http://jhmi.sums.ac.ir
Health Man & Info Sci
2026, 13(2), 139–153

HMIS
Health Management and Information Science

Original Article

A deep learning framework for evaluating dynamic sports movements via video-based motion analysis for digital health

Khosro Rezaee[1*], Fatemeh Monshizadeh[1]
[1]Department of Biomedical Engineering, Meybod University, Meybod, Iran

Abstract:

Introduction: Video-based assessment of exercise technique can support coaching, injury prevention, and remote rehabilitation, yet many methods stop at action recognition or require laboratory motion capture. To develop and evaluate a deep-learning framework that classifies execution quality as correct vs. incorrect form from instructional videos.

Methods: We compiled 270 YouTube coaching videos spanning ten exercises (~20,500 labeled frames after removing ~15% non-movement content). Clip-level technique labels (correct/incorrect) were assigned based on coaching/biomechanical criteria and propagated to sampled frames. A marker-less pose-estimation model (name/version reported) produced 2D key points and kinematic descriptors (e.g., joint angles and velocities). These signals were encoded with a ResNet-50 backbone, and a multi-head attention Transformer modeled temporal dependencies. Training used a video-stratified 80/20 split, with 5-fold cross-validation on the training portion only for model selection; class-weighted loss mitigated imbalance, and evaluation reports per-class precision/recall/F1 and confusion matrices. We also define a biomechanical complexity index (0–1) combining joints engaged, angular sensitivity, range of motion, and balance demand to relate movement difficulty to performance.

Results: The full CNN–attention–Transformer achieved 97.56% accuracy and F1=0.956 on the held-out test set. Per-exercise performance remained high for low-to-intermediate complexity movements, while the most challenging exercises showed reduced F1—for example, deadlift and crunch exhibited the lowest scores (≈0.88–0.90), yielding an overall F1 range of 0.875–0.945 across exercises.

Conclusion: This framework enables multi-exercise, video-based technique assessment (correct vs. incorrect) from consumer footage. Attention-weight patterns and error-case analyses provide practical insight into model decisions for intelligent coaching and tele-rehabilitation.

Keywords: Deep learning, Transformer, Video-based motion analysis, Sports exercise quality evaluation, Markerless human pose estimation.

Article History:
Received: 10 November 2025
Accepted: 22 February 2026

Please cite this paper as:
Rezaee Kh, Monshizadeh F. A deep learning framework for evaluating dynamic sports movements via video-based motion analysis for digital health. Health Man & Info Sci. 2026; 13(2): 139-153. doi: 10.30476/JHMI.2026.109616.1334

***Correspondence to:**
Khosro Rezaee, Yahyazadeh Blvd., Khorramshahr Blvd., Meybod, Iran.
Tel: +98-35-32357500 | Fax: +98-35-32353004 | Email: kh.rezaee@meybod.ac.ir

Introduction

Sports exercises are coordinated, multi-joint movements in which execution quality is directly linked to tissue loading and injury risk, especially when technique faults persist across repetitions or training sessions (1,2). In both performance and rehabilitation contexts, even small deviations in alignment, range of motion, or trunk stability can shift joint moments and muscle recruitment patterns, increasing cumulative overload and, in some cases, the likelihood of acute injury events (2). Consequently, there is a growing need for consistent and scalable assessment of movement quality that operates outside controlled laboratory environments and can be integrated into routine training, tele-coaching, and remote digital-health workflows.

Classical biomechanics provides precise descriptions of kinematics and kinetics, but marker-based motion capture systems, force plates, and advanced imaging remain costly and largely confined to laboratory settings (1,3). To reduce these barriers, recent studies increasingly rely on consumer-grade RGB video, combined with markerless pose estimation and learning-based models, to analyze human movement using

ordinary cameras (4,5). However, much of the existing literature still focuses on recognizing which action is performed, rather than evaluating how well it is executed, and many methods are validated on datasets recorded under controlled conditions with fixed viewpoints and limited real-world noise (e.g., clutter, occlusions, and variable camera placement) (6-9). This limits their robustness and practical deployment in realistic training environments characterized by clutter, occlusions, and variable camera placement.

From a methodological perspective, advances in movement analysis highlight the importance of combining pose estimation, signal stabilization, and temporal modeling. Multiview toolkits can reconstruct three-dimensional skeleton trajectories from synchronized cameras, while filtering approaches such as Kalman-type filters can reduce measurement noise and improve temporal consistency for downstream analysis (10,11). Rather than replacing physics-based biomechanics, learning-based methods have also been explored as surrogate models that approximate specific finite-element simulation outputs, enabling faster estimation when full biomechanical simulations are computationally impractical (12). In parallel, deep neural networks have demonstrated strong capability in extracting structure from high-dimensional kinematic data (13), and convolutional neural networks (CNNs) remain effective at learning spatial representations relevant to body posture and anatomical configuration (14,15).

Assessing movement quality, however, is inherently a temporal problem. Many execution errors are defined not only by static posture but by how joint configurations evolve, including timing, coordination, and control across movement phases. Attention mechanisms and temporal models can emphasize informative regions and frames while suppressing noise (16). Transformer-based architectures further enable explicit modeling of long-range temporal dependencies and complex motion patterns, complementing CNN-based spatial feature extraction and offering advantages over purely recurrent temporal models in variable-length sequences (17,18). Prior work has shown that deep temporal-learning pipelines combining CNN feature extraction with sequence models are effective for activity recognition and sports-teaching video analysis, supporting the feasibility of learning-based temporal analysis in unconstrained video settings (19,20).

Despite these advances, there remains a clear need for frameworks that (1) rely solely on readily available consumer video, (2) exploit pose-derived descriptors to reduce sensitivity to background appearance, (3) jointly model spatial configuration and temporal evolution of movement, and (4) are explicitly designed to assess exercise execution quality rather than action identity (4,21). In particular, the majority of existing studies prioritize action recognition (i.e., detecting what movement is being performed) and are evaluated in relatively controlled settings, whereas comparatively fewer works address how well the movement is executed across diverse exercises under realistic, noisy conditions (variable viewpoints, clutter, and occlusions).

Unlike most prior work centered on action identification, we explicitly target execution-quality assessment (correct vs. incorrect) from consumer videos by coupling pose-derived descriptors with CNN-based spatial encoding and Transformer-based temporal modeling. Addressing this gap, the present study introduces a deep-learning framework that combines a CNN backbone with multi-head attention and a Transformer encoder to evaluate dynamic body movement in workout videos. In the proposed pipeline, markerless pose estimation produces skeleton-based descriptors from RGB footage; these descriptors include 2D joint coordinates and simple kinematic features (e.g., joint angles and velocities) computed from keypoint trajectories. The CNN captures local spatial relationships among joints and body segments; attention mechanisms prioritize biomechanically informative and error-prone regions; and the Transformer models' long-range temporal dependencies across movement phases to classify each exercise instance as correctly or incorrectly performed (16–19,22). Compared with recurrent units, the Transformer's self-attention can model longer-range dependencies across repetitions and movement phases with fewer constraints on sequence length, which is advantageous for unconstrained consumer videos. By aligning model design with biomechanical considerations and recent advances in temporal representation learning, this framework aims to support intelligent coaching, injury-risk awareness, and rehabilitation support outside laboratory settings.

Methods

We designed a multi-stage deep-learning pipeline for automatic assessment of exercise

execution quality in unconstrained videos. The dataset consists of 270 instructional YouTube videos across 10 exercises, yielding approximately 20,500 refined frames after removing non-movement segments (≈15%) and assigning binary labels (correct vs. incorrect) as described in Section 2.1. Each video was temporally trimmed and sampled into frame sequences. Markerless pose estimation was performed using MediaPipe Pose (23) (BlazePose; MediaPipe v0.10.14) in static-image mode, extracting 33 per-frame 2D landmarks with visibility scores; detections used minimum detection and tracking confidences of 0.5. Extracted keypoints were mapped to image coordinates and used to compute pose-derived kinematic descriptors (e.g., joint angles and joint velocities) from landmark trajectories. Spatial features were extracted using a ResNet-50 backbone, and a multi-head attention Transformer encoder modeled long-range temporal dependencies across movement phases to classify each exercise instance as correctly or incorrectly performed.

The model outputs a binary probability score (incorrect vs. correct). Data were split using an 80/20 video-stratified train/test split, and five-fold cross-validation on the training portion was used for model selection, followed by evaluation on the held-out test set. An overview of the complete processing workflow is presented in Figure 1.

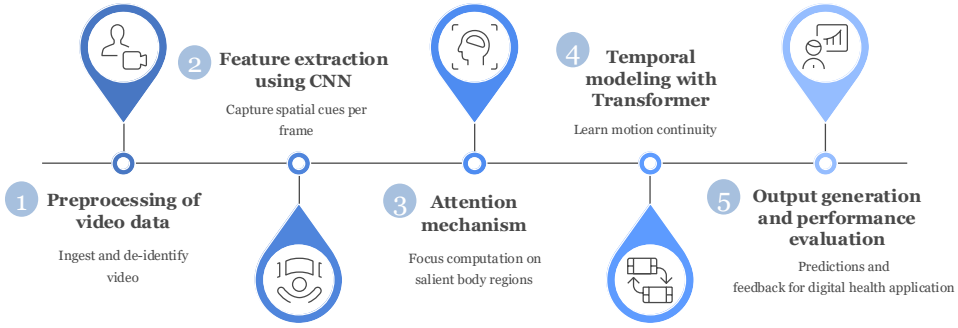

Fig. 1. Overview of the proposed method.

Datasets

We curated a video dataset to train and evaluate the proposed framework for binary assessment of exercise execution quality (correct vs. incorrect). Data were collected through systematic web exploration, primarily from publicly available instructional YouTube videos, supplemented by other open online educational sources. The selection emphasized diversity in recording environments, camera viewpoints, performers, and exercise styles, reflecting realistic consumer-grade conditions rather than controlled laboratory settings. A statistical overview of the dataset is summarized in Table 1, with additional details provided in the Supplementary Materials.

Exercise categories

The dataset covers ten strength and conditioning exercises selected for biomechanical diversity and relevance to performance and injury risk: squat, plank, push-up, lunge (including reverse-lunge demonstrations), deadlift, crunch (abdominal exercise), bench press, pull-up, TRX chest fly, and jump squat. These movements span both relatively stable kinematics (e.g., plank) and exercises with higher sensitivity to trunk control, joint coordination, and timing (e.g., deadlift and jump squat).

Table 1. Summary of the curated exercise video dataset.

Description	**Measure**
Total number of videos	270
Number of exercise categories	10
Exercise types	Squat, plank, push-up, lunge, deadlift, crunch, bench press, pull-up, TRX chest fly, jump squat
Labeling granularity	Manual clip-level labels (correct/incorrect), propagated to frames
Effective frame sampling rate	~10 fps
Average frames per clip	~90
Total raw frames extracted	~24,000

Final refined frames	~20,500
Number of classes	2 (correct/incorrect)
Primary data sources	Public web videos, primarily YouTube instructional content

Labeling procedure

Execution-quality labels were assigned manually at the clip level as *correct* or *incorrect*, based on visual inspection guided by common coaching cues and biomechanical principles (e.g., joint alignment, trunk stability, and movement control). Labels were not derived from video titles, metadata, or automated heuristics. After temporal sampling, frame-level labels were inherited from the corresponding clip-level annotation. This binary labeling strategy aligns with the instructional nature of the source material, in which correct demonstrations are typically followed by demonstrations with common technique faults.

Frame extraction and temporal organization

Videos were decoded and uniformly sub-sampled at an effective rate of approximately 10 frames per second, corresponding to retaining roughly one frame out of every two to three original frames. This yielded an average of ~90 frames per exercise clip and approximately 24,000 raw frames across the dataset. Frames were temporally grouped into short, ordered segments representing individual exercise instances, enabling sequence-based modeling of movement quality. Precise repetition boundaries were not manually annotated; instead, segment boundaries were determined automatically based on motion continuity, reflecting realistic deployment scenarios where fine-grained repetition labels are unavailable.

Dataset refinement and statistics

Non-informative segments (e.g., verbal explanations, pauses, or frames without visible exercise motion) were removed during preprocessing, resulting in a refined set of approximately 20,500 frames used for model training and evaluation. Details of the frame-filtering procedure and the 15% content reduction are provided in the Supplementary Materials to preserve methodological transparency while keeping the main text concise.

Class balance and variability

Because instructional videos typically demonstrate both correct and incorrect executions, the dataset is approximately balanced at the clip level, with minor deviations at the frame level due to differences in segment duration. All exercise categories contain samples from both classes. To mitigate potential bias, performance is reported using per-class precision, recall, and F1-score in addition to overall accuracy.

Distributed computing platform

All experiments were executed on a single workstation rather than a distributed computing cluster. The computing environment consisted of an ASUS TUF A17 (2023) laptop equipped with an AMD Ryzen™ 7040-series multi-core processor, 32 GB system RAM, and an NVIDIA® RTX™ 4070 Laptop GPU, running Windows 11. This clarification replaces the earlier description of a distributed platform with worker nodes and a central coordinator, which does not apply to the present implementation.

The full pipeline was implemented in Python. Video decoding, frame sampling, and preprocessing were performed using standard scientific and computer-vision libraries (e.g., OpenCV and NumPy). Markerless pose estimation was carried out using MediaPipe Pose (BlazePose) in static-image mode, which outputs 33 per-frame 2D body landmarks together with per-landmark visibility (confidence) scores. Extracted pose information was written to disk as frame-aligned tabular files containing landmark coordinates and visibility values, enabling reproducible downstream processing and feature computation. Minimum detection and tracking confidence thresholds of 0.5 were used during pose extraction, consistent with the extraction configuration.

Model training and inference were implemented using PyTorch, with GPU acceleration enabled for all deep-learning components. Standard data loaders were used to stream preprocessed samples from disk into the training loop, avoiding repeated video decoding and pose extraction.

The term *training shards* refers to local, file-based packaging of preprocessed data on a single machine, rather than distributed storage. Each

shard groups temporally ordered frame sequences, their associated pose-derived descriptors (e.g., joint positions, angles, and velocities), and binary execution-quality labels. Shards are stored on local disk and loaded sequentially during training, improving I/O efficiency and ensuring consistent reuse of preprocessed samples across experiments.

Noise-robust filtering denotes deterministic pose-quality control operations applied before feature construction to mitigate errors introduced by compression artifacts, motion blur, occlusions, and partial out-of-frame joints common in consumer videos. Specifically, landmark visibility scores are used to identify unreliable detections; low-confidence frames are down-weighted or excluded; short gaps caused by occlusion are repaired via temporal interpolation; jitter is reduced using lightweight temporal smoothing; and implausible inter-frame landmark jumps (e.g., unrealistically high joint velocities) are treated as outliers and corrected through local interpolation. These steps stabilize kinematic descriptors before temporal modeling.

Because the pipeline runs on a single machine, there is no distributed scheduler, parameter server, or inter-node communication. Coordination occurs within a single training process (with optional multi-worker data loading), and concerns such as network bottlenecks or single points of failure across nodes do not apply. Standard checkpointing is used to resume training in the event of interruption.

Attention-based CNN–Transformer model

Figure 1 depicts the overall pipeline. The proposed system performs sequence-based execution-quality assessment: each input sample is an ordered clip of consecutive frames, not a single image, so that both spatial posture cues and temporal dynamics are modeled explicitly.

Data summary, labeling, and partitioning

We curated 270 publicly available YouTube coaching videos across ten exercises and retained ~20,500 labeled frames after preprocessing. Videos/clips are assigned a binary execution-quality label (correct vs. incorrect) following the labeling protocol in Section 2.1. To avoid leakage, data are split at the video level (no frames/clips from the same source video can appear in multiple splits). We use an 80/20 video-stratified train/test split. Hyperparameters are selected via 5-fold cross-validation performed only on the training portion, and the held-out test set is used once for final reporting.

Motion segmentation and clip construction

Instructional videos contain pauses and explanations; therefore, we first isolate motion-bearing intervals using a simple frame-difference motion filter and remove non-exercise segments. From each retained interval, we uniformly sample clips of T = 32 consecutive frames at an effective rate of 10 fps (i.e., selecting every 3rd frame from 30 fps source videos). Longer intervals yield multiple windows; shorter ones are zero-padded, and a temporal mask prevents padded positions from contributing to Transformer self-attention. Training augmentations are limited to lightweight photometric/geometry changes (brightness/contrast, small rotations, and random cropping).

Pose estimation and kinematic descriptors (MediaPipe)

Markerless pose estimation is performed with MediaPipe Pose [23] (BlazePose; v0.10.14) in static-image mode, yielding 33 2D landmarks per frame plus per-landmark visibility scores. Landmarks are normalized for scale/camera variability. From landmark trajectories, we compute biomechanically interpretable descriptors: normalized joint positions, inter-joint angles, and first-order joint velocities via finite differences across consecutive frames. Crucially, keypoints/kinematics are not predicted by the CNN: they are obtained from MediaPipe and analytical post-processing only.

Robustness to low-quality video and noisy keypoints

To address motion blur, compression artifacts, occlusions, and out-of-frame joints, we apply pose quality control before feature fusion: (1) visibility-based exclusion of unreliable frames/joints, (2) short-gap temporal interpolation for occlusions, (3) lightweight temporal smoothing to reduce jitter, and (4) outlier correction for implausible inter-frame landmark jumps and unrealistically high velocities. This stabilizes angles/velocities under non-ideal recording conditions.

Spatial encoding, pose-guided attention, and fusion (ResNet-50)

Each frame is resized to 224×224×3 and encoded by an ImageNet-pretrained ResNet-50 to obtain a visual embedding c_t. In parallel, the

pose/kinematic vector $p_t$is computed as above. We then apply a pose/joint-level multi-head attention module over p_t to reweight joints/segments (e.g., knee–hip chain in squats; shoulder–hip alignment in plaks), producing $\tilde{p}_t$. This attention operates on pose descriptors (not pixel-space "regions"). The final per-frame token is formed by concatenation and linear projection:

$$F_t = \text{Proj}([c_t \parallel \tilde{p}_t]) \quad (1)$$

This design ensures biomechanical descriptors are computed once (from MediaPipe) and fused with CNN features, eliminating any interpretation of "double feature extraction" or an internal regression branch for keypoints.

Temporal modeling and output definition

The token sequence $\{F_t\}_{t=1}^{T}$is processed by a Transformer encoder with positional encoding and padding masks. Temporal self-attention captures dependencies across movement phases under variable pacing, without the sequential constraints of RNNs; accordingly, no bidirectional recurrent unit (LSTM/GRU) is used in the proposed model. The network outputs a single probability for incorrect vs. correct execution; no explicit sub-scores (range of motion, speed, coordination) are separately estimated—these factors are implicitly represented in the learned spatiotemporal embedding.

Evaluation protocol and ablation fairness

All ablation variants (CNN-only, CNN + pose attention, CNN + Transformer, and the full CNN–pose-attention–Transformer) are trained with the same video-level split, cross-validation procedure, and metrics, and are evaluated on the same held-out independent test set (accuracy, precision, recall, F1-score, and confusion matrices).

Results

The attention-based CNN–Transformer model was evaluated on a dataset of real-world training videos in which each exercise instance was labeled as correctly or incorrectly executed (binary execution-quality classification), consistent with recent video-based systems for exercise analysis and recognition (24,25). After preprocessing and motion segmentation, a ResNet-50 backbone extracted frame-wise spatial embeddings, and a Transformer encoder modeled temporal dependencies across fixed-length frame sequences for execution-quality assessment in unconstrained sports scenes (26,27). Performance was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC analysis.

To improve methodological clarity and avoid ambiguity, we emphasize that pose-derived kinematic descriptors (e.g., joint coordinates, inter-segment angles, and velocities) are computed from an external pose-estimation module during preprocessing, whereas ResNet-50 is used solely to extract visual embeddings from RGB frames. Accordingly, kinematic descriptors are not "produced" by ResNet-50, and no regression branch is used to predict keypoints within the network. In addition, the attention used in our ablation models refers to a pose/feature reweighting module applied before (or independent of) temporal modeling, while the Transformer itself performs temporal self-attention across time steps. The resulting sequence representation is passed to a binary output layer predicting correct versus incorrect execution. The main architectural and training settings are summarized in Table 2.

Table 2. Model architecture and training settings, including ResNet-50 backbone, Transformer and attention configuration, optimization details, and K-fold train/validation/test splits

Parameter	Value / Method
Optimizer	Adam
Initial learning rate	0.0001 (dynamic Reduce-on-Plateau schedule)
Batch size	16
Number of epochs	50 (early stopping enabled)
Dropout rate	0.3 in attention and dense layers
Loss function	Class-weighted cross-entropy (emphasizing the *incorrect* movement)
Overfitting mitigation	Dropout and early stopping

The dataset was split into 80% training and 20% held-out test data at the video level. Five-fold cross-validation (K = 5) was performed only on the training portion for hyperparameter selection, and the final model was evaluated once on the independent test set. Training used the Adam optimizer with an initial learning rate of 0.0001 (Reduce-on-Plateau schedule), a batch size of 16, and up to 50 epochs with early stopping. Dropout (0.3) was applied in attention and dense layers. Because incorrect executions were less frequent, class-weighted cross-entropy increased sensitivity to the incorrect-form class, which is important for injury-risk monitoring (28). Misclassifications were also inspected qualitatively to understand failure modes.

An ablation study quantified the contribution of the convolutional backbone, the attention module, and temporal modeling. Five variants—CNN-only, CNN + attention, CNN + Transformer, Transformer with hand-crafted features, and the full CNN + Transformer + attention model—were trained under the same protocol and evaluated using identical metrics (Table 3).

Table 3. Ablation results showing the impact of the CNN, attention module, and Transformer on model performance.

Model variant	Modification/description	Overall accuracy (%)	Precision	Recall	F1 score	Relative training time
CNN only (no Transformer, no attention)	Spatial feature extraction only; no temporal learning	88.61	0.861	0.887	0.873	Low
CNN + attention (no Transformer)	Focus on key regions without temporal modeling	91.32	0.894	0.910	0.902	Medium
CNN + Transformer (no attention)	Temporal modeling without emphasis on critical regions	93.28	0.906	0.931	0.918	Moderately high
Transformer with hand-crafted features	Sequence modeling over simple, hand-engineered features	90.12	0.876	0.908	0.891	Medium
Proposed method: CNN + Transformer + attention	Full combination: feature extraction, attention, and sequence modeling	97.56	0.962	0.951	0.956	High

The full configuration achieved the best overall performance (97.56% accuracy, precision=0.962, recall=0.951, F1=0.956), indicating that combining spatial encoding, feature reweighting, and temporal modeling yields the most reliable execution-quality discrimination. The CNN-only baseline (88.61% accuracy, F1 = 0.873) confirms that purely frame-level cues are insufficient for robust quality assessment. Adding attention on top of the CNN increased performance to 91.32% accuracy and F1 = 0.902, consistent with improved focus on informative cues and reduced sensitivity to background clutter. The CNN + Transformer variant further improved temporal modeling (93.28% accuracy, F1=0.918), but still lagged behind the full model, supporting the complementary value of explicit feature reweighting. The Transformer with hand-crafted features achieved 90.12% accuracy and F1=0.891, demonstrating that engineered descriptors alone underperform learned visual representations. For reproducibility, the "hand-crafted features" baseline uses pose-derived descriptors only—specifically normalized 2D joint coordinates, selected inter-joint angles (e.g., hip–knee–ankle and shoulder–hip alignment), and their first-order temporal derivatives (velocities)—summarized over the input clip by simple statistics (mean and standard deviation) and then modeled as a temporal sequence by the Transformer. Overall, while simpler configurations train faster, the consistent drop in F1-score (Table 3) shows that the added complexity of the full hybrid architecture is warranted when precise assessment of execution quality is required.

Table 4 reports per-exercise performance and typical error patterns for the incorrect class. The reported per-exercise recall and F1-scores span 0.87–0.94 and 0.875–0.945, respectively, indicating strong performance on common calisthenics movements but reduced sensitivity for technically demanding lifts. Squats achieved the highest results (accuracy 97.71%, F1 = 0.945), with most incorrect cases characterized by anterior knee overtravel. Planks and push-ups remained robust (F1 = 0.925 and 0.915, respectively), while lunges were slightly more angle-sensitive (F1 = 0.901). As expected, deadlifts and crunches were the most challenging exercises (F1 = 0.885 and 0.875, respectively), consistent with subtle trunk-control and range-of-motion deviations that can be difficult to distinguish reliably from monocular video. These per-exercise results are fully consistent with the global test-set metrics in Table 3 and indicate that performance is not driven by a single dominant exercise type.

Table 4. Per-exercise performance of the proposed model, including common error patterns in the incorrect class and corresponding accuracy, precision, recall, and F1-score.

Exercise	Correct/incorrect execution	Test samples	Overall accuracy (%)	Precision	Recall	F1-score	Common error in the incorrect class	Model note
Squat	Correct / anterior knee overtravel	150	97.71	0.95	0.94	0.945	Excessive knee flexion	High-performing
Plank	Correct / sagging lower back	130	95.23	0.93	0.92	0.925	Reduced core tension	Reliable
Push-up	Correct/incomplete elbow angle	120	94.38	0.92	0.91	0.915	Arm asymmetry	Upper-mid performance
Lunge	Correct/incorrect knee alignment	110	92.56	0.91	0.89	0.901	Rear knee not flexing	Angle-sensitive
Deadlift	Correct/excessive lumbar flexion	100	91.80	0.89	0.88	0.885	Loss of trunk balance	Challenging
Crunch	Correct/incomplete movement	90	90.79	0.88	0.87	0.875	Neck laxity	Acceptable

To analyze behavior across exercise types, the model was evaluated separately on six movements—squat, plank, push-up, lunge, deadlift, and crunch—and exercise-wise metrics were computed, as summarized in Table 4. Squats and planks achieved the highest accuracies and F1-scores because typical errors (e.g., excessive knee flexion or reduced core tension) produce visually salient deviations from correct form. Push-ups and lunges also performed strongly, with the model consistently identifying asymmetries and incomplete ranges of motion. Deadlifts and crunches were more challenging due to subtler errors involving trunk stability, neck control, and shallow motion; nevertheless, performance remained acceptable, indicating that the model preserves discriminative power even when deviations are less pronounced or more localized (26,27).

Robustness and generalization were examined using confusion-matrix analysis on independent 20% held-out test subsets, consistent with recent work on video-based activity recognition [24,25]. Two random splits yielded accuracies of 97.66% and 97.44%, with only small variations in false positives and false negatives (Figure 2), indicating stable behavior and supporting the model's suitability for repeated deployment in practice.

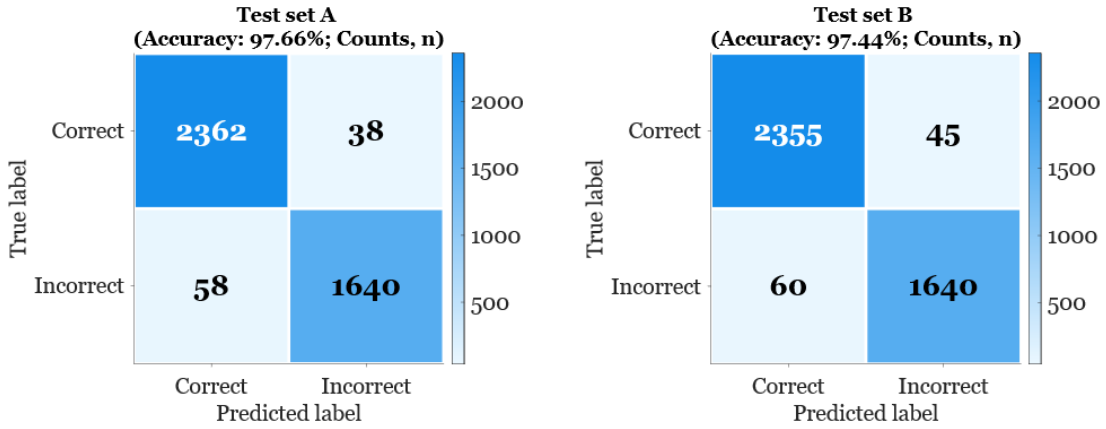

Figure 2. Confusion matrices of the proposed model on two distinct held-out test sets. Cell values report absolute counts (n); rows correspond to true labels and columns to predicted labels, illustrating stable performance in classifying correct vs. incorrect execution.

To probe how architectural choices affect error patterns, confusion matrices were also examined for four representative variants: CNN only, CNN + attention, CNN + Transformer, and Transformer + hand-crafted features. The CNN-only configuration produced the weakest class separation and the largest number of false positives. Adding attention reduced misclassifications by focusing the model on discriminative body-configuration cues, while combining CNN with a Transformer further decreased error rates by modelling temporal dependencies across frames. For the Transformer + hand-crafted-features baseline, we used pose-derived kinematic descriptors only, computed from the estimated 2D skeleton: (1) normalized joint coordinates (x, y) for key joints, (2) selected inter-joint angles (e.g., knee, hip, shoulder, and elbow flexion), (3) trunk inclination and shoulder–hip alignment, and (4) first-order temporal derivatives (angular and joint velocities).

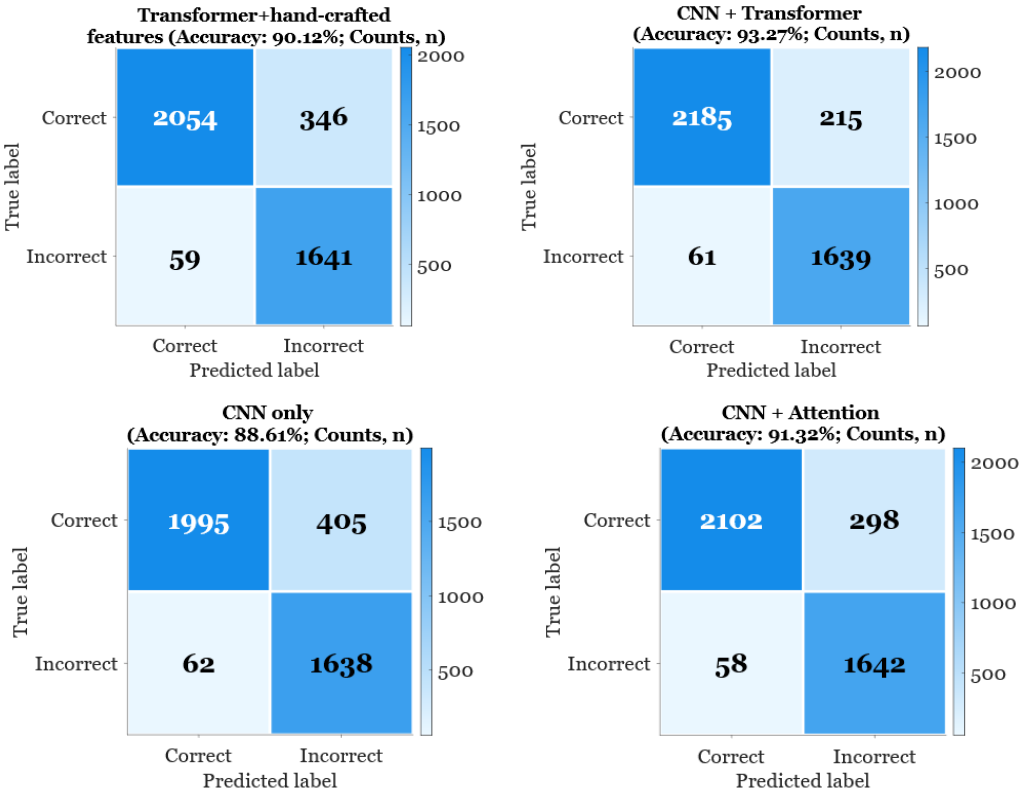

Figure 3. Confusion matrices for four model variants (cell values are counts (n); rows: true labels; columns: predicted labels), arranged left-to-right, top-to-bottom: (1) CNN only; (2) CNN + attention; (3) Transformer + hand-crafted features; (4) CNN + Transformer.

These per-frame descriptors were concatenated into a fixed-length vector (96-D per frame) and input to the Transformer as a temporal sequence. This variant underperformed relative to CNN-based models, confirming that hand-engineered descriptors alone cannot capture the fine-grained visual cues needed for reliable form assessment (24,25). As summarized in Figure 3, the proposed hybrid architecture achieves high aggregate accuracy while consistently lowering misclassification rates, providing a robust basis for video-based evaluation of exercise execution in realistic sports environments and a foundation for biomechanics-aware monitoring and injury-risk prediction (27,28).

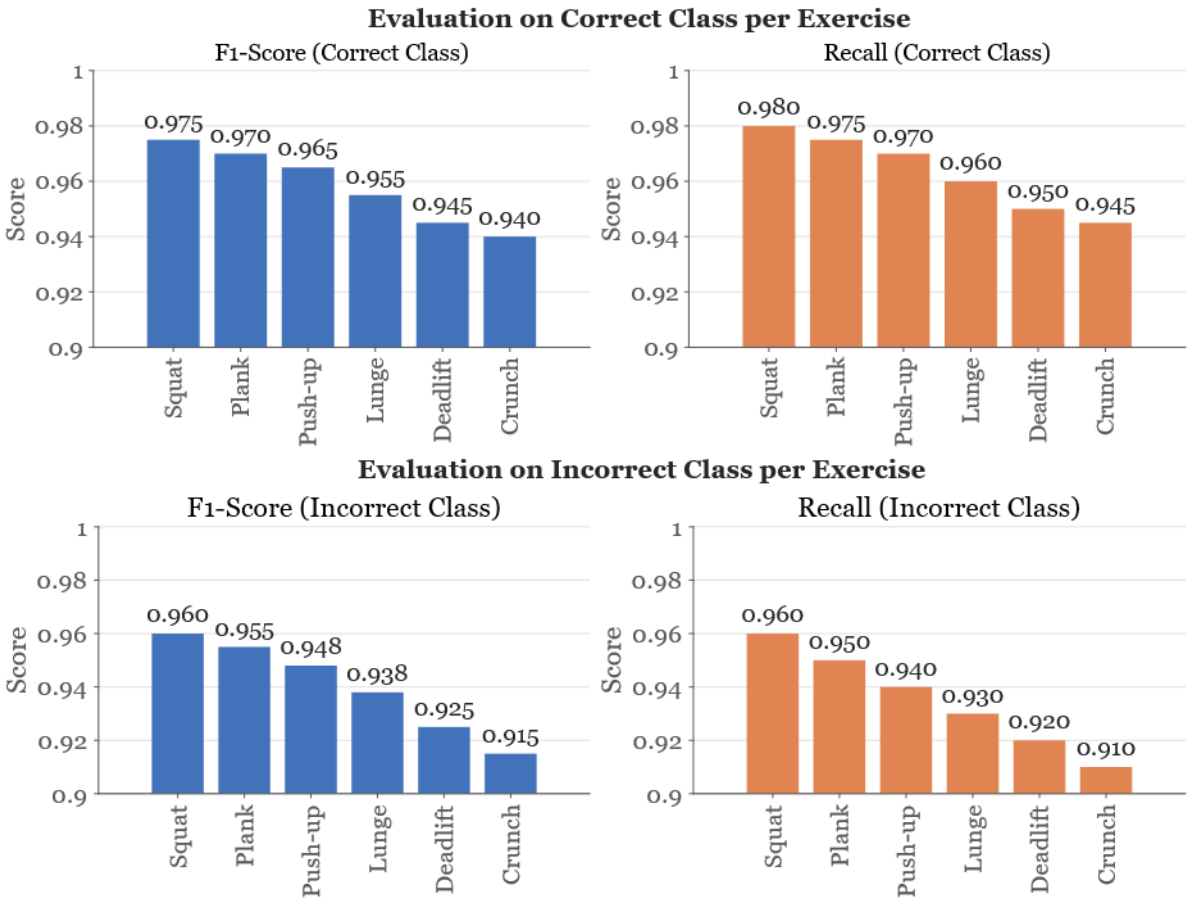

Figure 4. F1-score and recall of the proposed model for correct (top) and incorrect (bottom) executions across movement complexity (0–1 scale).

Discussion

The proposed three-stage architecture—convolutional feature extraction, attention over error-salient regions, and Transformer-based temporal modelling—shows that automatic discrimination between correct and incorrect execution of fitness exercises is achievable under realistic and noisy video conditions. The design remains interpretable in the sense that attention highlights regions associated with execution faults while temporal modelling aggregates evidence across movement phases.

Reliable handling of the incorrect class is essential for technique-assessment systems because missed detections may prolong exposure to harmful mechanics. Figure 4 reports class-wise performance for six representative exercises. For the correct class, F1-scores range from 0.940 to 0.975 and recall from 0.945 to 0.980, indicating consistently

strong recognition of correct execution. For the incorrect class, performance is systematically lower, with F1-scores from 0.915 to 0.960 and recall from 0.910 to 0.960. Even for the most challenging movements (deadlift and crunch), incorrect-class recall remains above approximately 0.91, suggesting practically useful sensitivity to faulty technique. Throughout this section, metric references explicitly distinguish between correct-class and incorrect-class results, eliminating ambiguity with respect to Figure 4.

Across independent held-out splits, confusion matrices and aggregate metrics vary only modestly, consistent with sampling variability rather than sensitivity to a particular partition. This stability suggests that the staged pipeline generalizes across subsets of the dataset under the same acquisition conditions.

To relate performance to movement difficulty, Figure 5 compares class-wise F1-scores with a Biomechanical Complexity Index (BCI). The BCI is a quantitative, normalized proxy intended to rank relative movement complexity rather than to estimate absolute biomechanical load. For each exercise e, four components are defined: the number of actively involved joints J_e, range of motion ROM_e (average peak-to-peak angular excursion of key joints), angular sensitivity S_e (number and variability of critical joint angles), and balance/stability demand B_e(trunk-control and multi-joint coordination requirements). Each component is min–max normalized across the exercise set:

$$\tilde{x}_e = \frac{x_e - \min(x)}{\max(x) - \min(x)} \quad (2)$$

The normalized components are combined as:

$$BCI_e = w_J \tilde{J}_e + w_{ROM} \widetilde{ROM}_e + w_S \tilde{S}_e + w_B \tilde{B}_e \quad (3)$$

with equal weights in this study for interpretability:

$$w_J = w_{ROM} = w_S = w_B = 0.25, \sum_k w_k = 1 \quad (4)$$

This yields $BCI_e \in [0,1]$. As shown in **Figure 5**, increasing BCI is associated with decreasing F1-score, with a more pronounced reduction for the incorrect class.

Incorrect execution is harder to detect because it is not a single pattern but a heterogeneous set of deviations. Many incorrect cases preserve the global appearance of correct form while differing only in small joint-angle changes, increasing overlap between class distributions. In monocular instructional videos, viewpoint variation and partial occlusion can further obscure key joints, reducing the observability of alignment errors. In addition, several faults are phase-specific (e.g., brief loss of trunk stability near transitions or incomplete depth at peak flexion), so short clips or suboptimal timing can weaken the temporal evidence available to the model. These factors explain the lower incorrect-class performance for deadlifts and crunches in Figures 4 and 5. For deadlifts, trunk stability and hip–spine coordination may be difficult to infer reliably from 2D appearance. For crunches, the limited range of motion and subtle compensations produce weak visual signatures. Despite these challenges, the incorrect-class recall remains above approximately 0.91, highlighting practical value while indicating where richer inputs (e.g., explicit joint-angle supervision or improved temporal segmentation) are most likely to yield further gains.

Relative to prior work that focuses on pose estimation without explicit quality assessment, injury-risk prediction without form-level interpretability, or single-exercise form classification under controlled conditions, the proposed framework supports multi-exercise form evaluation on realistic instructional videos and links exercise biomechanics (via BCI) to class-wise performance trends. This connection provides a principled way to prioritize additional data collection or specialized modelling for higher-complexity movements.

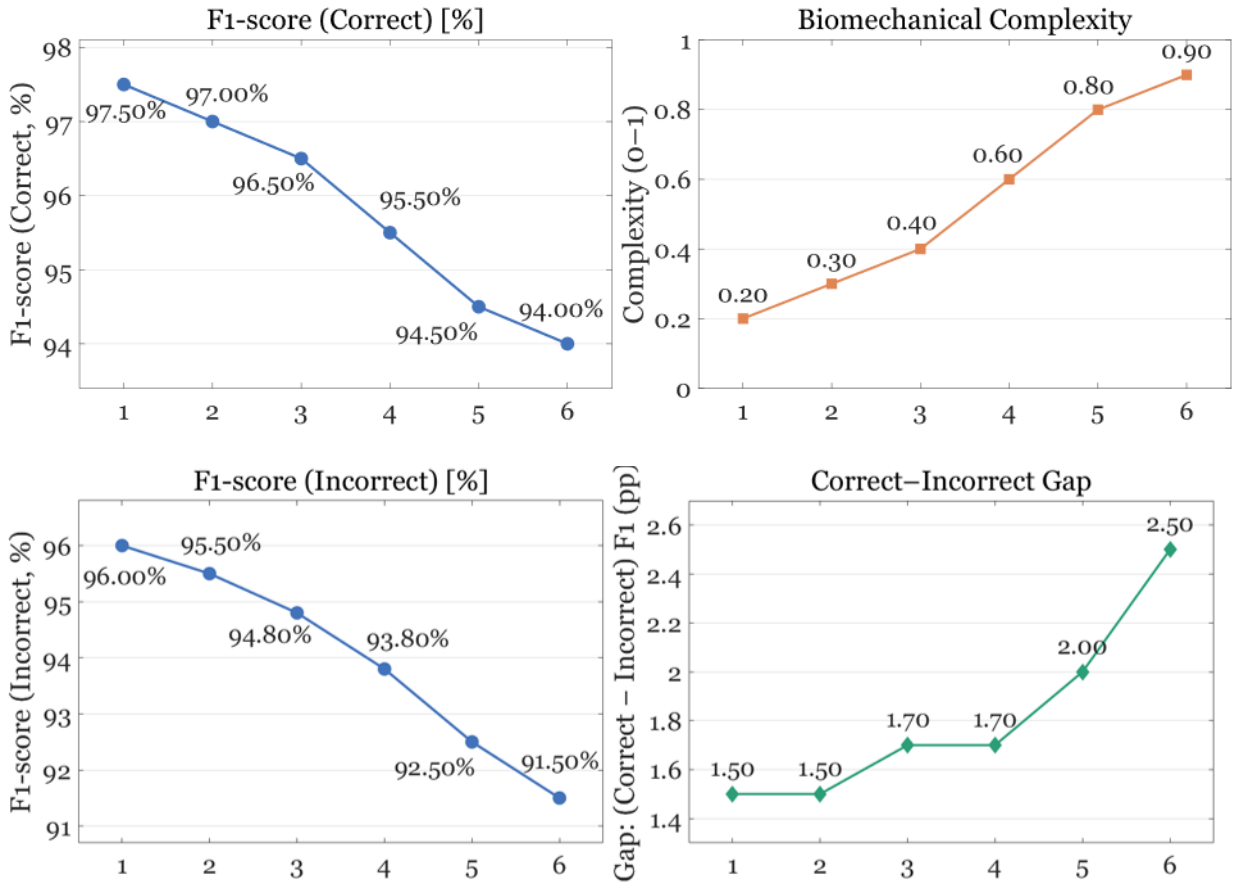

Figure 5. F1 scores and biomechanical complexity for six exercises by class (correct, top; incorrect, bottom). Left: per-exercise F1. Right: biomechanical complexity (joints, angular sensitivity, range of motion, balance), showing lower F1 with higher complexity.

The six exercise case studies translate these trends into practical insight. In Figure 4, correct-class performance is uniformly high for squat, plank, and push-up (F1 ≈ 0.965–0.975; recall ≥ 0.97), consistent with visually stable patterns and low inter-subject variation. The incorrect class, however, is harder to detect because "incorrect execution" is heterogeneous and often subtle. For squat and plank, typical faults (e.g., knee overtravel, lumbar sag) produce relatively clear spatial cues, yielding strong incorrect-class performance as well. Push-ups and lunges form an intermediate group: common deviations such as arm asymmetry, incomplete elbow flexion, and knee misalignment can be small in magnitude and sensitive to viewpoint, which increases overlap between correct and incorrect samples and reduces incorrect-class recall compared with simpler movements.

Deadlifts and crunches represent the most challenging regime in Figures 4–5, particularly for the incorrect class (e.g., recall around 0.91–0.92). Many deadlift faults (mild lumbar flexion, loss of trunk stiffness, altered hip–spine coordination) and crunch faults (short ROM, subtle cervical/trunk compensation) can be phase-specific and difficult to infer reliably from monocular 2D video, especially under occlusion or off-axis viewpoints. Despite this, incorrect-class recall remains above ~0.91, indicating practical sensitivity to potentially harmful technique while also identifying the exercises most likely to benefit from richer supervision (e.g., explicit joint-angle features), improved temporal segmentation, or multi-view inputs.

These observations motivate the BCI used in Figure 5 as a quantitative proxy for relative exercise difficulty. By combining normalized terms for the number of active joints, range of motion, angular sensitivity, and balance/stability demand (with the weighting scheme described in the BCI formulation), BCI formalizes the expectation that higher coordination and tighter kinematic tolerances reduce class separability—particularly for the incorrect class—and provides a practical guide for prioritizing denser sampling,

targeted augmentation, or exercise-specific modelling for high-BCI movements.

Comparison with existing approaches clarifies the position of the proposed framework. Prior work in automated sports-movement analysis typically emphasizes (1) pose/keypoint estimation without explicit quality classification, (2) injury-risk prediction without form-level error discrimination, or (3) correct/incorrect classification limited to a single exercise under constrained conditions. In contrast, the present method performs multi-exercise form assessment on real instructional videos and explicitly links exercise biomechanics (via BCI) to class-wise performance patterns (Table 5; Figures 4–5).

Architecturally, assigning distinct roles to the convolutional backbone, attention module, and Transformer makes spatial feature extraction, emphasis on error-salient regions, and temporal relation modelling complementary. Ablation experiments show that removing any component leads to a measurable drop in F1-score, while the overall design remains modular and deployable without specialized motion-capture systems.

Table 5. Comparative summary of our method and six related studies by research objective, input modality, model architecture, form-assessment capability, and biomechanical analysis scope.

Study	Research objective	Data type	Model type	Form evaluation	Biomechanical analysis
Ding et al. [16]	Estimating motion from skeletal data using attention	Keypoint (skeleton) sequences	Convolutional neural network + attention mechanism	Continuous motion signals	None
Bazarevsky et al. [23]	Real-time body-pose estimation for fitness applications	Lightweight/low-latency video	Lightweight CNN (BlazePose)	None (keypoint extraction only)	None
Sadr et al. [28]	Predicting sports-injury risk by combining CNN and RNN	Video + sensor data	CNN + recurrent neural network	None (focus on injury prediction)	Partial (via injury labels)
Rangari et al. [25]	Classifying correct vs. incorrect plank form	YouTube plank videos	OpenPose + LSTM	Yes (plank only)	None
Chen and Yang, [29]	Exercise posture correction + feedback using pose estimation	RGB video → 2D pose keypoints (OpenPose)	Geometry-based heuristics + ML classifier	Yes (correct/incorrect + feedback), multi-exercise	Rule-based kinematic cues (angles/distances); no explicit complexity index
Hsu et al., [30]	Classify correct vs. incorrect lifting posture (smartphone camera)	Smartphone RGB video (multi-angle) → OpenPose keypoints + derived features	BiLSTM sequence classifier	Yes (correct vs incorrect posture)	Uses biomechanical/kinematic features (e.g., joint angles + trunk alignment proxy); no complexity index
Proposed method (this study)	Classifying correct/incorrect form across diverse exercises	Instructional YouTube videos	CNN + Transformer + attention	Yes (correct/incorrect), multi-exercise	Yes (linking biomechanical complexity to accuracy)

Conclusion

This work investigated whether exercise execution quality can be assessed directly from ordinary monocular video using a hybrid deep-learning framework for dynamic movements. The proposed pipeline combines convolutional feature extraction, attention to emphasize error-salient regions, and Transformer-based temporal

modelling, enabling both exercise recognition and correct/incorrect form assessment within a single architecture. The model achieved an overall accuracy of 97.56% and an F1-score of 0.956, and ablation results confirmed that the full CNN–attention–Transformer design provides the most stable and accurate performance among the tested variants. We also introduced a quantitative Biomechanical Complexity Index (BCI) and observed an inverse relationship between complexity and performance, highlighting exercises where additional supervision or targeted modelling is most beneficial. To strengthen the significance of these results, future work will include a systematic benchmark against stronger state-of-the-art methods in the same domain (e.g., pose-based temporal models and recent video-Transformer architectures), evaluated under identical splits and class-wise metrics. In addition, we will extend coverage to more exercises and assess robustness in real gym and rehabilitation settings, potentially incorporating multi-view/3D cues or wearable sensors for high-complexity movements.

Acknowledgments

We thank the sport/clinical advisors and annotators. AI tools were used only for language polishing.

Funding

No specific funding was received.

Data Availability

The code and dataset for this paper are available from the corresponding author upon reasonable request.

Conflict of Interest

The authors declare no competing interests.

Approval of Ethics for Studies Involving Animals and Human Participants

Analysis used publicly available instructional videos; no new human or animal data were collected, and no private identifiers were analyzed.

References

1. Penichet-Tomas A. Applied biomechanics in sports performance, injury prevention, and rehabilitation. Appl Sci. 2024;14(24):11623.
2. Zhao F. The application of sports biomechanics in sports injury prevention and rehabilitation. Front Sport Res. 2024 May 20;6(3):142-7.
3. Pardos A, Tziomaka M, Menychtas A, Maglogiannis I. Automated posture analysis for the assessment of sports exercises. In: Proceedings of the 12th Hellenic Conference on Artificial Intelligence; 2022 Sep 7; (pp. 1-9).
4. Jayaweerage I, Rajamanthri K, Kothalawala A, Gunawardana N, Induranga A, Weerakkody P, et al. Motion capturing in cricket with bare minimum hardware and optimised software: a comparison of MediaPipe and OpenPose. In: 2024 First International Conference on Software, Systems and Information Technology (SSITCON); 2024 Oct 18; (pp. 1-8).
5. Niu Z. A lightweight two-stream fusion deep neural network based on ResNet model for sports motion image recognition. Sens Imaging. 2021;22(1):26.
6. Cust EE, Sweeting AJ, Ball K, Robertson S. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. J Sports Sci. 2019;37(5):568-600.
7. Ma J, Ma L, Ruan W, Chen H, Feng J. A wushu posture recognition system based on MediaPipe. In: 2022 2nd International Conference on Information Technology and Contemporary Sports (TCS); 2022 Jun 24; (pp. 10-13).
8. Ashok A, Surendiran B, Babu S. Yoga pose estimation using MoveNet deep learning models. Sadhana. 2025;50(3):125.
9. Dayapoglu M, Savirdi E, Akcaba I, Alp S, Kucuk S. 3D skeleton-based sport action recognition system via deep learning. In: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU); 2024 Oct 16; (pp. 1-6).
10. Pagnon D, Domalain M, Reveret L. Pose2Sim: an open-source Python package for multiview markerless kinematics. J Open Source Softw. 2022;7(77):4362.
11. Assa A, Janabi-Sharifi F, Plataniotis KN. Sample-based adaptive Kalman filtering for accurate camera pose tracking. Neurocomputing. 2019;333:307-18.
12. Ziche L, Lenzig B, Bieck R, Neumuth T, Schoenfelder S. Image-based deep learning of finite element simulations for

fast surrogate biomechanical organ deformations. In: Current Directions in Biomedical Engineering. 2022;8(2):528-31.
13. Boyle A, Ross GB, Graham RB. Machine learning and deep neural network architectures for 3D motion capture datasets. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2020 Jul 20; (pp. 4827-30).
14. Pan X, Luo Z, Zhou L. Comprehensive survey of state-of-the-art convolutional neural network architectures and their applications in image classification. Innov Appl Eng Technol. 2022 Mar 11;:1-6.
15. Wu Z. Application of CNN classic model in modern image processing. J Adv Eng Technol. 2024 Nov 13;1(3):1-6.
16. Ding G, Plummer A, Georgilas I. Deep learning with an attention mechanism for continuous biomechanical motion estimation across varied activities. Front Bioeng Biotechnol. 2022;10:1021505.
17. Yang J, Ismail AW. A review: deep learning for 3D reconstruction of human motion detection. Int J Innov Comput. 2022;12(1):65-71.
18. Liu R, Chen Y, Gai F, Liu Y, Miao Q, Wu S. Local and global spatial-temporal transformer for skeleton-based action recognition. Neurocomputing. 2025;634:129820.
19. Ahmad T, Wu J, Alwageed HS, Khan F, Khan J, Lee Y. Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection. IEEE Access. 2023;11:33148-59.
20. Zhao X. Research on athlete behavior recognition technology in sports teaching video based on deep neural network. Comput Intell Neurosci. 2022;2022(1):7260894.
21. Pandey S, Kumar N. Enhancing human action recognition in high-resolution videos using ConvLSTM and LRCN model. In: 2023 4th International Conference on Communication, Computing and Industry 6.0 (C2I6); 2023 Dec 15; (pp. 1-5).
22. Al-Timemy AH, Castellini C, Escudero J, Khushaba R, Muceli S. Current trends in deep learning for movement analysis and prosthesis control. Front Neurosci. 2022;16:889202.
23. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M. BlazePose: on-device real-time body pose tracking. arXiv [preprint]. 2020 Jun 17:arXiv:2006.10204.
24. Cheng SH, Sarwar MA, Daraghmi YA, Ik TU, Li YL. Periodic physical activity information segmentation, counting and recognition from video. IEEE Access. 2023;11:23019-31.
25. Rangari T, Kumar S, Roy PP, Dogra DP, Kim BG. Video based exercise recognition and correct pose detection. Multimed Tools Appl. 2022;81(21):30267-82.
26. Li Y, He H, Zhang Z. Human motion quality assessment toward sophisticated sports scenes based on deeply-learned 3D CNN model. J Vis Commun Image Represent. 2020;71:102702.
27. Zhang M, Li Y, Cui Y. The use of deep learning in intelligent athlete motion recognition: integrating biological mechanisms. Molecular & Cellular Biomechanics. 2025;22(1):670.
28. Sadr MM, Khani M, Tootkaleh SM. Predicting athletic injuries with deep learning: evaluating CNNs and RNNs for enhanced performance and safety. Biomed Signal Process Control. 2025;105:107692.
29. Chen S, Yang RR. Pose Trainer: correcting exercise posture using pose estimation. arXiv [preprint]. 2020:arXiv:2006.11718.
30. Hsu GSJ, Wu JS, Huang YKD, Chiu CC, Kang JH. Automatic detect incorrect lifting posture with the pose estimation model. Life (Basel). 2025;15(3):358.