Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 13 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i13.62395

PAPER

A Mobile Multimodal Interactive System for Sports Skill Assessment in Educational Contexts

Dan Lv, Huijun Li (✉)

Inner Mongolia University, Hohhot, China

lhj111971110@163.com

ABSTRACT

Traditional sports skill assessment has been constrained by subjective bias, delayed feedback, and limited evaluation dimensions, thereby restricting improvements in instructional quality. The integration of mobile technologies and multimodal interaction has introduced new opportunities for precise skill assessment; however, existing systems commonly suffer from inadequate mobile adaptability, inefficient multimodal data collaboration, and limited suitability for authentic teaching scenarios. In this study, a mobile multimodal interactive skill assessment system tailored for sports education was designed to enable real-time, accurate, and multidimensional evaluation of sports skills in support of instructional optimization. A four-layer architecture comprising mobile terminals, data transmission, cloud-based processing, and application services was established. Multimodal data acquisition modules were integrated with lightweight fusion algorithms, and system effectiveness was examined through requirement analysis, architectural design, prototype development, and teaching-oriented experimental validation. The results demonstrate that the proposed system exhibits robust stability in mobile environments and achieves high multimodal data fusion accuracy. Compared with traditional assessment approaches and existing systems, assessment accuracy is significantly improved, and reliable application performance is observed in practical sports teaching contexts. The primary contribution lies in the development of a lightweight multimodal acquisition and real-time assessment framework specifically adapted to sports education scenarios. This work enriches the theoretical foundation of mobile intelligent assessment and provides critical technical support for advancing the intelligent transformation and scientific assessment of sports teaching.

KEYWORDS

mobile multimodal interaction, sports skill assessment, mobile intelligent systems, sports education, real-time data fusion

1 INTRODUCTION

The global transformation toward educational digitalization has become irreversible [1–3]. As a critical branch of this transformation, the intelligent development of

Lv, D., Li, H. (2026). A Mobile Multimodal Interactive System for Sports Skill Assessment in Educational Contexts. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(13), pp. 69–83. https://doi.org/10.3991/ijim.v20i13.62395

Article submitted 2026-03-19. Revision uploaded 2026-05-02. Final acceptance 2026-05-20.

sports education [4, 5] reflects both policy orientations and urgent practical demands. Traditional sports skill assessment has long relied on manual and subjective judgment, resulting in widespread limitations such as scoring bias, single-dimensional data representation, and delayed feedback. Consequently, learners' skill proficiency has been difficult to quantify accurately, thereby constraining improvements in instructional quality [6, 7].

The rapid advancement of mobile technologies [8] provides a viable pathway for addressing these challenges. The widespread adoption and portability of mobile terminals, including smartphones and wearable devices, have removed temporal and spatial constraints inherent in conventional assessment settings [9, 10], enabling real-time acquisition and transmission of skill-related data during sports instruction. Concurrently, the emergence of multimodal interaction technologies has further expanded the evaluative scope of sports skill assessment [11, 12]. Through the coordinated analysis of multimodal data—such as visual information, movement patterns, and physiological signals—the core characteristics of sports skills can be represented more comprehensively and objectively. This multimodal perspective effectively overcomes the limitations associated with single-modality assessment and provides robust technical support for precision-oriented evaluation.

Existing studies have preliminarily explored the application of mobile technologies in sports education, with reported outcomes primarily concentrated on instructional resource delivery and basic data logging functions [13]. However, most existing systems lack targeted sports skill assessment capabilities and exhibit insufficient integration of multimodal data, thereby failing to meet the requirements of accurate evaluation. Within the domain of multimodal interaction and skill assessment, fusion algorithms based on deep learning and machine learning have demonstrated considerable potential. Nevertheless, current research has largely focused on laboratory-controlled environments. The resulting models are often characterized by high computational complexity and excessive power consumption, rendering them unsuitable for the lightweight, low-latency requirements of mobile application scenarios.

From the perspective of overall research on sports skill assessment systems, traditional manual evaluation approaches, although grounded in extensive practical experience, are characterized by strong subjectivity and a low degree of standardization. Existing technology-based systems have improved assessment objectivity; however, they commonly exhibit insufficient alignment with instructional scenarios, limited real-time performance, and suboptimal user friendliness. As a result, the dynamic characteristics of sports teaching environments and authentic instructional requirements have not been adequately accommodated. In summary, a mobile-oriented multimodal interactive sports skill assessment solution specifically adapted to sports education scenarios has not yet been established. The synergistic advantages of mobile technologies and multimodal interaction remain underexploited, and this study gap constitutes the central entry point of the present study.

The primary objectives of this study are threefold. First, a multimodal interactive skill assessment architecture adapted to mobile sports education scenarios is designed to achieve deep integration between instructional contexts and technical frameworks. Second, key challenges related to efficient multimodal data acquisition, real-time transmission, and accurate data fusion are addressed in order to enhance processing efficiency and precision. Third, a prototype system is developed, and its assessment performance is validated through teaching-oriented experiments, thereby providing empirical support for practical instructional application.

2 SYSTEM DESIGN AND IMPLEMENTATION

2.1 Overall system architecture

Based on the results of the requirement analysis, a four-layer distributed architecture comprising a mobile terminal layer, data transmission layer, cloud-based processing layer, and application service layer was adopted. Data interaction and functional coordination across layers are achieved through standardized interfaces. The core design principle is to enhance system scalability and adaptability to mobile instructional scenarios through layered decoupling while simultaneously ensuring real-time performance and accuracy of data processing. The overall system framework is illustrated in Figure 1. The core functionalities, technical configurations, and interaction logic of each layer are described below.

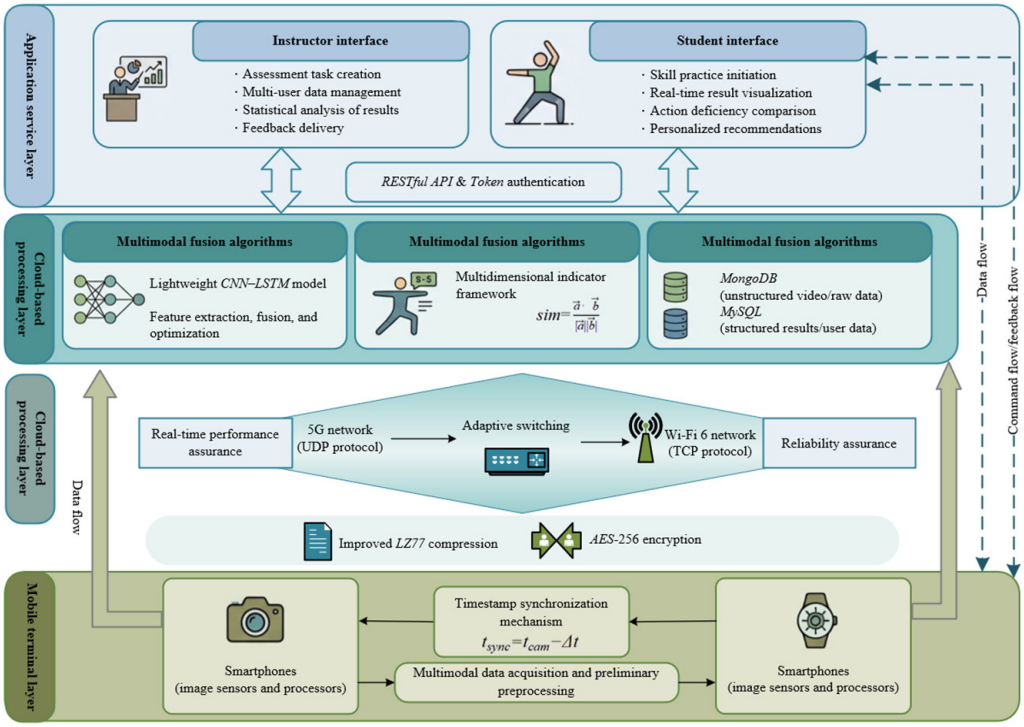

Fig. 1. Mobile multimodal interactive skill assessment system for sports education

The mobile terminal layer, serving as the primary component for data acquisition and front-end interaction, is responsible for raw multimodal data collection, preliminary preprocessing, and user command response. A collaborative hardware configuration combining smartphones and wearable sensors is employed. Smartphones equipped with high-performance image sensors and processors are selected, while lightweight six-axis inertial measurement unit (IMU) modules are used as wearable sensors. The multimodal data acquisition module achieves driver-level adaptation of the camera and IMU sensors through a hardware abstraction layer (HAL). A timestamp synchronization mechanism is implemented to ensure spatiotemporal consistency among heterogeneous data sources. The synchronization relationship is expressed as $t_{sync} = t_{cam} - \Delta t$, where t_{cam} denotes the camera acquisition timestamp and Δt represents the fixed temporal offset between sensors, which is controlled within 1 ms. After preliminary preprocessing, the collected data are transmitted to

the data transmission layer, while user interaction commands from the application service layer are concurrently handled.

The data transmission layer is responsible for enabling efficient and secure data exchange between mobile terminals and the cloud, with primary emphasis placed on addressing network fluctuations, limited bandwidth, and data security challenges inherent to mobile environments. An adaptive protocol switching strategy is adopted for data transmission. Specifically, the user datagram protocol (UDP) is selected under 5G network conditions to ensure real-time performance, whereas the transmission control protocol (TCP) is employed in Wi-Fi 6 environments to guarantee transmission reliability. Protocol switching is automatically triggered when the network packet loss rate exceeds 3%. Data compression is performed using an improved LZ77 algorithm. By constructing a domain-specific dictionary tailored to sports motion data, compression efficiency is enhanced. The compression ratio (CR), defined as the ratio of the original data volume to the compressed data volume, remains stable between 4 and 6, effectively reducing transmission bandwidth consumption. Data security is ensured through end-to-end encryption using the AES-256 algorithm, while data integrity is protected via a checksum verification mechanism. In addition, traffic control and retransmission mechanisms are integrated within the layer. When network latency exceeds a predefined threshold, data buffering and batch transmission are automatically activated to prevent data loss.

The cloud-based processing layer functions as the central computational node of the system and is responsible for multimodal data fusion, skill assessment, and data management. The multimodal fusion module adopts a lightweight, improved convolutional neural network–long short-term memory (CNN–LSTM) model. Model computational complexity is reduced by decreasing the number of convolutional kernels and simplifying the fully connected layer structure, thereby enabling efficient lightweight deployment in cloud environments. The fusion process is divided into three stages: feature extraction, feature fusion, and feature optimization, with a unified motion feature vector produced as the final output. The skill assessment model is constructed based on the sports education curriculum framework and incorporates a multidimensional indicator system. Quantitative scoring is achieved by computing the cosine similarity between the action features to be evaluated and a reference action template, expressed as:

$$sim = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{b}|} \tag{1}$$

The data storage and management module adopts a hybrid architecture combining MongoDB and MySQL. MongoDB is utilized for storing unstructured motion videos and raw sensor data, whereas MySQL is employed for managing structured assessment results, user profiles, and instructional statistical data. Query efficiency is enhanced through data indexing optimization.

The application service layer, serving as the primary interface between users and the system, provides differentiated functional services for the instructor interface and the student interface. Interface development is implemented using a responsive design framework, and cross-platform compatibility is achieved through Flutter-based development, ensuring display consistency across heterogeneous terminal devices. Core functions of the instructor interface include assessment task creation, multi-user data management, statistical analysis of assessment outcomes, and batch feedback delivery, with support for the generation of visualized data reports. Core functions of the student interface include skill practice initiation,

real-time assessment result visualization, comparative analysis of motion deficiencies through video playback, and access to personalized improvement recommendations. Within the layer, data interaction with the cloud-based processing layer is realized via RESTful Application Programming Interfaces (APIs). A token-based authentication mechanism is adopted to ensure secure user access, while user operation logs are recorded to provide data support for system optimization. Through standardized data formats and interface protocols, coordinated operation across all layers is achieved, forming an integrated workflow encompassing data acquisition, transmission, processing, and service delivery, thereby ensuring efficient system operation in sports education scenarios.

2.2 Implementation of core modules

Mobile multimodal data acquisition module. The primary objective of the mobile multimodal data acquisition module is to achieve efficient coordinated acquisition of visual and inertial data, while ensuring data quality and adaptability to mobile application scenarios. Visual data acquisition is performed using the built-in red–green–blue (RGB) camera of mobile devices and is governed by a dynamic parameter adaptation strategy. For sports activities characterized by different motion speeds—such as basketball shooting and short-distance track sprinting—the frame rate and resolution are automatically adjusted. Specifically, fast motion scenarios are configured with a resolution of 1080p at 60 frames per second (fps) to ensure that motion details are captured without blur, whereas slow technical movement scenarios are switched to a resolution of 720p at 30 fps to balance data precision with device power consumption. Camera drivers are encapsulated through HAL, and real-time joint keypoint extraction and preprocessing are implemented using the MediaPipe Pose framework. This process outputs the three-dimensional coordinates of 25 key skeletal joints, with an extraction latency not exceeding 15 ms.

Inertial data acquisition is conducted using lightweight six-axis IMU sensors. The sampling rate is fixed at 100 Hz, enabling synchronous collection of three-axis acceleration and angular velocity data. To reduce noise in the raw data, a second-order Butterworth low-pass filter is applied during preprocessing, with the cutoff frequency set to 10 Hz. The filtering process is defined as follows:

$$y(n) = 2y(n-1) - y(n-2) + 0.0007x(n) + 0.0014x(n-1) + 0.0007x(n-2) \quad (2)$$

High-frequency interference generated during motion execution is effectively attenuated through the applied filtering process. Data synchronization is implemented using a timestamp calibration mechanism, by which temporal alignment between the camera and IMU sensors is achieved through the mobile terminal system clock. Initially, temporal offset calibration is performed for both types of sensors to obtain a fixed offset value, denoted as $\Delta t0$. During data acquisition, a system timestamp is appended to each visual frame and each inertial data packet, and dynamic calibration is performed according to the following equation:

$$t_{unified} = t_{raw} + \Delta t_0 + \Delta t_{dynamic} \quad (3)$$

where, $\Delta t_{dynamic}$ represents the real-time clock drift compensation term computed during operation. After synchronization, the temporal deviation between multimodal data streams is maintained within 1 ms, thereby ensuring spatiotemporal consistency across multimodal data.

Lightweight multimodal fusion algorithm. The multimodal fusion algorithm is designed using a three-stage architecture comprising preprocessing, feature extraction, and lightweight fusion. The primary objective is to reduce computational complexity while preserving fusion accuracy, thereby enabling coordinated processing between mobile terminals and cloud-based resources. During the data preprocessing stage, in addition to low-pass filtering of inertial signals, normalization is applied separately to visual key point coordinates and inertial data. Visual data are normalized using min–max scaling, mapping coordinate values to the [0, 1] interval, expressed as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4}$$

Inertial data are normalized using Z-score normalization, defined as:

$$X_{norm} = \frac{X - \mu}{\sigma} \tag{5}$$

where, μ denotes the mean and σ represents the standard deviation. This normalization process effectively eliminates dimensional inconsistencies.

During the feature extraction stage, a modality-specific extraction strategy is adopted. Visual features are derived from skeletal key point coordinates, from which 18-dimensional spatiotemporal features are extracted, including motion trajectory curvature, joint angle variation rates, and pose similarity of key frames. Inertial features are extracted from tri-axial acceleration and angular velocity signals, yielding 12-dimensional time-domain features such as mean, variance, peak value, and kurtosis. As a result, a 30-dimensional unimodal feature vector is constructed. To enhance feature discriminability, principal component analysis (PCA) is applied to unimodal features for dimensionality reduction. Principal components with a cumulative explained variance of no less than 95% are retained, resulting in a reduction of visual feature dimensionality to 10 and inertial feature dimensionality to 8.

The fusion model is implemented using an improved lightweight CNN–LSTM architecture, in which computational complexity is reduced through structural optimization. The CNN component consists of two convolutional layers followed by a single max-pooling layer, enabling the extraction of local spatial correlations from visual features. The LSTM component employs a single bidirectional LSTM layer with 64 hidden units to capture temporal dependencies in inertial features. To achieve lightweight deployment, channel pruning is applied to the convolutional layers, and 8-bit integer (INT8) quantization is applied to the fully connected layers. As a result, the number of model parameters is reduced by 42%, and computational complexity is decreased by 51% compared with the original CNN–LSTM architecture. During the fusion stage, a feature-level fusion strategy is employed. Visual features output from the CNN and inertial features output from the LSTM are concatenated and subsequently fed into two fully connected layers. A 20-dimensional fused feature vector is produced as the final output, providing core data support for subsequent skill assessment.

Sports skill assessment model. The sports skill assessment model is constructed based on the sports education curriculum framework and incorporates a multidimensional quantitative indicator system encompassing three core dimensions: movement standardization, execution speed, and stability. The movement standardization indicator is defined as the degree of feature matching between the action to be evaluated and a reference template, reflecting compliance with technical execution requirements. The execution speed indicator is quantified by the deviation

between the actual action completion time and a predefined standard time, representing movement efficiency. The stability indicator is quantified by the feature variance across multiple repetitions of the same action, thereby characterizing movement consistency. The relative importance of each indicator is determined using the analytic hierarchy process (AHP). Weights of 0.6, 0.2, and 0.2 are assigned to movement standardization, execution speed, and stability, respectively, ensuring that the assessment focus remains aligned with core instructional objectives in sports education.

The construction of reference action templates follows a "professional sample–based dynamic alignment" strategy. Standard action data are collected from ten national-level athletes, and temporal alignment across multiple samples is performed using the dynamic time warping (DTW) algorithm to generate a normalized reference template. The DTW distance metric is defined as:

$$D(i, j) = |x_i - y_j| + min\,(D(i-1, j), D(i, j-1)\, D(i-1, j-1)) \tag{6}$$

where, x_i denotes the feature of the sample to be aligned and y_j represents the corresponding feature of the reference sample. After alignment, the temporal length of the template is uniformly normalized to 100 time steps.

Similarity between the action to be evaluated and the template is computed using the cosine similarity metric, defined as:

$$sim = \frac{\vec{F}_{test} \cdot \vec{F}_{ref}}{|\vec{F}_{test}| \cdot |\vec{F}_{ref}|} \tag{7}$$

where, $\vec{F}_{test}$ denotes the fused feature vector of the action to be evaluated and $\vec{F}_{ref}$ represents the feature vector of the reference template. A comprehensive scoring mechanism is implemented using a weighted summation formulation:

$$S = 0.6 \times sim \times 100 + 0.2 \times (1 - \frac{|\Delta t|}{t_{ref}}) \times 100 + 0.2 \times (1 - \sigma) \times 100 \tag{8}$$

where, Δt denotes the deviation in action completion time, t_{ref} represents the standard completion time, and σ corresponds to the normalized variance of stability. Assessment outcomes are categorized into four performance levels: excellent (90–100), good (80–89), qualified (60–79), and unqualified (<60). In addition, the three key action nodes exhibiting the largest deviations from the template are automatically identified, and targeted improvement recommendations are generated accordingly.

Mobile interaction interface design. The design of the mobile interaction interface follows the core principles of simplicity, efficiency, and interference resistance, with full consideration given to the operational demands of dynamic sports education scenarios. A high-contrast color scheme is adopted, while key functional buttons are implemented with enlarged rounded designs. The font size is maintained at no less than 14 pt to ensure visual clarity under strong outdoor lighting conditions. In parallel, touch interaction logic is optimized to support gloved operation and accidental-touch filtering, thereby reducing operational errors caused by body movement during physical activity.

The implementation of core functional modules is centered on maintaining continuity within the instructional workflow. The data acquisition initiation module adopts a "single-action trigger with automatic configuration" mechanism. Once the "Start Assessment" command is activated, acquisition parameters corresponding

to the current sports activity are automatically matched, eliminating the need for manual configuration. During data acquisition, transmission status and remaining battery capacity are displayed in real time, and assessment progress is presented intuitively through a progress bar. The assessment result visualization module employs graphical presentation techniques. Feature curves of the action to be evaluated and the reference template are compared using line charts, while motion deficiency regions are highlighted via heat maps. Quantitative scores and descriptions of key deficiencies are displayed synchronously. The feedback information delivery module supports both local real-time notification and cloud-based synchronization. Assessment reports can be batch-delivered from the instructor interface to the student interface, while the student interface receives personalized improvement recommendations and standard action demonstration videos.

Differentiated interface layouts are adopted for the instructor interface and the student interface. Core functions of the instructor interface include assessment task creation, class management, and statistical data analysis. Filtering of assessment data by sports item, class, and time interval is supported, enabling the generation of skill-level distribution histograms and performance improvement trend curves. Statistical reports can be exported in Excel format. Core functions of the student interface include individual skill practice, historical assessment record retrieval, and improvement recommendation review. Key action demonstration videos can be bookmarked, and targeted practice content is automatically recommended based on historical performance data. A responsive interface design is employed throughout the system. Cross-platform compatibility is achieved using the Flutter framework, ensuring consistent interaction logic and optimized visual presentation across mobile devices with varying screen sizes.

2.3 Prototype development

Prototype development is implemented using a cross-platform technology stack to ensure system compatibility and stability across mainstream mobile terminals. The system principle block diagram is illustrated in Figure 2. The development environment is configured below. Kotlin and Swift are used for front-end development, while the Flutter framework is employed to enable cross-platform reuse of core business logic. For back-end development, Python is adopted, and RESTful APIs are constructed based on the Django framework. Development tools include Android Studio Hedgehog, Xcode 15.0, and PyCharm Professional 2023.2, with Git utilized for version control. For the hardware platform, compatibility testing is conducted using two categories of mainstream mobile terminals. Smartphone platforms include Xiaomi 13 and iPhone 14, while a self-developed lightweight IMU wearable sensor module is employed.

The core functional demonstration of the prototype system covers the complete workflow of data acquisition, transmission, processing, assessment, and feedback. An assessment task for basketball free throws is created via the instructor interface, where the number of participants and the standard completion time are configured and subsequently distributed to the student interface. After task reception, data acquisition is initiated by activating the “Start Acquisition” function. The smartphone camera and wearable sensors are automatically triggered, enabling synchronized collection of visual and inertial data associated with the shooting action. Following compression and encryption, the acquired data are transmitted to the cloud via a 5G network. Cloud-based processing is performed by invoking the lightweight fusion

algorithm and the assessment model, through which quantitative scores and motion deficiency reports are generated. Assessment results are returned to the student interface in real time, allowing motion deficiency heat maps and targeted improvement recommendations to be reviewed. In parallel, class-level assessment data are synchronized to the instructor interface, where skill statistics reports are generated.

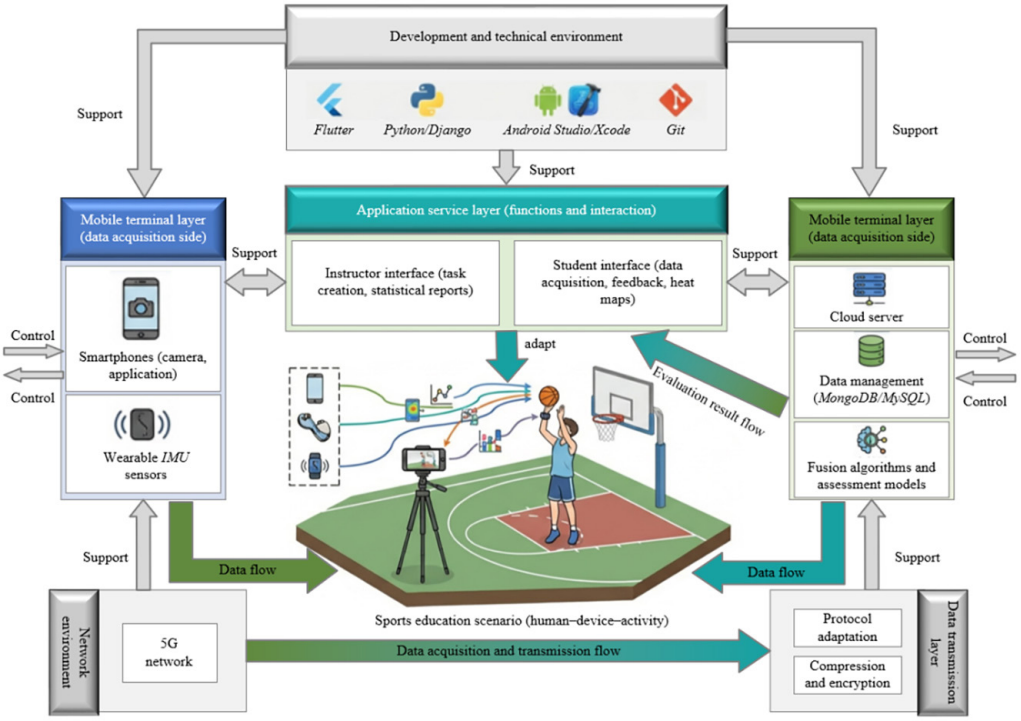

Fig. 2. Prototype development and technical environment

The hardware configuration and software module composition of the system are explicitly defined. The hardware part consists of smartphones, wearable IMU sensors, and cloud servers, with data interaction achieved through Bluetooth 5.2 and 5G networks. The software architecture is divided into a front-end application module, a data transmission module, a cloud-based processing module, and a data management module. The front-end application module comprises a multimodal data acquisition submodule, an interaction interface submodule, and a local caching submodule. The data transmission module includes protocol adaptation, compression and encryption, and traffic control submodules. The cloud-based processing module incorporates a fusion algorithm submodule, an assessment model submodule, and a task scheduling submodule. The data management module consists of structured data storage, unstructured data storage, and data backup submodules. Data interaction among all software modules is achieved through standardized JSON-based interfaces, thereby ensuring system scalability and maintainability.

3 EXPERIMENTS AND RESULTS

The results of system technical performance testing under different mobile scenarios are presented in Table 1. It is observed that the system satisfies all predefined performance criteria across all tested scenarios, demonstrating strong adaptability to diverse mobile sports teaching environments.

Table 1. System technical performance under different mobile scenarios

Test Scenario	Indoor Static	Indoor Dynamic	Outdoor Static	Outdoor Dynamic
Visual acquisition error (%)	1.2 ± 0.3	1.8 ± 0.4	1.5 ± 0.4	2.1 ± 0.5
Visual frame-rate stability (fps)	30.0 ± 0.5	29.5 ± 1.2	29.8 ± 0.7	28.9 ± 1.5
Inertial acquisition error (%)	0.8 ± 0.2	1.3 ± 0.3	1.0 ± 0.2	1.6 ± 0.4
Inertial sampling-rate stability (Hz)	100.0 ± 0.8	99.5 ± 1.2	99.8 ± 0.9	99.2 ± 1.5
Transmission latency (5G, ms)	28 ± 4	32 ± 5	30 ± 5	36 ± 6
Transmission latency (Wi-Fi 6, ms)	35 ± 5	41 ± 6	38 ± 6	45 ± 7
Continuous operating time (h)	9.2	8.7	8.5	8.1
Average power consumption (Wh)	6.8	7.5	7.2	8.1

Visual acquisition error is minimized under indoor static conditions and reaches its maximum under outdoor dynamic conditions; however, values remain below the predefined threshold in all scenarios. Visual frame rates are consistently maintained above 28.9 fps, ensuring complete capture of motion details. Inertial acquisition error remains below 1.6% across all scenarios, while sampling-rate stability is maintained at or above 99.2 Hz, indicating that the second-order Butterworth low-pass filter effectively suppresses high-frequency motion interference. With respect to data transmission, average latency ranges from 28 to 36 ms under 5G conditions and from 35 to 45 ms under Wi-Fi 6 conditions, both satisfying the real-time requirement of ≤50 ms. The lower latency observed under 5G conditions indicates superior suitability for dynamic instructional scenarios. Continuous operating time is maintained at or above 8.1 h across all scenarios, with average power consumption not exceeding 8.1 Wh, demonstrating that the system is capable of supporting complete teaching sessions.

Table 2. Comparison of assessment performance across different algorithms

Algorithm Type	Single Visual Modality (CNN)	Single Inertial Modality (LSTM)	Conventional Fusion (CNN–LSTM)	Proposed Lightweight CNN–LSTM
Basketball free-throw accuracy (%)	82.3 ± 2.1	78.6 ± 2.3	89.5 ± 1.5	93.2 ± 1.2
Standing long jump accuracy (%)	79.5 ± 2.4	83.2 ± 2.1	88.6 ± 1.6	92.8 ± 1.3
Basic Tai Chi movement accuracy (%)	80.1 ± 2.2	81.5 ± 2.0	89.2 ± 1.4	94.1 ± 1.1
Single-sample computation time (ms)	35 ± 3	28 ± 2	82 ± 5	42 ± 3
Model parameter size (MB)	8.6	4.2	22.4	6.8
Computational cost (GFLOPs)	1.2	0.8	3.5	1.5

The comparative assessment performance of the four algorithms is summarized in Table 2. It is evident that the proposed lightweight CNN–LSTM fusion algorithm achieves superior overall performance compared with the baseline methods.

With respect to assessment accuracy, performance exceeding 92% is achieved by the lightweight CNN–LSTM algorithm across all three sports skills. Improvements of 10.9–14.0 percentage points are observed relative to the single visual modality model, while gains of 9.6–15.5 percentage points are achieved compared with the single inertial modality model. When compared with the conventional CNN–LSTM fusion approach, accuracy improvements of 3.7–4.9 percentage points are obtained, thereby confirming the effectiveness of complementary multimodal data fusion. In terms of computational efficiency, the average single-sample processing time is limited to 42 ms, representing a reduction of 48.8% relative to the conventional fusion algorithm. Furthermore, the model parameter size and computational cost are reduced by 69.6% and 57.1%, respectively. These results demonstrate the effectiveness of the applied lightweight optimization strategies—specifically channel pruning and INT8 quantization.

Table 3. Comparison of assessment results across different evaluation methods

Evaluation Method	Proposed System	Professional Vicon System	Traditional Instructor Subjective Assessment
Basketball free-throw mean score	85.6 ± 4.2	86.3 ± 3.8	84.2 ± 5.1
Standing long jump mean score	83.8 ± 3.9	84.5 ± 3.6	82.3 ± 4.8
Basic Tai Chi movement mean score	86.2 ± 4.5	87.1 ± 4.1	84.7 ± 5.2
Correlation with Vicon (r)	0.94 ± 0.03	–	0.82 ± 0.05
Assessment consistency (Kappa)	0.87 ± 0.04	–	0.73 ± 0.06

Table 4. User satisfaction ratings from instructors and students (maximum score = 5)

Evaluation Dimension	Instructor Rating (n = 8)	Student Rating (n = 45)	Overall Mean Rating
Operational convenience	4.6 ± 0.3	4.5 ± 0.4	4.5 ± 0.4
Result accuracy	4.7 ± 0.2	4.4 ± 0.5	4.5 ± 0.4
Feedback timeliness	4.8 ± 0.2	4.6 ± 0.4	4.7 ± 0.3
Overall practicality	4.6 ± 0.3	4.5 ± 0.4	4.5 ± 0.4

The results of the teaching validation experiments are presented in Tables 3 and 4. A high level of consistency is observed between the assessment outcomes produced by the proposed system and those obtained using the professional Vicon motion capture system.

As shown in Table 3, the mean score difference between the proposed system and the Vicon system does not exceed 0.7 points across all evaluated sports skills. A strong correlation with the Vicon system is achieved (r = 0.94), and assessment consistency reaches a Kappa coefficient of 0.87. Both metrics are substantially higher than those obtained using traditional instructor subjective assessment, indicating that high-precision, standardized skill evaluation can be achieved without reliance on specialized motion capture equipment, while effectively mitigating subjective bias. The user satisfaction survey results presented in Table 4 indicate an overall mean score of 4.5 out of 5. Among all evaluation dimensions, feedback timeliness receives the highest rating. Both operational convenience and result accuracy

achieve mean scores of no less than 4.4, confirming strong adaptability to sports teaching scenarios and high practical applicability of the system.

The multidimensional comparison of algorithmic computational efficiency is illustrated in Figure 3. A radar chart is employed to visualize the comprehensive performance differences among the four algorithms across five core efficiency-related indicators. The conventional fusion algorithm exhibits inferior performance across all dimensions. Owing to its large model parameter size and high computational complexity, excessive memory consumption and power usage are induced, rendering the algorithm poorly suited to the resource constraints of mobile terminals. This limitation represents a primary bottleneck that has hindered the practical deployment of existing multimodal fusion algorithms in mobile sports teaching scenarios. The single inertial modality algorithm demonstrates the highest efficiency across computational dimensions; however, reliance solely on inertial data prevents accurate representation of fine-grained movement posture details, resulting in inherent deficiencies in assessment dimensionality. The single visual modality algorithm achieves moderate computational efficiency, yet its stability is compromised in dynamic motion assessment due to susceptibility to environmental interference, thereby limiting robustness in real-world teaching contexts.

The proposed lightweight CNN–LSTM algorithm achieves a balanced trade-off between efficiency and accuracy through the application of channel pruning and INT8 quantization strategies. As illustrated by the radar chart, performance across all efficiency-related dimensions is maintained within a favorable range. Although the single-sample computation time is marginally higher than that of the single inertial modality algorithm, the lightweight CNN–LSTM algorithm demonstrates substantial advantages over the conventional fusion algorithm in terms of model parameter size, computational cost, memory usage, and mobile terminal power consumption, with reductions of 69.6%, 57.1%, 52.5%, and 34.4%, respectively. These findings indicate that the lightweight design strategy effectively overcomes the resource constraint bottlenecks associated with traditional fusion algorithms. Consequently, compatibility with the collaborative processing architecture between mobile terminals and cloud-based platforms is achieved, providing essential technical support for stable system operation in dynamic sports education scenarios.

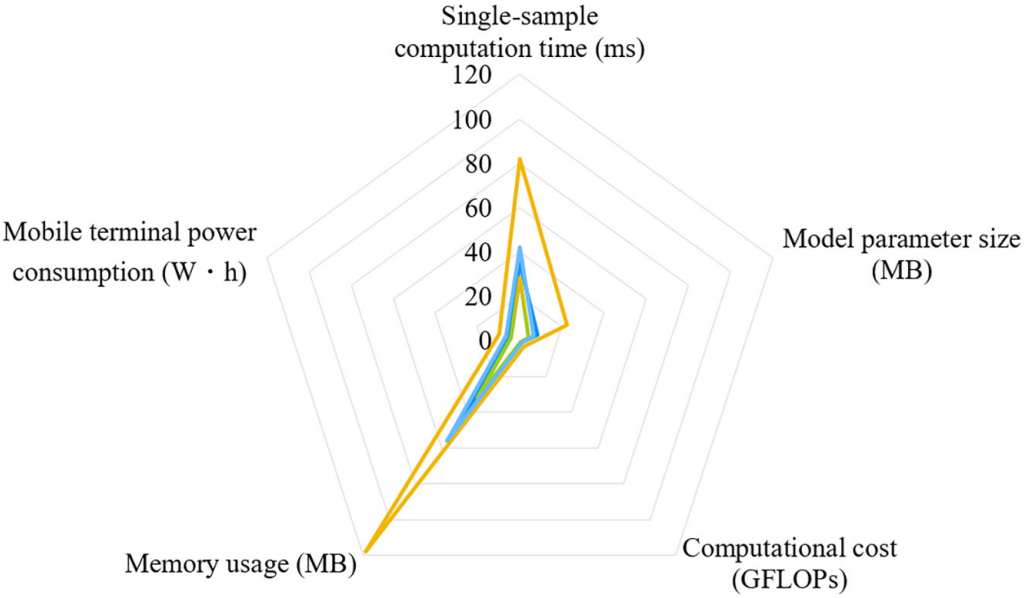

Fig. 3. Radar chart comparison of algorithmic computational efficiency

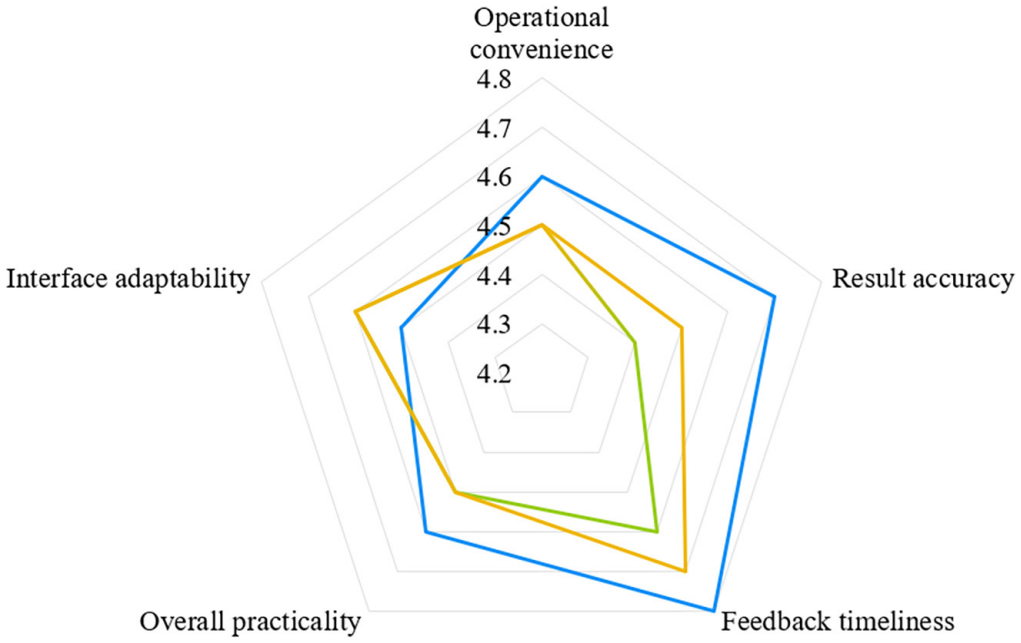

Fig. 4. Radar chart of user satisfaction ratings

The multidimensional evaluation results of user satisfaction are illustrated in Figure 4. The radar chart clearly presents the score distributions and overall differences between instructors and students across five core evaluation dimensions. Highly similar radar profiles are observed for the two user groups, indicating balanced system adaptability across different user populations. Overall, all evaluation dimensions achieve scores of no less than 4.4, with an overall mean score of 4.5, thereby confirming the instructional practicality and user friendliness of the system.

The feedback timeliness dimension receives the highest rating. This outcome is closely aligned with the system's real-time design objectives. Through the use of adaptive transmission protocols and lightweight fusion algorithms, feedback latency is controlled within 100 ms, enabling timely instructional guidance during sports teaching activities and effectively addressing the delayed feedback limitation associated with traditional assessment approaches. The result accuracy dimension attains a mean score of 4.5, with instructor ratings slightly exceeding those of students. This finding indicates that the standardization and reliability of the assessment outcomes are well recognized by teaching practitioners and that the system is capable of providing precise data support for instructional optimization. Scores for operational convenience and interface adaptability both exceed 4.5, demonstrating that the "single-action trigger with automatic configuration" interaction design and responsive interface layout are well suited to the operational demands of dynamic sports teaching scenarios while also reducing user learning and usage costs. The overall practicality dimension reaches a mean score of 4.5, indicating that the system effectively addresses key limitations of traditional sports skill assessment, including subjectivity and insufficient standardization. These results further demonstrate that a deployable technical solution is provided to support the intelligent transformation of sports education.

4 CONCLUSION

In response to the limitations of traditional sports skill assessment—namely subjectivity, delayed feedback, and single-dimensional evaluation—a mobile multimodal

interactive skill assessment system for sports education was designed and implemented. A four-layer distributed architecture comprising the mobile terminal layer, data transmission layer, cloud-based processing layer, and application service layer was established. By integrating multimodal data acquisition, lightweight fusion algorithms, and a precise assessment model, real-time, multidimensional, and standardized evaluation of sports skills was achieved.

Experimental validation demonstrates substantial technical advantages of the proposed system. The lightweight architectural design is compatible with mainstream mobile terminals and supports stable operation in complex mobile teaching scenarios while satisfying low-power and high real-time requirements. Through structural optimization, the multimodal data fusion algorithm achieves a favorable balance between accuracy and efficiency, with skill assessment accuracy exceeding 92%, significantly outperforming single-modality and conventional fusion approaches. Overall, strong adaptability to mobile scenarios is exhibited, and high consistency with professional motion capture systems is achieved, confirming the reliability and practical applicability of the system.

5 REFERENCES

[1] J. Jain, M. Kaur, and K. Sood, "Human touch in digital learning communities: Teacher's role in shaping learner satisfaction on EdTech platforms," *Central Community Development Journal*, vol. 6, no. 1, pp. 26–39, 2026. https://doi.org/10.56578/ccdj060103

[2] C. B. Xuan, L. H. Viet, and P. V. Thang, "Digital transformation in Vietnamese SMEs: Challenges, opportunities, and strategic implications," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 20, no. 7, pp. 84–96, 2026. https://doi.org/10.3991/ijim.v20i07.61093

[3] S. N. M. Roslan, K. Gohain, A. M. A. A. Mustafa, M. M. Ismail, and V. V. Kumaran, "Designing affordable urban ecosystems: A quantitative model to enhance the quality of life for the urban poor in Malaysia through employment, housing, and digital access," *Challenges in Sustainability*, vol. 13, no. 1, pp. 18–34, 2025. https://doi.org/10.56578/cis130102

[4] G. A. Meek and D. Behets, "Physical education teachers' concerns towards teaching," *Teaching and Teacher Education*, vol. 15, no. 5, pp. 497–505, 1999. https://doi.org/10.1016/S0742-051X(98)00061-4

[5] A. Huda, W. Febrianti, Firdaus, Y. Hendriyani, B. R. Fajri, and M. Sukmawati, "Designing digital modules in project-based learning-based printing graphic design subjects at SMK N 1 Koto Baru Dharmasraya," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 18, pp. 94–111, 2024. https://doi.org/10.3991/ijim.v18i18.50551

[6] J. Yao and Y. Li, "Youth sports special skills' training and evaluation system based on machine learning," *Mobile Information Systems*, vol. 2022, no. 1, p. 6082280, 2022. https://doi.org/10.1155/2022/6082280

[7] M. A. W. Alsoub, "Scientific and practical competencies, communication skills, evaluation and ability to work among the graduates of sports rehabilitation," *Revista iberoamericana de psicología del ejercicio y el deporte*, vol. 19, no. 1, pp. 38–44, 2024.

[8] B. Mak, R. C. Nickerson, and J. Sim, "Mobile technology dependence and mobile technostress," *International Journal of Innovation and Technology Management*, vol. 15, no. 4, p. 1850039, 2018. https://doi.org/10.1142/S0219877018500396

[9] H. Pieterse, M. Olivier, and R. Van Heerden, "Smartphone data evaluation model: Identifying authentic smartphone data," *Digital Investigation*, vol. 24, pp. 11–24, 2018. https://doi.org/10.1016/j.diin.2018.01.017

[10] M. T. Hamid and M. Abid, "Decision support system for mobile phone selection utilizing fuzzy hypersoft sets and machine learning," *Journal of Intelligent Management Decision*, vol. 3, no. 2, pp. 104–115, 2024. https://doi.org/10.56578/jimd030204

[11] Y. Zhang, "Application of Speech Recognition Technology Based on Multimodal Information in Human-Computer Interaction," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 9, pp. 101–111, 2024. https://doi.org/10.14569/IJACSA.2024.0150911

[12] G. Schiavone, D. Formica, F. Taffoni, D. Campolo, E. Guglielmelli, and F. Keller, "Multimodal ecological technology: From child's social behavior assessment to child-robot interaction improvement," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 69–81, 2011. https://doi.org/10.1007/s12369-010-0080-9

[13] Q. Wan, X. Yang, and G. Chen, "New scheduling algorithm for mobile teaching cloud resource push," *International Journal of Emerging Technologies in Learning*, vol. 13, no. 7, pp. 17–29, 2018. https://doi.org/10.3991/ijet.v13i07.8803

6 AUTHORS

Dan Lv graduated from Beijing Normal University in 2013 and currently work at the School of Physical Education, Inner Mongolia University, and serves as a Lecturer (E-mail: 111988003@imu.edu.cn).

Huijun Li is a Professor with a research focus on physical education and sports training (E-mail: lhj111971110@163.com).