Challenges in building machine learning models for movement recognition in sports caused by technique inconsistencies of beginners

Val Vec[1], Sašo Tomažič[1], Anton Kos[1], Anton Umek[1]
{val.vec@fe.uni-lj.si, saso.tomazic@fe.uni-lj.si, anton.kos@fe.uni-lj.si, anton.umek@fe.uni-lj.si}
[1]Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia

Abstract. This research is part of a broader effort to develop machine learning–based systems that provide real-time feedback to athletes during training. In this paper, we focus on inconsistencies in the movement techniques of beginner athletes, using dart throwing as a case study. In the first part of the paper, we demonstrate that a beginner's technique changes from day to day. We show that these changes are gradual rather than random, suggesting that both learning and forgetting play a role. In the subsequent sections, we investigate how this variability affects two tasks: identifying the individual and detecting poor technique using machine learning methods. Our findings suggest that the variability in beginners' techniques is substantial enough that methods designed for professional athletes are not directly applicable when building systems for novices.

Keywords: technique inconsistencies, biomechanical feedback, machine learning, person recognition

1 Introduction

As professional athletes strive for improvement, new technology that can assist them in their training is needed. For this purpose, researchers have developed biomechanical feedback loop systems [1] . These systems are designed to monitor athletes' movements, identify errors, and deliver feedback that helps improve their technique. They include four main components: the athlete performing the movement and responding to received feedback, sensors that capture motion data, a processing unit that analyses this data to extract relevant insights and generate feedback, and actuators that deliver this feedback to the athlete. Sensors are typically either inertial measurement sensors or cameras. Most of the studies utilising artificial intelligence for feedback use cameras [2], however, cameras have two significant disadvantages over IMU sensors. They require setup and calibration in the room where they are used. Because the setup requires time before each use, they are not ideal for professional athletes, whose free time is limited. Additionally, due to their higher cost, they are also unsuitable for cost-conscious amateurs. For these reasons, we focus all our research on solutions utilising IMU sensors.

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365256

When developing feedback applications, we must focus on a sport-specific motion technique. That can be a throw in basketball, a swing in golf or any other isolated motion. It is not possible to build a feedback system that would work universally, as which technique is correct is defined by the sport. Furthermore, in most sports, professional athletes have their own specific technique, which is a variation of the correct technique for that particular movement in sport [3]. Consistency in technique is key for professional athletes. However, beginners and amateurs are not as consistent.

For our experiments, we selected the dart throw as a representative example of a sport-specific movement technique. This motion was chosen because it is easy for complete beginners to understand, which allowed us to include more novice participants in the study. We expect our findings to be transferable to other motions in sports, where the duration is relatively short, such as throws in various sports, golf swings, strokes in swimming, strokes in tennis and so on.

In this paper, we explore variations in techniques that may complicate the development of effective feedback models. A recent review of machine learning applications in sports revealed that many studies rely on subject-specific models [4], underscoring the importance of developing personalised approaches. Some applications allow for feedback based on a universal technique, and others require personalised feedback to account for individual differences in performance. Furthermore, the differences between individuals are large enough that multiple studies manage to use machine learning to successfully identify the person based on recorded IMU signals [5, 6]. The models for person recognition can serve as proof that there are significant differences in technique between individuals. As we show later in the paper, these differences are often more significant and easier to identify than the differences between correct and incorrect movement, which is what we need to recognise in order to provide feedback.

The models that recognise a person are also useful by themselves. In order to build an application that provides personalised feedback to the athlete, such as [7], we must recognise which personalised model or non-machine learning algorithm to use for that individual.

In this paper, we explore the consistency of each individual's personal technique. When the athlete changes their technique from one recording session to another, this introduces difficulties when building useful models. For professional athletes, there should be no difference between each recording session, and therefore we should not be able to recognise from which recording session the recorded movement is. This is not the case for amateurs and especially beginners, and we can build machine learning models to recognise recording sessions easily, proving that there is significant variation in a person's technique from one day to another.

We analyse whether the technique of amateurs is consistent enough to recognise them across multiple recording sessions on various days, and whether the technique changes so much that we cannot build a model that would differentiate between good and bad technique across multiple days.

All experiments in this paper are conducted on a dataset of 3486 dart throws, all done by amateur players. It is expected that the throwing technique would be more consistent across recording sessions if we used professional players. However, this work outlines the difficulties we face when building a machine learning model to assist beginners in their training.

2 Dataset

The experiments in this study were conducted using a subset of a custom dataset, consisting of recordings from participants who took part in at least two measurement sessions. This subset contains 3486 recordings of dart throws collected from 13 participants. One recording represents a single throw and spans a duration of one second or 100 samples.

For two of the participants, we collected data across 7 and 8 measurement sessions, respectively, over a span of two months. They were instructed to complete as many as possible throws. All other participants completed two measurement sessions, with intervals ranging from three days to one week between sessions. They performed 60 throws per session each. The final number of measurements differs as we removed some measurements during the data clean-up process.

For experiments that require more than 2 measurement sessions, we use data from two participants, while for others, we use data from all participants. All participants were amateurs, self-reporting playing darts between never to once per month. When analysing the data, we need to consider that due to multiple measurement sessions for the two participants with more data, they had significantly more recent practice than the other participants, which can explain some of the results.

Throws were captured with a custom-built device worn on the back of the hand, featuring an Adafruit Feather M0 microcontroller and a Bosch BNO085 sensor. For each throw, we have captured multivariate time series of sensor data with a one-second length. Raw measured data collected included accelerometer, gyroscope, and magnetometer signals. In addition to those, the sensor also has an internal sensor fusion algorithm, which allows us to capture linear acceleration, which is 3D acceleration without the component of gravity, vector of gravity and the sensor's orientation. The sampling rate was 100Hz, which is enough to capture all information about the motion during throws.

In addition to measurements from sensors, we also collected the results of each throw and information regarding the sports activity of the participants. The latter was done by utilising an anonymised questionnaire, where we asked the participants about the frequency of their sport activity, the frequency of playing sports that involve a ball and the frequency of playing darts.

The experiment aimed to simulate natural dart-throwing conditions, using a standard-size dartboard and standard throwing distance, with participants throwing in sets of three.

3 Methods

We want our experiments to prove whether there is a detectable difference between classes. To do that, we will classify data using a 1D convolutional neural network, which is established as successful for the classification of time-series [8]. The model is successful in classifying between players and measurement sessions. We do not alter the methodology between experiments, because we want the negative results to mean that the differences between classes are less pronounced than in other experiments.

The used 1D CNN model comprises three convolutional layers with kernel sizes of 5, 5, and 3, respectively, each followed by batch normalisation and pooling. We used max pooling for the first two layers, and adaptive average pooling for the final layer. The final output is passed through a fully

connected layer and a SoftMax activation to yield class probabilities. The models were trained using the Adam optimizer with learning rate of 0.001 for 100 epochs with a batch size of 4. The dataset was randomly split into training and test sets using a 60/40 ratio. Accuracy confidence intervals (CI) were obtained by non-parametric bootstrapping (10 000 resamples of the test data).

For comparison, the first experiment was also conducted with two additional models: a session-wise centroid + 1-NN classifier with Dynamic Time Warping (DTW) distance and a K-Nearest Neighbour classifier, with k = 5.

Due to different number of recordings for different measurement sessions and subjects, we do not have balanced classes in all experiments. To account for class imbalance, class weights were incorporated into the loss function using weighted cross-entropy during training.

Ensuring consistent sensor placement across measurement sessions is inherently challenging. We placed the sensor device on the hand and aligned it with the participants knuckles, to achieve as consistent sensor placement as possible as shown in fig. 1. Even with this placement method, minor variations in sensor orientation when attaching the device to the hand can introduce session-specific information. This information could cause us to come to a potentially wrong conclusion that the technique of the person has changed between sessions. To address this issue, we use a preprocessing method for handling 3D linear acceleration data.

The BNO085 sensor provides linear acceleration measurements in its local coordinate frame, along with the gravity vector. Both are 3D vectors. Using this information, we compute acceleration components in the vertical and horizontal directions. This transformation yields two acceleration signals that are invariant to sensor orientation. Vertical acceleration is the component of acceleration in the vertical direction, with negative values meaning acceleration downwards. The horizontal component does not carry the information about the direction of the acceleration in the horizontal plane, but we assume it points towards the target. It is a non-negative scalar value.

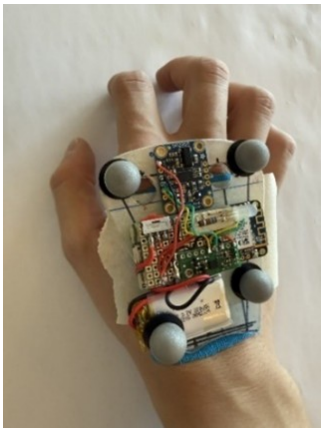

Fig. 1. Sensor device is attached to the back of the hand.

Table 1: Binary classification between the first and second measurement sessions for each subject.

Subject ID	Acc. 1D CNN (95% CI)	Acc. KNN	Acc. Nearest Centroid DTW	Test Samples
0, 3–10	1.00 [1.00, 1.00]	1.00	1.00	45–173
1	0.98 [0.95, 1.00]	0.94	0.93	72
11	0.93 [0.84, 1.00]	0.86	0.87	46
12	0.82 [0.71, 0.93]	0.73	0.71	46
13	0.85 [0.75, 0.93]	0.79	0.89	48

The two signals are calculated using the following formulas:

$$\mathrm{LA}_{\mathrm{ver}} = \frac{\vec{g} \cdot \vec{\mathrm{LA}}}{\|\vec{g}\|} \tag{1}$$

$$\mathrm{LA}_{\mathrm{hor}} = \left\| \vec{\mathrm{LA}} - \mathrm{LA}_{\mathrm{ver}} \cdot \frac{\vec{g}}{\|\vec{g}\|} \right\| \tag{2}$$

In all our experiments, we use only these two signals to train the model. The model input, therefore consists of two time series, each containing 100 time steps.

4 Results

4.1 Identifying the measurement session using machine learnings

The idea behind this experiment is that if the machine learning model can be trained to recognise a recording session for each person, we can conclude that the individual's throwing technique of the individual has significantly changed from one day to another. In Table 1, we provide results of classifying throws into either the first or second measurement session for each subject. If the person's technique was perfectly consistent, the accuracy would be 50% as there would be no information about the measurement session in the captured signal.

The results from Table 1 show that the technique of most participants changed significantly between measurement sessions, as the session could be identified perfectly. We have tested two additional models, and concluded, that the model selection did not have a meaningful impact on the results. All other experiments were therefore conducted only using the 1D CNN.

When we analysed what is different for participants where we could not recognise the measurements session perfectly, we found no pattern in their answers regarding the frequency of darts playing or sports activity. Participants with IDs 1, 11,12 and 13, which are the ones where we did not get 100% accuracy, were the only ones where both measurement sessions were conducted within 3 days, with all other participants having 6 days or more between both recording sessions.

This indicates that the amateurs who take an extended time between playing sport change their technique each time. This change in personal technique provides additional challenges to overcome when building models that rely on one's individual technique, which should be trained. However,

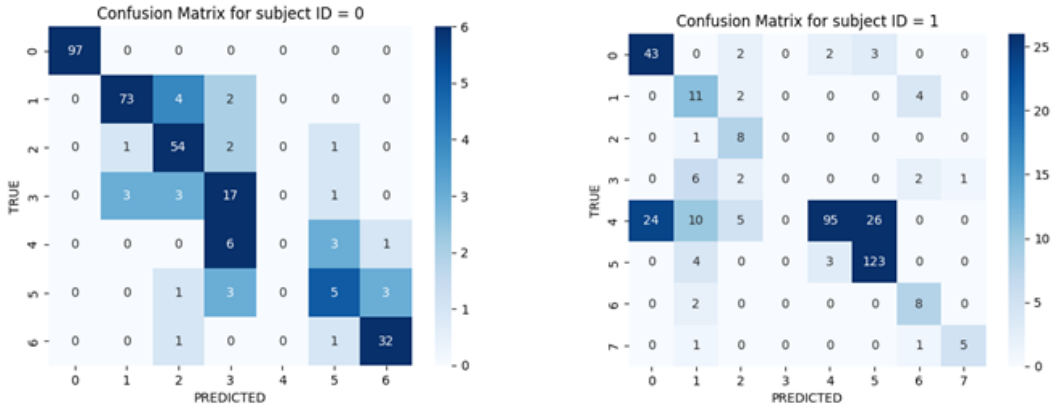

Fig. 2. Confusion matrices for subjects 0 and 1, respectively, show which measurement sessions get misclassified with which sessions. The scale is set to emphasise misclassified samples.

it is likely that the individuals who would use such systems train regularly and would have a more consistent personal technique.

For participants with IDs 0 and 1, we conducted more than two measurement sessions, and therefore, we can analyse which of the measurement sessions can be separated. Fig. 2 shows confusion matrices when building a model to classify measurement sessions for participants with IDs 0 and 1. Most of the measurement sessions can be separated well, with one of the participants showing gradual changes in technique, as consecutive sessions get mixed more often. For the other participant, this trend is less pronounced.

We built binary classifiers to separate each pair of measurement sessions for participants with IDs 0 and 1. The highest accuracy was 100% and the lowest was 67%

To understand, which factors influence how well a given session can be distinguished from others, we analysed various session-level metrics. These included the average throw accuracy measured as the average distance from the centre of the target, the session's order, such as first or second, and the number of sessions separating the two in each pair. We calculated the Pearson correlation coefficient [9] to determine the linear relationship between machine learning accuracy and the above-mentioned metrics. The correlations with machine learning accuracy are shown in Table 2. We calculated the correlation for each metric for each class in the pair. The number following each metric name in the table indicates whether it refers to the first or second class in the pair, with the first class always corresponding to the earlier of the two measurement sessions.

Results from Table 2 show that for subject with ID 0, we can statistically significantly more accurately separate earlier sessions and sessions with worse throwing accuracy. Sessions that happened in a shorter time frame are statistically significantly harder to separate. This is the expected result, as it shows that the technique stabilises and changes gradually. The worse throwing accuracy being correlated with better ML results indicates that being more consistent in technique has led to better throw outcomes. For person 1, none of the correlations are statistically significant. They show

Table 2: Correlation of machine learning accuracy with session-level features. Bold values indicate statistically significant correlations ($p < 0.05$).

Subject ID	Metric	Correlation with ML accuracy	p-value
0	Throw accuracy 1	**0.482**	**0.0270**
0	Throw accuracy 2	0.092	0.6908
0	Session's order 1	**-0.586**	**0.0052**
0	Session's order 2	-0.051	0.8255
0	Sessions in between	**0.535**	**0.0124**
1	Throw accuracy 1	0.276	0.1544
1	Throw accuracy 2	0.332	0.0842
1	Session's order 1	-0.132	0.5038
1	Session's order 2	0.057	0.7723
1	Sessions in between	0.191	0.3291

similar but weaker correlations than for subject 0.

These correlations however did not pass Holm adjustment [10], meaning that this may be a false positive driven by testing multiple features. Therefore, the results from Table 2 should be further validated on a larger dataset in future work.

We can conclude that the technique of amateurs changes from one measurement session to another. Consecutive sessions are harder to separate, indicating a gradual change in technique, possibly explained by the player learning to throw better. Given that there was a difference between subjects that had the first two measurement sessions in the same week versus those that had their second measurement session in the following week, we can infer that the change also happens with time while not playing, which could be explained by players forgetting the technique.

4.2 Effect of inter-session technique variability on person recognition

As shown in the previous section, the person's technique changes over time for amateur athletes. In this section, we analyse how much this affects person recognition.

If we train the person recognition model on data from all measurement sessions, we can achieve 98% accuracy. All participants can be recognised accurately. This is the baseline result.

When training the model on data from only half of the measurement sessions, trying to classify the recordings from the other sessions, the model fails to recognise most of the participants. The exceptions are subjects 0 and 1, for whom we had multiple sessions for training and subjects 10,11 and 12, whose technique was more consistent, as already shown in section IV.A, likely due to less time between measurement sessions. We provide the confusion matrix as the only metric because the conclusions can be drawn based on which classes are misclassified as which. With the model being more successful for certain classes, overall accuracy is not meaningful.

Fig. 3 and 4 show confusion matrices when either training on half of the measurement sessions for subjects 0 and 1, or training on only one measurement session. For subjects 2 through 12, training on half of the sessions is the same as training on one, as we only have 2 sessions.

We can interpret the results from Fig. 3 in two possible ways. One possible explanation is that the model struggles to learn person-specific features because it is not trained across multiple sessions per individual. As a result, it may fail to distinguish between session-related variability and the consistent characteristics that differentiate individuals. The other explanation would be that the differences in amateurs' technique between sessions are just as big as the differences between individuals. The fact that subjects 0 and 1, who have multiple measurement sessions, are more accurately recognised than the rest suggests that the first explanation is more likely. However, even if training on only the first measurement session, as shown in Fig. 4, the model can accurately identify theparticipant with ID 1, suggesting that it is possible that their technique was simply more consistent.

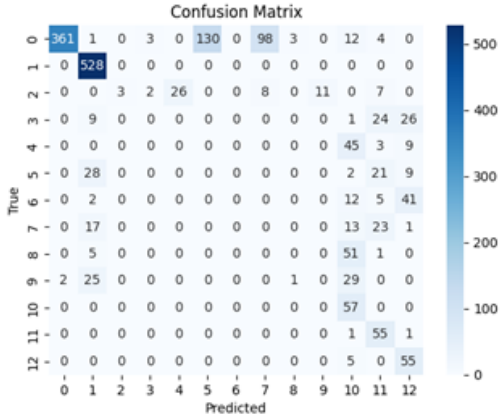

Fig. 3. Confusion matrix when classifying a person using a classifier trained on data from different measurement sessions.

When trying to build a classifier, with a pair of measurement sessions and subject ID as class, the model fails to differentiate between measurement sessions for subjects 0 and 1, suggesting that the differences between personal techniques are still larger than the variance in technique from one session to another. However, for all other participants, the model managed to separate measurement sessions.

All in all, the variability in the technique of amateurs makes building a person recognition model challenging. Data from only one measurement session is not sufficient, as the model trained on all data achieves 98% accuracy, while the model trained on only one session completely fails for a large number of subjects. This is expected to change for professional players, and our findings only apply to amateurs and beginners.

4.3 Effect of inter-session technique variability on throw outcome classification

Classifying throw outcomes is a much more difficult task. Many factors affect the outcome of the throw, which are not captured by our measurements, the largest of them being aim, which

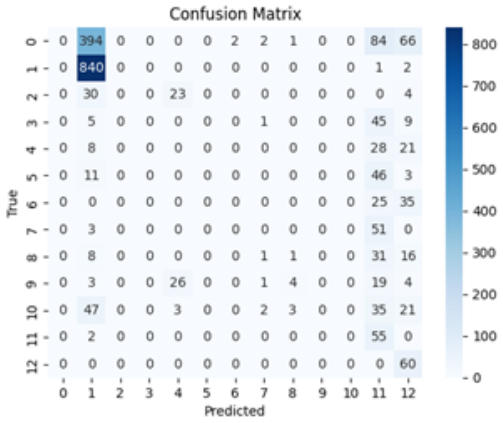

Fig. 4. Confusion matrix when classifying a person using a classifier trained on data from only the first measurement sessions.

cannot be captured by acceleration signals. For this reason, even in ideal scenario, with participant having consistent technique and using the perfect machine learning method, we cannot expect perfect accuracy. We can, however, compare accuracies across different scenarios. In this section, we want to explore whether the personalised model trained on all of the person's data outperforms the model trained on data from just one measurement session in recognising the good throws.

In target sports, performance is measured by precision, measured by how closely the hits cluster together and accuracy, which is measured by the hits' proximity to the target center. We define "good throws" as those with more consistent outcomes, meaning better throwing precision rather than accuracy, to minimise the influence of aiming, which cannot be measured with IMU sensors, and focus instead on variations caused by technique, which are measured. Another reason for selecting precision is that good precision, but poor accuracy means a systematic error you can correct with aim, while good accuracy but low precision means random errors that only better technique can fix. For each set of three throws, we first compute the centre point. Then, we calculate the average distance of the throws from this centre point. Throws with an average distance below the median distance across the entire dataset are labelled as good throws.

In Table 3, we compare the results between using data from only one measurement session and using all data for that person.

The results of classifying good versus bad throws are, in most cases, better than chance, but far from good enough to be useful for feedback applications. For both participants, most models that were built on one session outperformed the model built on all of that person's data. For the participant with ID 0, the model built on all data completely fails, which can be explained by a lack of consistency in this person's technique. For the participant with ID 1, who was significantly better at throwing darts, the model built on all data performs better than chance. Most session-specific models outperform it, suggesting that the technique variations are introducing significant amounts

Table 3: Comparison of machine learning results for models that classify good and bad throws, trained on different measurement sessions.

ID	Session	Samples	ML Acc.	AUC
1	all	963	0.55	0.57
1	0	372	0.50	0.58
1	1	141	0.56	0.66
1	2	327	0.58	0.58
1	3	39	0.40	0.50
1	4	30	0.54	0.73
1	5	21	0.87	1.00
1	6	33	0.83	0.72
1	7	21	0.38	0.53
0	all	783	0.50	0.53
0	0	234	0.56	0.59
0	1	198	0.46	0.44
0	2	138	0.53	0.60
0	3	57	0.73	0.74
0	4	24	0.50	0.57
0	5	42	0.33	0.28
0	6	90	0.58	0.60

of noise into our training data.

All in all, building a machine learning model to recognise mistakes in technique for beginners, who do not yet have an established technique, is not the right approach. Machine learning based feedback applications might only be suitable for athletes with higher skill and with smaller variations in technique.

5 Discussion and Conclusion

In this paper, we have concluded that the inconsistency in personal technique of amateur athletes is large enough to introduce problems when building ML model for movement recognition. We could perfectly separate most of the first measurement sessions from the second, except for those that happened in a short time frame. This can be explained by the athlete forgetting the technique they used last time. This part of the inconsistency is most likely random and detrimental to good performance in sports.

Results of ML separation of later measurement sessions suggest that as the technique stabilises with learning, indicated by ML accuracy being correlated with both the throw result and the session order. This indicates that there is another component to the inconsistency, which can be explained by the person learning proper technique.

These variations in technique are great enough that training models for person recognition based on only measurements from one session is not possible. The same goes for building the models

that would recognise good throws, where separating data by session leads to more successful results.

While professional athletes should have more consistent techniques and these concerns do not apply to them, our findings prove the need to rethink the approach of building personalised assistance systems for amateur athletes.

One way we could approach this is with an adaptive model that would change with time. For person recognition specifically, we could use a few-shot methods using only the last few measurements for each person and then update the dataset upon successful recognition.

As for feedback models, it is possibly a better idea not to use personalised models until the individual has a stable and effective technique. We have to ask ourselves what techniques we are trying to reinforce with feedback, if the throws with good outcomes come from different techniques on different days.

The main limitation of this paper is the dataset size. The tests that required more that two measurements sessions were all done on the data from 2 subjects. In the future our findings should be validated on data from more people. Furthermore, future work should include validation on different sports.

Acknowledgments

This research was funded in part by the Slovenian Research Agency within the research program ICT4QoL - Information and Communications Technologies for Quality of Life, grant number P2-0246.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT and InstaText in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Kos A, Umek A. Biomechanical Biofeedback Systems and Applications. Human–Computer Interaction Series. Cham: Springer International Publishing; 2018.

[2] Naik BT, Hashmi MF, Bokde ND. A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions. Applied Sciences. 2022;12(9):Art. no. 9.

[3] Schöllhorn W, Bauer H. Identifying individual movement styles in high performance sports by means of self- organizing Kohonen maps; 1998. .

[4] Vec V, Tomažič S, Kos A, Umek A. Trends in real-time artificial intelligence methods in sports: a systematic review. Journal of Big Data. 2024 Oct;11(1):148.

[5] Zhang Z, Zhang Y, Kos A, Umek A. Strain Gage Sensor Based Golfer Identification Using Machine Learning Algorithms. Procedia Computer Science. 2018 Jan;129:135-40.

[6] Yao ZM, Zhou X, Lin ED, Xu S, Sun YN. A novel biometric recognition system based on ground reaction force measurements of continuous gait. In: 3rd International Conference on Human System Interaction; 2010. p. 452-8.

[7] Vec V, Tomažič S, Kos A, Umek A. User interface design for orientation sensor-based biomechanical feedback in golf. In: 2024 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI); 2024. p. 172-7.

[8] Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing. 2021 Apr;151:107398.

[9] SciPy Community. pearsonr — SciPy v1.16.0 Manual; 2025. Accessed: July 30, 2025. https://docs.scipy.org/doc/scipy-1.16.0/reference/generated/scipy.stats.pearsonr.html.

[10] Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6(2):65-70.