frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 04 August 2025
DOI 10.3389/frai.2025.1611534

OPEN ACCESS

EDITED BY
Maria Concetta Carruba,
Pegaso University, Italy

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Levent Ceylan,
Hitit University, Türkiye

*CORRESPONDENCE
Hubert Makaruk
✉ hubert.makaruk@awf.edu.pl

RECEIVED 14 April 2025
ACCEPTED 21 July 2025
PUBLISHED 04 August 2025
CORRECTED 16 October 2025

CITATION
Makaruk H, Porter JM, Webster EK, Makaruk B, Tomaszewski P, Nogal M, Gawłowski D, Sobański Ł, Molik B and Sadowski J (2025) Artificial intelligence-enhanced assessment of fundamental motor skills: validity and reliability of the FUS test for jumping rope performance.
Front. Artif. Intell. 8:1611534.
doi: 10.3389/frai.2025.1611534

Artificial intelligence-enhanced assessment of fundamental motor skills: validity and reliability of the FUS test for jumping rope performance

Hubert Makaruk[1]*, Jared M. Porter[2], E. Kipling Webster[2], Beata Makaruk[1], Paweł Tomaszewski[3], Marta Nogal[1], Daniel Gawłowski[4], Łukasz Sobański[5], Bartosz Molik[6] and Jerzy Sadowski[1]

[1]Faculty of Physical Education and Health in Biala Podlaska, Józef Piłsudski University of Physical Education in Warsaw, Warsaw, Poland, [2]Department of Kinesiology, Recreation, and Sport Studies, University of Tennessee, Knoxville, TN, United States, [3]Faculty of Physical Education, Józef Piłsudski University of Physical Education in Warsaw, Warsaw, Poland, [4]Artificial Intelligence Department, DG Consulting, Wrocław, Poland, [5]Department of Research in Artificial Intelligence, Instat sp. z o.o., Wrocław, Poland, [6]Faculty of Rehabilitation, Józef Piłsudski University of Physical Education in Warsaw, Warsaw, Poland

Introduction: Widespread concerns about children's low fundamental motor skill (FMS) proficiency highlight the need for accurate assessment tools to support structured instruction. This study examined the validity and reliability of an AI-enhanced methodology for assessing jumping rope performance within the Fundamental Motor Skills in Sport (FUS) test.

Methods: A total of 236 participants (126 primary school students aged 7–14; 110 university sports students aged 20–21) completed jumping rope tasks recorded via the FUS mobile app integrated with an AI model evaluating five process-oriented performance criteria. Concurrent validity and inter-rater reliability were examined by comparing AIgenerated assessments with scores from two expert evaluators. Intra-rater reliability was also assessed through reassessment of video trials after a 3-week interval.

Results: Results revealed excellent concurrent validity and inter-rater reliability for the AI model compared with expert ratings (ICC = 0.96; weighted kappa = 0.87). Agreement on individual criteria was similarly high (Cohen's kappa = 0.83–0.87). Expertadjusted AI scores further improved reliability (ICC = 0.98). Intrarater reliability was also excellent, with perfect agreement for AIgenerated scores (ICC = 1.00; kappa = 1.00).

Conclusions: These findings demonstrate that AI-based assessment offers objective, reliable, and scalable evaluation, enhancing accuracy and efficiency of FMS assessment in education and research.

KEYWORDS

motor competence, fundamental movement skills, machine learning, mobile application, physical education

Introduction

Assessing and improving movement proficiency during childhood is crucial for supporting optimal physical development and building the foundational skills necessary for lifelong physical activity (Logan et al., 2015; Stodden et al., 2008). Fundamental motor skills (FMS), including jumping, running, and throwing, form the foundation for more complex motor tasks, facilitating children's effective participation in diverse sports and physical activities (Logan et al., 2018). A large body of research indicates that proficiency in FMS is associated with enhanced physical fitness and a reduced risk of lifestyle-related health issues, including obesity and cardiovascular conditions (Barnett et al., 2022; Robinson et al., 2015; Stodden et al., 2008). Conversely, children with insufficient FMS often face barriers for participation in sports and social play, which can negatively affect their physical, emotional, and social development (Robinson et al., 2015).

Process-oriented assessments are widely recognized as the paramount benchmark for evaluating FMS, offering detailed insights into how movements are performed (Logan et al., 2017; Watanabe et al., 2024). Unlike product-oriented tools that emphasize measurable outcomes such as distance or speed, process-oriented evaluations focus on coordination, control, and other qualitative factors, providing deeper insights into the mechanics and efficiency of motor performance (Logan et al., 2018). These tools are particularly valuable for identifying deficits in specific movement components, which are essential for advanced skill proficiency. For instance, O'Brien et al. (2016) highlighted that poor execution of fundamental behavioral components, such as bending the knees and extending the arms during take-off, was a common reason for failure in both vertical and horizontal jumps.

Despite their predictive value for assessing FMS proficiency, implementing process-oriented assessments poses substantial challenges, even for trained experts and researchers (Hulteen et al., 2018; Hulteen et al., 2023; Ward et al., 2020). Ensuring validity and reliability requires extensive training and strict adherence to protocols, making the process resource-intensive and time-consuming (Lander et al., 2015). Variability in scoring frequently arises from subjective interpretations of movement quality, particularly during real-time assessments, where observers must simultaneously track multiple criteria, compromising accuracy (Barnett et al., 2014; Ward et al., 2020). The issue becomes more pronounced with less experienced evaluators, emphasizing the need to improve training procedures, assessment frameworks, and clearly defined evaluation criteria (Lander et al., 2015; Palmer and Brian, 2016). Additionally, inherent constraints in process-oriented assessments limit the number of performance criteria that human observers can reliably evaluate. For example, widely-used tools, such as the Test of Gross Motor Development (Webster and Ulrich, 2017), address this limitation by removing criteria during an item-analysis phase if they cannot be consistently scored by human observers. However, this practice may result in essential movement characteristics remaining unassessed. Barnett et al. (2014) illustrate this problem, noting that inadequately defined criteria, such as not specifying elbow flexion, can cause incorrect movements (e.g., a "fling" or "sling") to be incorrectly scored as correct. Similarly, tools like the Victorian Fundamental Motor Skill Manual (Walkley et al., 1996) provide partial guidance but fail to fully represent the complexity and subtlety inherent in FMS, further complicating accurate assessment. These limitations collectively underscore the need for more comprehensive assessment strategies.

In many countries, early education (classroom) and physical education (PE) teachers are responsible for monitoring FMS development as part of the curriculum (Lander et al., 2016; Makaruk et al., 2024b). However, many teachers lack confidence in utilizing process-oriented tools, which often demand a deeper understanding of human movement and FMS development (Bourke et al., 2024; Harris et al., 2011; Lander et al., 2015). This knowledge gap can lead to inconsistent evaluations and an over-reliance on simplified tools that fail to capture the complexities of motor skills (Morley et al., 2019). Additionally, limited training and professional development opportunities contribute to these obstacles, leaving many teachers unprepared to effectively assess and address movement skill deficiencies within the constraints of their teaching schedules (Draper et al., 2019). Research highlights that teachers often prioritize instructional activities over assessments due to the significant time and logistical burdens associated with administering traditional FMS tools, particularly in large class settings (Draper et al., 2019; Morley et al., 2019). Morley et al. (2019) noted that these tools are not only time-intensive but also require specialized expertise, making their integration into daily teaching routines difficult. Furthermore, limited access to updated FMS testing methodologies and appropriate resources frequently forces teachers to rely on simplified or outdated tools, which fail to capture the full scope of movement skills (Bourke et al., 2024; Foweather et al., 2018; Morley et al., 2019). As a result, assessments are often infrequent and incomplete, reducing the effectiveness of identifying and addressing movement deficiencies in students.

Addressing these challenges requires innovative solutions that preserve the depth and accuracy of process-oriented assessments (Bisi et al., 2017; Hulteen et al., 2020; Ward et al., 2017) while enhancing their feasibility. Digital technology, such as tablet-based applications, has shown great potential in this regard and is well-received by primary school teachers. These tools provide functionalities like video recording and analysis, enabling educators to capture and review children's performances effectively (Browne, 2015; Draper et al., 2019). Digital tools streamline the assessment process and provide visual evidence of movement execution, enabling more accurate identification of skill deficiencies. Embedded video demonstrations further support teachers with limited expertise, offering references for skill performance and guidance for conducting assessments (Foweather et al., 2018; Makaruk et al., 2024a; O'Loughlin et al., 2013). One such example is Meu Educativo®, a platform designed to evaluate FMS using an expert-validated checklist and rating system (Garbeloto et al., 2024). The tool prioritizes ease of use, maintaining reliability with inter-rater reliability ranging from 0.63 to 0.93 and intra-rater reliability from 0.46 to 0.94. Another example is the FUS test app, which optimizes the evaluation process by providing teachers and researchers with tools to efficiently record, analyze, and score performances (Makaruk et al., 2024a). Validation studies demonstrated strong concurrent validity ($r = 0.92$–0.96) and excellent intra-rater reliability (ICC > 0.91), supporting its use as a reliable tool for assessing FMS.

Emerging technologies like artificial intelligence (AI) build on advancements in digital tools, may offer transformative potential to enhance the accessibility, reliability, and accuracy of FMS assessments (Vandevoorde et al., 2022). AI-powered tools can automate labor-intensive tasks, such as video analysis and scoring, reducing the need for extensive training and systematic observation scoring. For example, Zhang et al. (2024) validated the precision of AI-powered tools in analyzing biomechanical markers, effectively detecting subtle

deficits in balance and coordination that are often overlooked with conventional feedback from supervisors. Similarly, Sganga et al. (2023) confirmed the effectiveness of integrating smartphone-based inertial motion units with AI algorithms to accurately estimate the relative displacement between the center of mass and the center of pressure. Another study (Hajihosseini et al., 2022) established the feasibility of automated assessments for FMS like overhand throwing. The AI-powered system, which employs wearable inertial measurement units and the 'k-nearest neighbor' algorithm, reduced scoring time from 5 min to less than 30 s while maintaining high accuracy, with classification rates ranging from 76 to 93% across four specific criteria. AI-based tools have proven effective in evaluating key rope-jumping performance parameters, including jumping pace, flight time, touchdown time, and jump height, achieving excellent reliability (ICC > 0.9) (Yu and Hu, 2022). Collectively, these advancements underscore the transformative potential of AI in bridging the gap between traditional, resource-intensive methods and scalable solutions for fostering motor skill learning and development.

Despite increasing use of digital tools in PE, there remains a pressing need to validate AI-powered systems capable of delivering standardized, efficient, and scalable FMS assessments. The primary aim of this study was to evaluate the validity and reliability of an AI-enhanced methodology for assessing jumping rope performance within the FUS test framework. This approach was intended to address key limitations of traditional process-oriented methods, such as scoring variability, time constraints, and the need for specialized expertise, by examining the AI model's capacity to deliver objective and consistent movement analysis. We hypothesized that integrating AI systems could standardize FMS evaluations, reduce logistical barriers, and support broader implementation in both educational and research settings. These assumptions underpin a broader vision of improving the accessibility and quality of motor skill assessments while advancing evidence-based practice through data-driven innovation.

Methods

Participants

A total of 236 students participated in this study, comprising 126 primary school students aged 7–14 years (54% female) and 110 university sports students aged 20–21 years (45% female). Older participants were included to validate the assessment protocol across a broader age and skill spectrum, extending beyond the originally targeted age range of the FUS test. Participants were eligible if they were actively enrolled in physical activity-related educational programs offered by their institution and had no medical conditions, musculoskeletal injuries, or neurological disorders affecting jumping performance. Written informed consent was obtained from all participants or their legal guardians, and the study protocol was approved by the institutional Research Ethics Committee.

Apparatus and technology

The FUS test app, developed for use on mobile phones and tablets, was employed to assess proficiency in six sports-related tasks, including jumping rope performance. The app facilitated video recording, analysis, and scoring based on predefined performance criteria. To evaluate the task of jumping rope, an AI-driven assessment system was incorporated to provide an automated and objective evaluation of movement proficiency. The AI assessment utilized the MoveNet model, an open-source deep neural network developed by Google (TensorFlow, 2021). The model tracked the horizontal and vertical coordinates of 17 body parts at a frequency of 10 frames per second, providing precise motion data for detailed analysis. The anatomical landmarks included the nose, left and right eyes, left and right ears, left and right shoulders, left and right elbows, left and right wrists, left and right hips, left and right knees, and left and right ankles.

Jump rope proficiency was evaluated by analyzing five key movement characteristics aligned to the FUS (reference) scoring criterion, each assessed using a machine learning model specifically developed for the corresponding criterion: continuity (criterion 1), rhythmicity (criterion 2), arm and wrist positioning (criterion 3), hip and knee flexion (criterion 4), and central positioning (criterion 5). Continuity was modeled using the Extra Trees Classifier (Geurts et al., 2006), an ensemble learning algorithm that analyzed temporal patterns to identify whether jumps were performed continuously without interruptions. Rhythmicity, representing the consistency and timing of jumps, was evaluated with a Gradient Boosting Classifier (Friedman, 2001) designed to detect deviations in timing across consecutive cycles. Arm and wrist positioning was assessed using another Gradient Boosting Classifier, which analyzed the spatial and temporal precision of arm and wrist movements during the swinging phase. Hip and knee flexion was modeled using a Multilayer Perceptron Neural Network Classifier (Rumelhart et al., 1986), a network capable of detecting subtle variations in joint angles. Finally, central positioning was assessed with a Gradient Boosting Classifier, which evaluated spatial alignment and postural control to ensure jumps were executed within the designated area while maintaining an upright trunk position. Each of these models was designed to provide automated, objective assessments of movement proficiency, scoring individual criteria on a binary scale where 1 indicated correct execution and 0 indicated incorrect execution. In cases where the AI system failed to detect the participant's movements with sufficient confidence, such as poor video quality issues or obstruction, a score vector of $(-1, -1, -1, -1, -1)$ was returned, indicating a failed recognition attempt. The AI model assessment utilized open-source libraries, including TensorFlow (2021), licensed under Apache 2.0, and scikit-learn (Pedregosa et al., 2011), licensed under the BSD License.

Data collection was conducted using a Lenovo Tab P11 (2nd Gen) tablet, equipped with a 13 MP high-resolution camera capable of recording 1080p HD videos at 30 frames per second. The tablet's 2,000 × 1,200 resolution screen provided high-quality playback, supporting accurate video analysis.

Procedure

The jumping rope task in the FUS test required participants to perform rhythmic and continuous jumps over the rope for 10 s. The app, however, allowed for a 15-s video recording, capturing both the preparation phase (1–2 s) and the task execution. The following standardized procedure was implemented to ensure consistency across

trials. At the start of each trial, the participant stood directly in front of the test administrator, positioned approximately 4 m from the camera, which was aligned to face the center of the "X" marked on the floor. The administrator provided clear instructions: "Jump to the rhythm of the rope hitting the ground," and ensured the participant was prepared to proceed. Upon confirmation, the administrator activated the recording function in the application and instructed the participant to begin the task by saying "Go." The app automatically stopped recording after 15 s, ensuring standardized durations across all trials.

Before beginning the warm-up trial, participants were instructed to adopt an upright stance, holding the handles of the rope behind their body. Arms were positioned close to the trunk, with elbows bent and externally abducted. The length of the rope was adjusted according to the participant's height, ensuring that, when folded in half, it extended from the floor to the shoulder. All jumps were performed on a flat wooden surface to ensure safety and consistency in testing conditions. To preserve optimal video analysis and minimize potential distractions, no other individuals were permitted near or in the background of the participant during the task. Additionally, the background was required to provide sufficient contrast with the participant's clothing (e.g., light-colored clothing against a dark background) to enhance visibility and ensure accuracy in movement assessment.

To ensure clarity and correct execution, the test supervisor presented the task before the trials commenced. Participants observed the demonstration from a position directly in front of the supervisor. Following the instructional phase, each participant completed a warm-up trial to familiarize themselves with the task. Subsequently, two test trials were conducted, with a minimum rest interval of 3 min between trials to mitigate fatigue. No feedback was provided during testing.

The criteria for jumping rope are as follows: criterion 1. jumps are performed continuously (without stopping); criterion 2. jumps are rhythmic and single, with short ground contact time and landing on the ball of the feet; criterion 3. arms are bent and held close to the trunk, and the rope is moved using the rotation of forearms and wrists; criterion 4. knees and hips are slightly bent during flight and landing; criterion 5. jumps are performed vertically with jumps initiating in the same designated area, with the trunk upright, feet parallel at a hip width apart.

Each criterion was scored as "1" if met or "0" if unmet, with points awarded only when performance clearly satisfied the respective criterion. The higher-scoring attempt was considered for further analysis.

Concurrent validity and inter-rater reliability between FMS experts and AI model

In this study, concurrent validity evaluated how well the AI model replicated expert evaluations, recognized as the benchmark for jumping rope proficiency. Inter-rater reliability analyzed the consistency of scores between the AI model and human assessors. Both psychometric evaluations employed the same statistical measures. Two experienced assessors independently scored 236 video-recorded performances, resolving disagreements through a consensus process to ensure accuracy. Prior to consensus, the assessors achieved at least 90% agreement on total scores. A third assessor utilized the AI model to independently evaluate the same performance criteria. The AI-generated scores, produced automatically by the software's predefined algorithms, were then directly compared with scores assigned by the human experts. This comparison allowed for an evaluation of both the validity and reliability of the AI assessment system.

Inter-rater reliability between FMS experts correcting AI-generated scores

A second analysis was conducted to assess the inter-rater reliability of the AI model after adjustments by human experts. Two independent evaluators reviewed the AI-generated scores for the same set of video-recorded performances ($n = 236$). When the AI-generated scores differed from the expert's judgment, the experts manually adjusted (corrected) these scores to align them with their expert evaluation. Corrections were made only when the expert was clearly satisfied that the AI-generated point did not accurately reflect the actual observed performance.

Intra-rater reliability of the AI model and intra-rater reliability between FMS experts correcting AI-generated scores

The intra-rater reliability of the AI model, operated independently by an expert, and the AI model corrected by two expert raters was independently evaluated by comparing scores from the initial and follow-up assessments of the same video-recorded jumping rope performances ($n = 236$) after a three-week interval. Corrections were applied only when the expert was fully confident that a point should or should not be awarded for a specific criterion.

Statistical analysis

Descriptive statistics were reported as means and standard deviations for all variables. Concurrent validity, inter-rater and intra-rater reliability were assessed using the percentage of observed agreements, Cohen's kappa coefficients for individual criteria, and weighted kappa coefficients for total points, along with intraclass correlation coefficients (ICCs). Cohen's kappa values were interpreted based on the classification proposed by Landis and Koch (1977), where values <0 indicate poor agreement, 0.01–0.20 indicate slight agreement, 0.21–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate substantial agreement, and 0.81–1.00 indicate almost perfect agreement. ICC values were interpreted as follows: values <0.5 indicate poor reliability, 0.5–0.75 indicate moderate reliability, 0.75–0.9 indicate good reliability, and >0.9 indicate excellent reliability (Koo and Li, 2016). For both Cohen's kappa and ICC values, 95% confidence intervals (CIs) were calculated to provide a measure of precision. The statistical significance threshold was set at alpha = 0.05 for all analyses. Data were analyzed using SPSS Statistics version 27 for Windows (SPSS Inc., Chicago, USA).

Results

Concurrent validity and inter-rater reliability

The total scores for the jumping rope assessment were nearly identical between the FMS experts (3.12 ± 1.80) and the AI model

(3.12 ± 1.77). As presented in Table 1, the observed agreement for individual performance criteria ranged from 92.8% (criteria 4 and 5) to 94.5% (criterion 2). Cohen's kappa values ranged from 0.83 (criterion 5) to 0.87 (criterion 2), indicating almost perfect inter-rater agreement. The total score demonstrated strong consistency between raters, with an observed agreement of 77.1%, a Cohen's kappa of 0.87 (95% CI: 0.84–0.91), and an excellent ICC of 0.96. Minor discrepancies between expert and AI scores occurred evenly across all criteria, with no consistent direction of bias.

Concurrent validity and inter-rater reliability between FMS experts correcting AI-generated scores

Table 2 summarizes the results for AI-generated scores adjusted by FMS experts. Observed agreements were consistently high, ranging from 96.6% (criterion 4) to 98.7% (criterion 1). Cohen's kappa coefficients varied between 0.93 (criteria 4 and 5) and 0.97 (criterion 1), consistently reflecting almost perfect agreement. For the total score, observed agreement was 89.0%, Cohen's kappa was 0.94, and the ICC for the total score was excellent at 0.98 (95% CI: 0.98–0.99). The two evaluators showed strong consistency across all criteria, with minor discrepancies evenly distributed and no consistent scoring bias identified. Differences between evaluators ranged from 3 to 8 cases per criterion, with total score agreement in 210 of 236 assessments.

Intra-rater reliability for AI model and intra-rater reliability between FMS experts correcting AI-generated scores

The AI model exhibited perfect consistency, achieving 100% observed agreement and a kappa coefficient of 1.00 across all criteria and the total score (Table 3). After correction by FMS experts, intra-rater agreement remained high, with observed agreements ranging from 98.3 to 99.1% for individual criteria and from 92.4 to 94.5% for the total score. Kappa coefficients ranged from 0.95 to 0.98 for individual criteria and were 0.97 and 0.96 for the total score, depending on the expert. The ICC for the total score was consistently excellent, reaching 0.99 for both experts.

Both experts maintained high consistency between initial and subsequent assessments across all criteria, with only minor, non-systematic discrepancies observed. Differences between sessions ranged from 2 to 5 cases per criterion, and total score consistency was high (expert 1: 223 out of 236 cases; expert 2: 218 out of 236 cases).

Discussion

This study confirmed the validity and reliability of an AI-enhanced methodology for assessing jumping rope performance within the FUS test framework. The AI model closely aligned with expert evaluations, as shown by high correlation coefficients and near-perfect agreement across both individual criteria and total scores. When refined by expert input, the model's inter-rater reliability improved further, demonstrating its capacity for iterative enhancement. Intra-rater reliability was consistently high, with the AI model alone achieving perfect agreement (kappa = 1.00, ICC = 1.00); after expert corrections, reliability remained robust, with only a minor reduction. These results support the use of AI systems as accurate, consistent tools for FMS evaluation, with clear implications for research and PE practice.

First, the strong agreement between the AI model and expert scores demonstrates its capacity to replicate expert evaluations with high accuracy and consistency. This advancement addresses long-standing limitations of traditional assessment methods, such as evaluator bias, fatigue-induced errors, and inter-rater variability (Hulteen et al., 2023). The AI system applies standardized criteria uniformly, thereby reducing the influence of subjective judgment and eliminating inconsistencies across evaluators. This is especially

TABLE 1 Concurrent validity and inter-rater reliability for jumping rope assessment between FMS experts and AI model (*n* = 236).

Jumping rope	Percentage of observed agreements (%)	Cohen's kappa (95% CI)
Criterion 1	93.2	0.86 (0.79–0.92)
Criterion 2	94.5	0.87 (0.81–0.94)
Criterion 3	93.6	0.87 (0.80–0.93)
Criterion 4	92.8	0.86 (0.79–0.92)
Criterion 5	92.8	0.83 (0.76–0.91)
Total score	77.1	0.87 (0.84–0.91)

TABLE 2 Inter-rater reliability for jumping rope assessment between experts correcting AI-generated scores (*n* = 236).

Jumping rope	Percentage of observed agreements (%)	Cohen's kappa (95% CI)
Criterion 1	98.7	0.97 (0.94–1.00)
Criterion 2	97.5	0.94 (0.90–0.99)
Criterion 3	97.5	0.94 (0.90–0.99)
Criterion 4	96.6	0.93 (0.88–0.98)
Criterion 5	97.0	0.93 (0.88–0.98)
Total score	89.0	0.94 (0.92–0.97)

TABLE 3 Intra-rater reliability for jumping rope assessment for AI model and experts correcting AI-generated scores.

Jumping rope	Percentage of observed agreements (%)	Cohen's kappa (95% CI)	ICC (95% CI)
Criterion 1			
AI model	100.0	1.00 (1.00–1.00)	
Corrected AI-Expert 1	99.1	0.98 (0.96–1.00)	
Corrected AI-Expert 2	98.7	0.97 (0.94–1.00)	
Criterion 2			
AI model	100.0	1.00 (1.00–1.00)	
Corrected AI-Expert 1	98.7	0.97 (0.94–1.00)	
Corrected AI-Expert 2	98.3	0.96 (0.93–1.00)	
Criterion 3			
AI model	100.0	1.00 (1.00–1.00)	
Corrected AI-Expert 1	98.3	0.96 (0.93–1.00)	
Corrected AI-Expert 2	98.7	0.97 (0.94–1.00)	
Criterion 4			
AI model	100.0	1.00 (1.00–1.00)	
Corrected AI-Expert 1	98.3	0.97 (0.93–1.00)	
Corrected AI-Expert 2	97.9	0.96 (0.92–0.99)	
Criterion 5			
AI model	100.0	1.00 (1.00–1.00)	
Corrected AI-Expert 1	99.1	0.98 (0.95–1.00)	
Corrected AI-Expert 2	97.9	0.95 (0.91–0.99)	
Total			
AI model	100.0	1.00 (1.00–1.00)	1.00 (1.00–0.1.00)
Corrected AI-Expert 1	94.5	0.97 (0.96–0.99)	0.99 (0.99–0.99)
Corrected AI-Expert 2	92.4	0.96 (0.94–0.98)	0.99 (0.98–0.99)

valuable in multi-site studies and large-scale educational settings, where consistent conditions are essential (Barnett et al., 2014). Moreover, the model supports post-hoc analysis of video recordings, enhancing transparency and reliability in performance evaluations (Atkinson and Nevill, 1998; Leukel et al., 2023; Ward et al., 2020). These features position the AI model as a viable and scalable alternative to traditional human-led assessments, particularly in contexts where time, training, and resources are constrained.

Second, the AI model shows strong potential for iterative refinement, leveraging performance data to continuously improve its scoring accuracy—an advantage over human experts, who often rely on fixed methods and experiential judgment (Song et al., 2021; Vandevoorde et al., 2022). Unlike human raters, whose perceptual limitations may lead to inconsistency in identifying nuanced movement errors, AI systems can detect subtle deficits with precision. Barnett et al. (2014) illustrated this limitation in their analysis of complex motion sequences like the overhand throw, where evaluators frequently disagreed during rapid movement phases such as the wind-up or follow-through. These discrepancies, compounded by the cognitive load of observing multi-joint actions in real time, underscore the limits of human perception. In contrast, AI systems can process detailed kinematic data objectively and without fatigue, ensuring greater accuracy and consistency in movement evaluation.

Third, the alignment between AI-derived evaluations and expert assessments highlights the system's adaptability to applied settings, including PE classes. For teachers with limited time or specialized training, the AI model offers a practical and efficient solution, providing consistent evaluations without reliance on subjective judgment. By addressing scoring variability and logistical constraints, it facilitates timely, individualized interventions based on accurate performance feedback. Morley et al. (2019) similarly observed that digital tools enhance assessment efficiency and reduce resource demands, enabling educators to prioritize evidence-based instruction. These benefits position AI-enhanced tools as valuable assets for improving the quality, accessibility, and standardization of motor skill assessments, ultimately supporting physical literacy and motor competence development.

Finally, the validated AI model offers new opportunities to democratize access to high-quality FMS assessments, particularly in resource-limited educational contexts. While this study provides an initial demonstration of the model's utility, broader implementation could transform PE by enabling scalable, AI-powered evaluation systems. Kok et al. (2020) showed that digital tools, such as self-controlled video feedback, enhance motor learning and self-efficacy by promoting autonomy and self-regulation. Building on this foundation, AI systems may empower students to assess and reflect on their performance independently, reducing reliance on direct teacher instruction while

sustaining accuracy and engagement. Moreover, such tools can support peer-to-peer assessments through structured, video-based applications enriched with instructional cues. This collaborative model fosters a motivating, student-centered learning environment and helps accommodate challenges like limited instructional capacity and large class sizes. Through the refinement of traditional assessment practices, AI-enhanced tools have the potential to promote both physical literacy and educational equity. Future research should explore the integration of adaptive feedback features to optimize motor learning and encourage student agency in FMS development.

Inter-rater reliability results confirmed the AI model's capacity to replicate expert judgments with high consistency, with near-perfect agreement across criteria and excellent ICC values. These outcomes were comparable to or exceeded inter-rater reliability levels reported in studies involving human experts (Makaruk et al., 2023; Zamani et al., 2024), underscoring the system's potential as a standardized evaluation tool. Notably, previous research has highlighted difficulties even among trained assessors. For instance, Hulteen et al. (2023) reported component-level agreement as low as 40.5%, particularly for tasks like skipping, while Ward et al. (2020) observed criterion-level accuracy ranging from 35 to 100%, often due to attentional lapses and inconsistent interpretation of scoring criteria. Incorporating expert oversight into the AI model improved inter-rater reliability further, suggesting that a hybrid approach—combining algorithmic precision with expert refinement—can elevate assessment quality. This process reduces subjective biases and ensures consistent application of criteria, thereby enhancing both the reliability and scalability of motor skill evaluations in research and applied settings.

Intra-rater reliability analysis reinforced the robustness of the AI model, particularly for longitudinal monitoring of motor skill development. The system achieved perfect agreement across all criteria and total scores, with kappa and ICC values reaching 1.00—a level virtually impossible to attain for human experts in intra-rater FMS assessments. This level of precision supports the use of AI tools for progress tracking, timely intervention, and performance optimization in educational and developmental contexts. Even after expert corrections, reliability remained high, exceeding benchmarks reported in prior studies (Makaruk et al., 2023; Zamani et al., 2024). These findings suggest that AI-assisted assessment can enhance data-driven decision-making in both research and applied practice, while maintaining consistency over time.

Discrepancies between the AI model and expert evaluations were generally minor and varied across criteria, indicating non-systematic biases rather than consistent directional errors. For instance, the AI model occasionally overestimated performance in Criterion 1 by assigning higher scores when participants quickly resumed jumping after an error, overlooking brief interruptions in continuity. In contrast, it adopted a more conservative approach in Criterion 2, underestimating performance due to prolonged ground contact. This bidirectional pattern indicates stochastic variation rather than a systematic scoring bias. Additionally, slight inconsistencies across repeated expert assessments suggest that some variation may stem from the subjective nature of human judgment rather than model limitations. These findings reinforce the importance of ongoing algorithmic refinement to better align with expert standards and ensure robust applicability in real-world assessment contexts.

Several limitations should be acknowledged. First, the model's accuracy depends on controlled testing conditions, including precise camera alignment, sufficient lighting, and a contrast-enhancing background—which may be challenging to reproduce in certain PE class environments. Second, the study focused solely on jumping rope, which, while informative, limits the generalizability of findings to other FMS with differing biomechanical and coordination demands. Future research should extend the framework to a broader range of skills. Lastly, minor over-and underestimations by the AI in specific criteria highlight the need for continued algorithmic refinement to improve sensitivity to nuanced movement variations and ensure alignment with expert standards.

Conclusion

This study provides robust evidence supporting the validity and reliability of an AI-enhanced methodology for assessing jumping rope performance within the FUS test framework. The AI model effectively replicated expert-level evaluations, and its performance was further strengthened through expert refinement. These results highlight the system's potential to overcome key limitations of traditional assessment methods by offering scalable, objective, and consistent evaluations. Beyond its research value, the model presents a practical solution for educational and sport contexts, particularly in settings with limited resources. With further development, such tools may help democratize access to high-quality motor skill assessments and support broader efforts to enhance motor competence and health outcomes across populations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by the study (no. SKE 01-19/2022) was approved by a university Institutional Review Board—The Józef Piłsudski University of Physical Education in Warsaw. This study was performed in line with the principles of the Declaration of Helsinki. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

HM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. JP: Conceptualization, Writing – original draft, Writing – review & editing, Methodology, Supervision, Validation. EW: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Methodology, Supervision, Validation. BeM: Conceptualization, Writing – original draft, Writing – review & editing, Investigation. PT: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Validation. MN:

Conceptualization, Writing – original draft, Writing – review & editing, Investigation. DG: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Software. ŁS: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Software. BaM: Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Funding acquisition, Resources, Supervision. JS: Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Funding acquisition, Supervision.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research was financed from the program "WF with AWF-Active Return to School after the Pandemic" awarded by the Polish Ministry of Science and Higher Education. Agreement No. MEiN/2022/DPI/37.

Conflict of interest

DG was employed by DG Consulting. LS was employed at Instat sp. z o.o.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Correction note

A correction has been made to this article. Details can be found at: 10.3389/frai.2025.1710897.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1611534/full#supplementary-material

References

Atkinson, G., and Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.* 26, 217–238. doi: 10.2165/00007256-199826040-00002

Barnett, L. M., Minto, C., Lander, N., and Hardy, L. L. (2014). Interrater reliability assessment using the test of gross motor development-2. *J. Sci. Med. Sport* 17, 667–670. doi: 10.1016/j.jsams.2013.09.013

Barnett, L. M., Webster, E. K., Hulteen, R. M., De Meester, A., Valentini, N. C., Lenoir, M., et al. (2022). Through the looking glass: a systematic review of longitudinal evidence, providing new insight for motor competence and health. *Sports Med.* 52, 875–920. doi: 10.1007/s40279-021-01516-8

Bisi, M. C., Panebianco, G. P., Polman, R., and Stagni, R. (2017). Objective assessment of movement competence in children using wearable sensors: an instrumented version of the TGMD-2 locomotor subtest. *Gait Posture* 56, 42–48. doi: 10.1016/j.gaitpost.2017.04.025

Bourke, M., Haddara, A., Loh, A., Saravanamuttoo, K. A., Bruijns, B. A., and Tucker, P. (2024). Effect of capacity building interventions on classroom teacher and early childhood educator perceived capabilities, knowledge, and attitudes relating to physical activity and fundamental movement skills: a systematic review and meta-analysis. *BMC Public Health* 24:1409. doi: 10.1186/s12889-024-18907-x

Browne, T. (2015). A case study of student teachers' learning and perceptions when using tablet applications teaching physical education. *Asia Pac. J. Health Sport Phys. Educ.* 6, 3–22. doi: 10.1080/18377122.2014.997858

Draper, N., Jusary, R., and Marshall, H. (2019). The validity and reliability of video assessment for the dragon challenge. *Int. J. Spa Wellness* 2, 154–165. doi: 10.1016/j.humov.2024.103302

Foweather, L., Van Rossum, T., Richardson, D., Hayes, S., and Morley, D. (2018). Primary teachers' recommendations for the development of a teacher-oriented movement assessment tool for 4-7 year children. *Meas. Phys. Educ. Exerc. Sci.* 23, 124–134. doi: 10.1080/1091367X.2018.1552587

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Garbeloto, F., Pereira, S., Tani, G., Chaput, J. P., Stodden, D. F., Garganta, R., et al. (2024). Validity and reliability of meu Educativo®: a new tool to assess fundamental movement skills in school-aged children. *Am. J. Hum. Biol.* 36:e24011. doi: 10.1002/ajhb.24011

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Hajihosseini, S., Sheikh, M., Ghaiur Najafabadi, M., Houminiyan Sharif Abadi, D., Veisi, H., and Bagherzadeh, F. (2022). Automated assessment of developmental stage of overhand throwing skill. *J. Adv. Sport Technol.* 6, 99–111. doi: 10.22098/jast.2022.2119

Harris, J., Cale, L., and Musson, H. (2011). The effects of a professional development programme on primary school teachers' perceptions of physical education. *Prof. Dev. Educ.* 37, 291–305. doi: 10.1080/19415257.2010.531973

Hulteen, R. M., Barnett, L. M., True, L., Lander, N. J., del Pozo Cruz, B., and Lonsdale, C. (2020). Validity and reliability evidence for motor competence assessments in children and adolescents: a systematic review. *J. Sports Sci.* 38, 1717–1798. doi: 10.1080/02640414.2020.1756674

Hulteen, R. M., Morgan, P. J., Barnett, L. M., Stodden, D. F., and Lubans, D. R. (2018). Development of foundational movement skills: a conceptual model for physical activity across the lifespan. *Sports Med.* 48, 1533–1540. doi: 10.1007/s40279-018-0892-6

Hulteen, R. M., True, L., and Kroc, E. (2023). Trust the "process"? When fundamental motor skill scores are reliably unreliable. *Meas. Phys. Educ. Exerc. Sci.* 27, 391–402. doi: 10.1080/01942638.2025.2463350

Kok, M., Komen, A., van Capelleveen, L., and van der Kamp, J. (2020). The effects of self-controlled video feedback on motor learning and self-efficacy in a physical education setting: an exploratory study on the shot-put. *Phys. Educ. Sport Pedagog.* 25, 49–66. doi: 10.1080/17408989.2019.1688773

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Lander, N. J., Barnett, L. M., Brown, H., and Telford, A. (2015). Physical education teacher training in fundamental movement skills makes a difference to instruction and assessment practices. *J. Teach. Phys. Educ.* 34, 548–556. doi: 10.1123/jtpe.2014-0043

Lander, N., Morgan, P. J., Salmon, J., and Barnett, L. M. (2016). Teachers' perceptions of a fundamental movement skill (FMS) assessment battery in a school setting. *Meas. Phys. Educ. Exerc. Sci.* 20, 50–62. doi: 10.1080/1091367X.2015.1095758

Landis, J. R., and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33, 363–374. doi: 10.2307/2529786

Leukel, C., Leuders, T., Bessi, F., and Loibl, K. (2023). Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics. *Front. Educ.* 8:1162499. doi: 10.3389/feduc.2023.1162499

Logan, S. W., Barnett, L. M., Goodway, J. D., and Stodden, D. F. (2017). Comparison of performance on process-and product-oriented assessments of fundamental motor skills across childhood. *J. Sports Sci.* 35, 634–641. doi: 10.1080/02640414.2016.1183803

Logan, S. W., Ross, S. M., Chee, K., Stodden, D. F., and Robinson, L. E. (2018). Fundamental motor skills: a systematic review of terminology. *J. Sports Sci.* 36, 781–796. doi: 10.1080/02640414.2017.1340660

Logan, S. W., Webster, E. K., Getchell, N., Pfeiffer, K. A., and Robinson, L. E. (2015). Relationship between fundamental motor skill competence and physical activity during childhood and adolescence: a systematic review. *Kinesiol. Rev.* 4, 416–426. doi: 10.1123/kr.2013-0012

Makaruk, H., Porter, J. M., Webster, E. K., Makaruk, B., Bodasińska, A., Zieliński, J., et al. (2023). The FUS test: a promising tool for evaluating fundamental motor skills in children and adolescents. *BMC Public Health* 23:1912. doi: 10.1186/s12889-023-16843-w

Makaruk, H., Porter, J. M., Webster, E. K., Makaruk, B., Niznikowski, T., and Sadowski, J. (2024a). Concurrent validity and reliability of the FUS test app for the measurement of fundamental motor skills. *J. Sport Exerc. Psychol.* 46:S36.

Makaruk, H., Webster, E. K., Porter, J., Makaruk, B., Bodasińska, A., Zieliński, J., et al. (2024b). The fundamental motor skill proficiency among polish primary school-aged children: a nationally representative surveillance study. *J. Sci. Med. Sport* 27, 243–249. doi: 10.1016/j.jsams.2023.12.007

Morley, D., Van Rossum, T., Richardson, D., and Foweather, L. (2019). Expert recommendations for the design of a children's movement competence assessment tool for use by primary school teachers. *Eur. Phys. Educ. Rev.* 25, 524–543. doi: 10.1177/1356336X17751358

O'Brien, W., Belton, S., and Issartel, J. (2016). Fundamental movement skill proficiency amongst adolescent youth. *Phys. Educ. Sport Pedagog.* 21, 557–571. doi: 10.1080/17408989.2015.1017451

O'Loughlin, J., Chróinín, D. N., and O'Grady, D. (2013). Digital video: the impact on children's learning experiences in primary physical education. *Eur. Phys. Educ. Rev.* 19, 165–182. doi: 10.1177/1356336X13486050

Palmer, K. K., and Brian, A. (2016). Test of gross motor development-2 scores differ between expert and novice coders. *J. Mot. Learn. Dev.* 4, 142–151. doi: 10.1123/jmld.2015-0035

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Robinson, L. E., Stodden, D. F., Barnett, L. M., Lopes, V. P., Logan, S. W., Rodrigues, L. P., et al. (2015). Motor competence and its effect on positive developmental trajectories of health. *Sports Med.* 45, 1273–1284. doi: 10.1007/s40279-015-0351-6

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sganga, M., Rozmiarek, P., Ravera, E., Akanyeti, O., and Povina, F. V. (2023). Automatic balance assessment using smartphone and AI. In: Proceedings of the 41st Conference on Computer Graphics and Visual Computing (CGVC 2023). Eurographics Association.. doi: 10.2312/cgvc.20231206

Song, S., Kidziński, Ł., Peng, X. B., Ong, C., Hicks, J., Levine, S., et al. (2021). Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *J. Neuroeng. Rehabil.* 18:126. doi: 10.1186/s12984-021-00919-y

Stodden, D. F., Goodway, J. D., Langendorfer, S. J., Roberton, M. A., Rudisill, M. E., Garcia, C., et al. (2008). A developmental perspective on the role of motor skill competence in physical activity: an emergent relationship. *Quest* 60, 290–306. doi: 10.1080/00336297.2008.10483582

TensorFlow. (2021). Next-generation pose detection with MoveNet and TensorFlow.js. TensorFlow Blog. Available online at: https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html (Accessed July 24, 2025).

Vandevoorde, K., Vollenkemper, L., Schwan, C., Kohlhase, M., and Schenck, W. (2022). Using artificial intelligence for assistance systems to bring motor learning principles into real world motor tasks. *Sensors* 22:2481. doi: 10.3390/s22072481

Walkley, J., Holland, B. V., Treloar, R., and O'Connor, J. (1996). Fundamental motor skills: a manual for classroom teachers. Victoria: Department of Education.

Ward, B., Thornton, A., Lay, B., Chen, N., and Rosenberg, M. (2020). Can proficiency criteria be accurately identified during real-time fundamental movement skill assessment? *Res. Q. Exerc. Sport* 91, 64–72. doi: 10.1080/02701367.2019.1646852

Ward, B. J., Thornton, A., Lay, B., and Rosenberg, M. (2017). Protocols for the investigation of information processing in human assessment of fundamental movement skills. *J. Mot. Behav.* 49, 593–602. doi: 10.1080/00222895.2016.1247033

Watanabe, M., Hikihara, Y., Aoyama, T., Wakabayashi, H., Hanawa, S., Omi, N., et al. (2024). Associations among motor competence, health-related fitness, and physical activity in children: a comparison of gold standard and field-based measures. *J. Sports Sci.* 42, 1644–1650. doi: 10.1080/02640414.2024.2404781

Webster, E. K., and Ulrich, D. A. (2017). Evaluation of the psychometric properties of the test of gross motor development – third edition. *J. Mot. Learn. Dev.* 5, 45–58. doi: 10.1123/jmld.2016-0003

Yu, C., and Hu, H. (2022). A method to evaluate rope-jumping skills based on smartphone and open pose. In: Proc. 3rd Int. Conf. Artif. Intell. Inf. Process. Cloud Comput. (AIIPCC 2022).

Zamani, M. H., Hashemi, A., Siavashi, E., Khanmohamadi, R., and Saeidi, H. (2024). Validity and reliability of the fundamental motor skills in sports (FUS) test for Iranian children and adolescents. *Int. J. Sch. Health* 11, 157–169. doi: 10.30476/intjsh.2024.101089.1368

Zhang, Y., Li, H., and Huang, R. (2024). The effect of Tai Chi (Bafa Wubu) training and artificial intelligence-based movement-precision feedback on the mental and physical outcomes of elderly. *Sensors* 24:6485. doi: 10.3390/s24196485