of oar handles is a demanding skill and a key to success. In general, the difference between experts and non-experts in a given domain can become clearer with increasing task difficulty (e.g., Arsal et al., 2016). Thus, future studies may want to explore the use of ML based on motion capture from trials on water to help discriminate between elite and non-elite rowers.

Practical implications

ML appears to be of value during talent identification and analysis in sports. As illustrated by the current findings, relatively limited motion capture data features and ML training can be used to indicate different skill levels among athletes. Further, despite our current mixed findings, such an approach can conceivably help analysts and coaches in distinguishing between crucial and less important movement features. One approach would be to use ML followed by other analytical approaches. The analyst, after gathering motion data, could initiate the analytical process by running ML models to identify which joints or motion features that seem to be relevant, for example by using a similar approach to what we did in the current study. Next, relevant features could be analyzed further via more targeted investigations and statistical tests, similar to what was done with the current motion capture data by Herrebrøden et al. (2023). During such a process, the coach can also make analytical decisions based on experience or

domain-specific knowledge. An interplay between ML and human judgement might be more thorough and effective than relying on human judgement alone.

ML should also have the potential to assist in training. As mentioned in the Introduction, the expert performance approach can be useful as it can enable athletes to learn from skilled performers and adopt their strategies (Williams & Ericsson, 2005). Observational learning (i.e., watching a performance) can be a useful training supplement in this venture (Ste-Marie et al., 2020). However, applying observational learning effectively may not always be straightforward, for a number of reasons. Not all motion features are important, and thus not all aspects of an expert performance are worth mimicking. Further, athletes are different, and a novice rower may, for example, be ill-advised to adopt movement patterns from an elite rower with incompatible body proportions and technical preferences. ML could help clarify the compatibility between athletes' technical styles. For instance, Chen et al. (2023) used ML to identify which rowers had similar postures during ergometer trials. Such information could be useful for athletes about to form a team or novice athletes about to observe elites.

Conclusion

ML using sport-specific motion capture data can be utilized to distinguish between elite and non-elite athletes. Both our GRU and GRU-CNN models, and the simpler MLP model, achieved encouraging testing accuracies, with the MLP model using all input features achieving an accuracy of 100%. Various feature combinations, with markers placed on different parts of rowers' bodies and the ergometer, were able to allow different ML models to classify skill levels substantially above chance level. Despite the current study's limitations, our research joins a line of research to suggest that sports coaches and analysts should consider adopting ML in order to optimize information provided to athletes.

Acknowledgements

We are grateful to Frank Veenstra, Benedikte Wallace, Farzan Majeed Noori and Zia Uddin for helpful feedback during the project. This work is partially supported by The Research Council of Norway (RCN) as a part of the projects: Collaboration on Intelligent Machines (COINMAC) project, under grant agreement no. 309869, Predictive and Intuitive Robot Companion (PIRC) under grant agreement no. 312333 and through its Centres of Excellence scheme, RITMO with project no. 262762.

References

Arsal, G., Eccles, D. W., & Ericsson, K. A. (2016). Cognitive mediation of putting: Use of a think-aloud measure and implications for studies of golf-putting in the laboratory. *Psychology of Sport and Exercise, 27,* 18–27. https://doi.org/10.1016/j.psychsport.2016.07.008

Bartlett, R., Wheat, J., & Robins, M. (2007). Is movement variability important for sports biomechanists?. *Sports Biomechanics, 6*(2), 224-243. https://doi.org/10.1080/14763140701322994

Bosch, S., Shoaib, M., Geerlings, S., Buit, L., Meratnia, N., & Havinga, P. (2015). Analysis of indoor rowing motion using wearable inertial sensors. *Proceedings of the 10th EAI International Conference on Body Area Networks*, 233-239. http://dx.doi.org/10.4108/eai.28-9-2015.2261465

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. Journal of Artificial Intelligence Research, 73, 1285-1322. https://doi.org/10.1613/jair.1.13509

Chen, C. C., Lin, C. S., Chen, Y. T., Chen, W. H., Chen, C. H., & Chen, I. C. (2023). Intelligent performance evaluation in rowing sport using a graph-matching network. *Journal of Imaging, 9*(9), 181. https://doi.org/10.3390/jimaging9090181

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*. https://doi.org/10.48550/arXiv.1409.1259

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. https://doi.org/10.48550/arXiv.1412.3555

Cordo, P. J., & Gurfinkel, V. S. (2004). Motor coordination can be fully understood only by studying complex movements. In *Progress in Brain Research* (Vol. 143, pp. 29–38). Elsevier. https://doi.org/10.1016/S0079-6123(03)43003-3

Cust, E. E., Sweeting, A. J., Ball, K., & Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of Sports Sciences, 37*(5), 568-600. https://doi.org/10.1080/02640414.2018.1521769

Herrebrøden, H., Jensenius, A. R., Espeseth, T., Bishop, L., & Vuoskoski, J. K. (2023). Cognitive load causes kinematic changes in both elite and non-elite rowers. *Human Movement Science, 90*, 103113. https://doi.org/10.1016/j.humov.2023.103113

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(5), e1380. https://doi.org/10.1002/widm.1380

Kleshnev, V. (2005). Comparison of on-water rowing with its simulation on Concept2 and Rowperfect machines. *International Symposium on Biomechanics in Sports, Conference Proceedings Archive, 23.* https://ojs.ub.uni-konstanz.de/cpa/article/view/853

Knudson, D. (1999). Validity and reliability of visual ratings of the vertical jump. *Perceptual and Motor Skills, 89(*2), 642-648. https://doi.org/10.2466/pms.1999.89.2.642

Lorenz D. S., Reiman M. P., Lehecka B. J., Naylor A. (2013). What performance characteristics determine elite versus nonelite athletes in the same sport? *Sports Health, 5*(6), 542-547. https://doi.org/10.1177/1941738113479763

Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F., & Baca, A. (2023). Machine learning application in soccer: a systematic review. *Biology of Sport, 40*(1), 249-263. https://doi.org/10.5114/biolsport.2023.112970

Rindal, O. M. H., Seeberg, T. M., Tjønnås, J., Haugnes, P., & Sandbakk, Ø. (2017). Automatic classification of sub-techniques in classical cross-country skiing using a machine learning algorithm on micro-sensor data. *Sensors, 18*(1), 75. https://doi.org/10.3390/s18010075

Ross, G. B., Dowling, B., Troje, N. F., Fischer, S. L., & Graham, R. B. (2020). Classifying elite from novice athletes using simulated wearable sensor data. *Frontiers in Bioengineering and Biotechnology, 8*, 814. https://doi.org/10.3389/fbioe.2020.00814

Smith, R. M., & Loschner, C. (2002). Biomechanics feedback for rowing. *Journal of Sports Sciences, 20*(10), 783-791. https://doi.org/10.1080/026404102320675639

Soper, C., & Hume, P. A. (2004). Towards an ideal rowing technique for performance. *Sports Medicine*, pp. 825-848. https://doi.org/10.2165/00007256-200434120-00003

Ste-Marie, D. M., Lelievre, N., & St. Germain, L. (2020). Revisiting the applied model for the use of observation: a review of articles spanning 2011–2018. *Research Quarterly for Exercise and Sport, 91*(4), 594-617. https://doi.org/10.1080/02701367.2019.1693489

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics, 8*, 1-15. https://doi.org/10.1186/s40634-021-00346-x

Williams, A. M., & Ericsson, K. A. (2005). Perceptual-cognitive expertise in sport: Some considerations when applying the expert performance approach. *Human Movement Science, 24*(3), 283-307. https://doi.org/10.1016/j.humov.2005.06.002

Wood, D., Reid, M., Elliot, B., Alderson, J., & Mian, A. (2023). The expert eye? An inter-rater comparison of elite tennis serve kinematics and performance. *Journal of Sports Sciences, 41*(19), 1779-1786. https://doi.org/10.1080/02640414.2023.2298102