

Real-time Detection of Basketball Free-throw Technique Movements Using a Deep Learning Model Based on Smartphone Video

Jinyu Wang*

School of Physical Education, Chongqing Preschool Education College, Chongqing 404047, China

**Corresponding author: Jinyu Wang.*

Abstract

This study aims to develop a deep learning model based on smartphone video recordings to enable real-time detection and feedback of basketball free-throw technique execution. Traditional analysis methods rely on manual observation, which suffers from feedback latency and high subjectivity. In this research, free-throw videos of 10 athletes were recorded using smartphones, and certified coaches classified the videos according to standard technical criteria; error categories included unstable stance, improper knee flexion, and excessive elbow abduction. The MoveNet model was employed to extract 17 key body points, and 26 features—including joint angles and inter-joint distances—were computed. The datasets were enhanced using techniques such as Gaussian noise, jitter, scaling, translation, and temporal distortion. Four models were trained and compared; on the balanced dataset, the dual-stream convolutional neural network model achieved the best overall performance, with an accuracy of 93.8%, precision, recall, and F1-score of 93.2%, 92.5%, and 92.2%, respectively, and a logarithmic loss of 0.23. Five-fold cross-validation yielded an accuracy of $93.8\% \pm 0.5\%$, demonstrating excellent stability. Confusion matrix analysis revealed that the model correctly predicted 804 out of 876 samples, exhibiting optimal class balance. This study validates the feasibility of this approach in providing athletes with instantaneous quantitative feedback; future work may explore the integration of edge computing and multimodal fusion to enhance robustness.

Keywords

Basketball free-throw; Deep learning; Motion capture

1. Introduction

Free throws in basketball are a pivotal factor in determining the outcome of a game. Traditional analysis methods primarily relied on manual observation and slow-motion video replay, which were plagued by issues such as feedback latency and high subjectivity. With the widespread adoption of smartphone cameras and advancements in deep learning computing capabilities, it has now become possible to employ markerless motion capture technology to achieve real-time, objective movement analysis. This approach not only assists athletes in refining their technique but also provides coaches with quantified training data.

In recent years, real-time detection of basketball athletic movements using smartphone video has emerged as a cutting-edge research direction integrating computer vision and sports training. By leveraging deep

learning models to analyze video footage captured on mobile devices, this approach enables precise action recognition and real-time decision-making, making it a prominent research focus in this field. Early studies centered on object detection—e.g., employing the YOLO series of models combined with lightweight networks to achieve efficient basketball action detection [1, 2]. With advancing technological capabilities, research efforts have shifted toward more refined pose estimation and spatiotemporal feature extraction, such as utilizing 3D convolutional neural networks (3D-CNN) to capture continuous dynamic information during actions like shooting [3].

However, deploying deep learning models on mobile devices faces two major challenges: computational resource constraints and complex shooting environments. To address these issues, existing research has proposed solutions such as model lightweighting and enhanced environmental robustness. Meanwhile, the construction of high-quality datasets—such as the Basketball-51 dataset for competition scenarios and the MultiSubjects dataset for training scenarios [4]—has provided critical support for model training and evaluation.

Although existing research has made progress in terms of recognition accuracy and real-time performance—satisfying preliminary application requirements—there remains considerable room for further exploration in the analysis of the specific and nuanced technical movement of basketball free throws. In particular, the focus of this study is on constructing a complete closed-loop system, encompassing data acquisition, processing, model training, and feedback, to provide athletes with immediate and precise corrective guidance on their movement techniques. This study aims to build, from scratch, a deep learning model specifically designed for analyzing basketball free-throw technical movements, leveraging smartphone video input to deliver real-time feedback to athletes for improving their free-throw accuracy. Subsequent sections will detail our research methodology, data processing pipeline, model architecture, and experimental results.

2. Correlation Studies

Real-time detection of basketball athletic techniques using smartphone video is a cutting-edge research direction integrating computer vision and sports training. By employing deep learning models to analyze video footage captured on mobile devices, this approach enables precise action recognition and real-time assessment. In basketball, pose estimation models are commonly utilized to predict shooting accuracy; adopting correct shooting postures can significantly improve shooting success rates, making this area a prominent research focus in the interdisciplinary application of computer vision within sports engineering. This section of the literature review examines studies that employ diverse techniques to track shooting techniques, including the application of computer vision (Human Pose Estimation), object detection, and biomechanics in basketball.

In early research on real-time detection of basketball technical movements, the core technical approach was to leverage object detection to drive action localization, thereby transforming basketball movement recognition into an object detection task. By constructing specific action templates and detecting key frames of these actions, researchers achieved precise localization and recognition of basketball technical movements. This methodology features a relatively simple model architecture, making it easier to deploy in a preliminary stage on mobile devices with limited computational resources; consequently, it has become a foundational research direction in this field. Within this technical framework, the YOLO series of models has emerged as the predominant choice in such studies, owing to their combination of rapid detection speed and high recognition accuracy. Numerous studies have achieved efficient real-time detection of basketball movements by integrating YOLO with lightweight networks. For instance, one study employed YOLOv3-tiny combined with the lightweight MobileNetV3 network and incorporated background subtraction techniques specifically for shot detection, ultimately achieving a real-time recognition accuracy of 91.4% [5]. Another study integrated YOLO with a deep fuzzy LSTM network to investigate the recognition of fundamental basketball actions—such as walking, dribbling, and shooting—achieving a recognition accuracy exceeding 96%, while maintaining detection efficiency that meets real-time requirements for mobile devices [6]. In 2025, this research direction achieved new breakthroughs, with related studies proposing a technical strategy that integrates 3D convolutional neural networks (3D-CNN) with a single-source multi-box detector (SSD). By leveraging SSD to perform precise real-time localization of key player body parts, this approach delineates accurate feature regions for subsequent action classification. This methodology elevates the basketball action recognition rate

to approximately 95%, achieves a processing speed of 95 ms per frame, and far exceeds the threshold requirements for real-time detection on mobile devices [7], thereby significantly enhancing both the accuracy and computational efficiency of basketball action recognition driven by object detection.

With the advancement of hardware computing capabilities and the continuous enrichment of specialized basketball movement annotation datasets, the core technical approach in this field has gradually shifted from coarse-grained action localization toward a direction that integrates fine-grained pose estimation with spatiotemporal feature extraction; this has become the prevailing research paradigm for real-time detection of basketball technical movements. This approach focuses on modeling the spatiotemporal dynamic characteristics of basketball technical movements, enabling refined identification of such movements through precise representation of action poses and spatiotemporal features. Among these methods, 3D convolutional networks have emerged as the cornerstone technique for spatiotemporal feature extraction; they can directly extract dual temporal and spatial features from continuous video sequences captured via smartphones, thereby accurately capturing the sequential dynamic information of basketball actions such as shooting and passing. In 2024, existing studies have already utilized 3D-CNNs to achieve precise modeling of typical basketball technical movements, providing critical technical support for refined action identification [8]. Meanwhile, optical flow and two-stream convolutional networks represent classic approaches to spatiotemporal feature extraction; they capture dynamic information of moving subjects via optical flow and combine it with spatial features from static images to perform action recognition. Although these methods exhibit high recognition accuracy, their overall computational demands are excessively high, making them difficult to deploy directly on smartphones—devices constrained by limited computing resources and power consumption. Nevertheless, they offer valuable insights for subsequent research on spatiotemporal feature fusion [9].

Addressing the two core challenges—computational bottlenecks in smartphone mobile deployments and the complexity of real-world shooting environments—existing research has proposed targeted technical solutions, thereby laying a solid foundation for the practical deployment of deep learning models on mobile devices. In the realm of model lightweight deployment, researchers have achieved technological breakthroughs primarily through two approaches: first, employing model optimization techniques such as network pruning and knowledge distillation—where pruning removes redundant neurons and connections from deep learning models, or knowledge distillation transfers the feature extraction capabilities and recognition knowledge of large models to lightweight architectures like MobileNet and GhostNet; this approach significantly reduces model parameters and computational requirements while effectively preserving detection accuracy; second, developing network architectures specifically tailored for mobile devices—such as MobileViT and GhostNet, which are designed for edge computing—these architectures fully leverage the hardware characteristics of mobile devices, enabling efficient feature extraction under conditions of low power consumption and limited computational resources, making them preferred architectures for deploying real-time basketball action detection models on mobile devices. To address robustness in complex environments, this study enhances the environmental adaptability of the model by integrating multiple techniques. Building upon action detection, it incorporates the DeepSORT multi-object tracking algorithm, employs spatiotemporal correlation analysis to filter and refine detection results, and eliminates false detections caused by occlusion or camera jitter, thereby overcoming the instability inherent in single-frame detection [10]. Simultaneously, it leverages optical flow and background subtraction techniques: optical flow is used to estimate the camera's motion trajectory and perform motion compensation, thereby correcting image shifts induced by jitter; background subtraction is then applied to separate moving subjects from static backgrounds in the video, reducing the interference of background information on action feature extraction, which further improves detection accuracy [11].

The training and performance evaluation of deep learning models heavily rely on large-scale, high-quality annotated datasets. To address the limited adaptability of generic human action datasets to basketball scenarios, researchers have developed several specialized basketball action detection datasets, providing standardized sample support for model training, optimization, and performance evaluation. Among these, the Basketball-51 dataset is tailored to basketball game scenarios; its samples are extracted from live basketball game videos, with particular annotations for various scoring actions such as three-point shots and free throws. This dataset is primarily used to assess the accuracy of models in identifying basketball scoring actions within game contexts, serving as a critical data foundation for competitive-level basketball action detection research [12]. The MultiSubjects dataset, in contrast, targets basketball training scenarios; it is acquired in actual basketball court environments and comprises multi-view videos of individual basketball techniques. This dataset

effectively addresses the challenge of diverse camera angles in real-world training settings, providing application-aligned samples for training real-time basketball action detection models on smartphone devices [13].

In current research, a series of technical approaches—including lightweight optimization of deep learning models, spatio-temporal feature fusion, and robustness handling in complex environments—have largely addressed the core technical challenge of achieving sub-second latency for basketball movement detection on smartphones. Existing models now demonstrate recognition accuracy and real-time performance sufficient to meet the application requirements of routine basketball training and preliminary match analysis, thereby realizing the initial practical deployment of computer vision technologies in the basketball domain. Given the emerging trends in cutting-edge technologies such as edge computing and multimodal fusion, future research on real-time basketball movement detection using smartphone video will advance further along two key directions: edge computing integrated with cloud collaboration, and multimodal information fusion. On one hand, by leveraging the integrated application of edge computing and cloud computing, computationally intensive feature extraction tasks—such as 3D convolutional spatiotemporal feature modeling—can be offloaded to the cloud. This approach harnesses the cloud's powerful computing capabilities to achieve precise feature analysis, while the smartphone merely handles video acquisition and real-time reception and feedback of detection results. This strategy not only fulfills the critical requirement for real-time detection on mobile devices but also further enhances the accuracy of motion recognition. On the other hand, by conducting research on multimodal information fusion modeling—specifically, by integrating data from the smartphone's built-in inertial measurement unit (IMU) with video information captured via vision systems—a multimodal fusion detection model can be established. This model utilizes inertial data to compensate for information gaps in visual features under challenging conditions, such as occlusion or poor lighting, thereby further improving the precision of player posture estimation and the robustness of the model in complex shooting environments [14]. Consequently, this approach enables fine-grained and comprehensive detection of basketball technical movements, providing more complete and accurate motion data to support scientific training in basketball.

3. Research Technique

This study pioneers a deep learning model for analyzing basketball free-throw technique movements, initiating the development process from fundamental aspects. Through data collection and data processing, the study ensures that the model aligns with its research objectives. The research focuses on training a functional model that primarily utilizes video input, providing feedback to help basketball players improve their free-throw accuracy. To achieve these goals, several key steps must be completed, including dataset collection and model training.

First, the researchers fixed the camera at a specific position and angle to record videos of athletes performing free throws. Next, the researchers edited the videos into short video clips containing only a single repetitive motion, removing extraneous movements before and after the athlete's action. Subsequently, basketball coaching instructors analyzed each individual video clip and classified them into predefined categories. Once the video classification was complete, the MoveNet pose estimation model was applied to each video; this model returned an array comprising 17 human body key points and the tracked basketball. Using these coordinates, the researchers calculated joint angles and joint distances to assess the athlete's body posture during free throws, and tracked the basketball to determine the release point height, ball acceleration, and ball trajectory; these data points were then added to the array, as they are critical for evaluating free throw accuracy. Following this, the video data underwent enhancement processing through temporal and spatial perturbations. The final data processing step involved normalizing the key points to the range of 0 to 1. Upon completion of data processing, the researchers proceeded to the training phase, during which the dataset was split into a training set and a test set in an 8:2 ratio. Finally, the selected model was trained and evaluated. Each step of this methodology is detailed in the following subsections.

3.1 Data Collection

This study aims to train a model that provides feedback to athletes who are conducting free-throw practice on improving their shooting accuracy. Accordingly, the data collection methodology must be capable of replicating the athlete's state during free-throw attempts, while capturing biomechanical posture data during

the preparation phase, release phase, and execution phase of the free throw, thereby ensuring the accuracy and consistency of the feedback. For this purpose, the data collection protocol for athletes in this study is as follows:

The video segments must capture basketball players during the execution of free throws. By establishing specific shooting angles, researchers facilitate the standardized extraction of video content. Selecting appropriate angles is crucial for capturing the physiological posture of the athlete during the preparation phase, the ball release phase, and the release phase. Shooting angles aligned with the athlete's sagittal and frontal planes enable comprehensive observation of the athlete's full body posture during free throws. The extracted video segments encompass the complete sequence of the athlete's free throw: the preparation phase, the ball release phase, the release phase, and the scoring phase.

When performing video segment extraction, it is necessary to communicate the extraction requirements to the researchers. Given the large volume of video data to be processed, some segments may fail to meet the specified criteria; however, these non-compliant segments are not deleted, enabling the model to generate classification results indicating “incorrect shooting angles,” thereby helping athletes undergoing penalty shootout training to refine their shooting techniques.

During the video extraction process for players performing free throws in training sessions, the captured videos exhibited variations in camera angle and background environment. Additionally, the athletes wore different uniforms during the training sessions. Given that the pose estimation model employed in this study possesses strong robustness, these environmental variations did not adversely affect the quality of the results. This study utilized the MoveNet model to capture athletes' poses; MoveNet is a single-person, 2D pose estimation model developed by Google, which demonstrates superior performance in detecting diverse human poses across both static images and videos compared to PoseNet, OpenPose, and Mediapipepose.[15] To ensure exceptional robustness, only a single athlete was recorded performing a free throw in each video; the presence of multiple athletes or occlusions within a video could result in multiple body key points, which might interfere with the detection performance of the MoveNet model.

In this study, 10 basketball players were filmed, with each player recording 20–30 free-throw videos. The participants ranged in age from 19 to 27 years, and their skill levels varied from those with limited game experience to those with extensive competitive experience; all had received basketball training. This foundational characteristic was crucial for collecting reliable and consistent motion data. The participants included coaches and players who had competed in collegiate basketball leagues; following the video recordings, the coaches conducted technical analyses.

3.2 Data Classification

The videos are categorized by E-level basketball coaches certified by the Chinese Basketball Association. Established in 1956, the Chinese Basketball Association serves as the top-level designer, standard-setter, and core organizer of coach training programs. This certification indicates that the participating coaches are capable of conducting youth basketball instruction and training, and can provide technical guidance to athletes during free-throw practice sessions. The video categorization process follows a “training workflow,” which includes a dedicated “error handling system” enabling coaches to classify technical errors according to priority—prioritizing the correction of errors where significant technical deviations lead to low free-throw accuracy rates. This is followed by the “feedback and correction” phase, where coaches must provide athletes with clear, concise, and actionable feedback to correct their technical movements.

The methodology employed by the coach for video classification in this study involves identifying the primary errors in the athlete's movements and categorizing them to provide feedback prompts for correcting the technical aspects of the athlete's performance. In this study, if an athlete exhibited multiple errors in their free-throw technique, only the primary error components were categorized for correction, in accordance with the classification criteria established by the E-level basketball coach.

The cues and classification categories selected in this study are based on the accurate descriptions of free-throw technique provided by coaches in “BASKETBALL FREE THROW: A WRITTEN TECHNICAL REPORT” [16], as well as the descriptions of correct free-throw technique offered by Veitia Fenty et al. [17]. According to these references, the correct technical execution of a free throw encompasses the following aspects:

- Regarding positioning and foot placement, keep your feet shoulder-width apart, with the supporting foot—the opposite side of your dominant foot—slightly stepped back to create a stable triangular base. Additionally, point your toes slightly toward the center of the basket to ensure your body is aligned horizontally with the basket's center line.
- When the knees and hips are flexed, after standing up, quickly squat down with the knees flexed at approximately 30° to 45° and the hips also in a flexed position, thereby creating a “spring-like” energy-storing state. If the flexion range is too limited, the available force will be insufficient; if it is too wide, it may disrupt the rhythm of the shot.
- Regarding the upper body and arm positioning, keep the upper body slightly tilted forward with the chest facing the basket to maintain balance. When shooting, position your elbow close to your body to create an elbow angle of approximately 90° . The non-shooting hand should assume a supporting position or lightly touch the ball for assistance.
- Regarding wrist and finger control, during the release phase of a shot, the wrist should rapidly “pop” forward and upward, while the fingers naturally spread apart, allowing the fingertips to make contact with the ball. The wrist extension angle should be approximately 90° .
- During the release and follow-through phases, leverage the elasticity of the lower limbs to generate upward drive, combined with arm extension and wrist snap, to complete the release. In the follow-through motion, allow the arm to rise naturally while flipping the palm toward the basket to create a “rope” posture, which helps the ball maintain a stable arc.

Based on the aforementioned elements of correct technical execution in free throws, the video footage is categorized as follows:

- An unstable stance—characterized by feet positioned too close together or toes not pointing toward the basket—can cause the body to sway side-to-side during a shot, which in turn negatively impacts the shooting percentage.
- There may be issues with either insufficient or excessive knee flexion; if the knee flexion angle is too small, the elastic energy reserve will be inadequate, leading to weaker shooting power; whereas excessive knee flexion shifts the center of gravity downward, making it difficult to rebound quickly upon release, which in turn disrupts the shooting rhythm.
- There is also the issue of elbow flare: when shooting, the elbow opens outward, causing the shot trajectory to deviate toward either the left or the right side, making it difficult to establish a consistent shooting arc.
- Incorrect timing in wrist action—whether the wrist snaps forward prematurely or releases too late—often results in the ball's trajectory being too low or too high, or insufficient spin.
- The follow-through motion is absent; after shooting, the arm is immediately pulled back, lacking a natural upward swing and palm rotation. This results in insufficient backspin of the ball and a decrease in shooting accuracy.
- The visual focus is unclear; when shooting, the player's gaze is directed toward the perimeter of the basket or the ball itself.
- An excessive forward tilt of the upper body can cause the center of gravity to shift forward during a shot, resulting in a shallower release angle and a decrease in shooting accuracy.
- Interference by a non-shooting hand: the non-shooting hand exhibits instability or makes contact with the ball at the moment of shooting, thereby disrupting the balance and rhythm of the shot.

3.3 Data Handling

Categorized videos must be converted into digital arrays to be utilized in deep learning models. In this study, we do not require full pixel arrays; instead, we extract a set of key human body part data from each video. The Movenet model is available via TensorFlow Hub, and its most notable feature is the balance between “speed and accuracy.” Its highly lightweight architecture enables it to run on ordinary smartphones and tablets, with

per-frame inference latency below 5 ms and frame rates exceeding 50 FPS. Its high accuracy also yields excellent performance on the COCO skeletal keypoint detection dataset. The model offers two variants: the Lightning version, suitable for scenarios requiring real-time response, albeit with slightly lower accuracy; and the Thunder version, designed for professional applications where high pose accuracy is critical, offering superior accuracy at the cost of relatively slower inference speeds. In this study, given that latency requirements were not stringent, while high accuracy contributed to improved classification performance, we adopted the latest Thunder version available at the time of the research.

For each frame, the MoveNet model outputs an array containing 14 key points. Each key point includes information such as its normalized y-coordinate, x-coordinate, and confidence score. These coordinates are typically normalized relative to the input image dimensions (within the range of 0–1), allowing them to be directly mapped back to the pixel coordinates of the original video frame.

When analyzing basketball free-throw techniques, MoveNet accurately captures the coordinates of key body landmarks, enabling the computation of critical angular and distance parameters. First, it calculates joint angles across multiple body segments to gain deep insights into the body's alignment and dynamic movement during the execution of the shot. These angular measurements are crucial for assessing whether an athlete maintains proper biomechanical posture—particularly during the complex technical phases of shooting. Second, it incorporates the inter-joint distances along parallel axes and the torso length; these two metrics are equally vital for posture evaluation. For instance, the distance from the ankle to the wrist can be used to determine the release height and analyze the trajectory of the body flexion phase; an excessively low release height or insufficient body flexion may result in a shot that lands just outside the rim.

MoveNet predicts local offsets—i.e., the subtle displacements of each key point relative to the balance center (the athlete's left and right hips). These data can be used to compute a suite of quantitative metrics that serve as the foundation for scientific training. For example, calculating the elbow joint angle allows for the assessment of the geometric configuration of the shooting arm; calculating the knee joint angle enables the evaluation of squat depth and force generation. In the absence of direct labeling of the basketball's position, the wrist position can serve as a surrogate indicator. By tracking the alignment of the wrist relative to the body center, one can roughly infer the straightness of the basketball trajectory, the release height, and detect deviations that may compromise movement execution. Table 2 lists the complete set of features.

Table 2: Features extracted from pose evaluation, along with additional angular and distance data

0. Tip of the nose	9. Left wrist	18. Right elbow joint angle
1. Left eye	10. Right wrist	19. Left shoulder joint angle
2. Right eye	11. Left buttock	20. Right shoulder joint angle
3. Left ear	12. Right buttock	21. Left knee joint angle
4. Right ear	13. Left knee	22. Right knee joint angle
5. Left shoulder	14. Right knee	23. Distance from the left wrist to the left ankle
6. Right shoulder	15. Left ankle	24. Distance from the right wrist to the right ankle
7. Left elbow	16. Right ankle	25. Distance between the knee joints
8. Right elbow	17. Left elbow joint angle	26. Distance between the ankles

Next, normalize all values in each frame's digital array to the range of 0–1 to avoid negative values.

3.4 Data Augmentation

When applying the keypoint data captured by MoveNet to train a high-precision basketball free-throw analysis model, relying solely on raw data often fails to encompass all possible shooting scenarios. Data augmentation is a critical approach to addressing sample scarcity and enhancing model robustness.

3.5 Excitation

For Gaussian noise, Gaussian noise with a standard deviation of 0.01 is added to the values at key points, thereby introducing a mild degree of random variation.

Regarding jitter, a uniform jitter with a standard deviation of 0.01 is applied to key points, thereby simulating subtle positional variations.

3.6 Time and Space Adjustment

For the scaling operation, the key point values are multiplied by a random factor ranging from 0.9 to 1.1, thereby simulating the variations in the distance between the athlete and the camera.

For the translation operation, key points are adjusted along the horizontal and vertical axes, respectively, to random values within the range of -0.05 to 0.05, thereby simulating the slight movement of an athlete relative to the camera position.

The horizontal flip operation involves mirroring key points along a horizontal axis to ultimately generate a reversed image of the athlete's movement.

As for time distortion manipulation, a distortion factor ranging from 0.8 to 1.2 is applied to slow down or accelerate an athlete's movement speed, thereby creating temporal variations in the movement.

3.7 Enhance the Application of Technology

The dataset is split into a training set and a test set in an 8:2 ratio. The augmentation process for the training set is as follows: first, the target class of each video is determined. If the video belongs to the most prevalent class, "correct posture," no perturbation is applied. For the least representative classes, two perturbation techniques are employed simultaneously to generate two new feature sets for each sample. For all other classes, one of the two perturbation techniques is randomly selected to generate an additional feature set.

After completing the perturbation processing, a randomly selected temporal or spatial adjustment operation is performed for each original feature set and each augmented feature set, thereby doubling the number of feature sets. To ensure data integrity, for each original feature set, no more than two such adjustments are performed, preventing excessive deviations from the original data.

Finally, all computational features mentioned in the previous section—such as angles and distances—are added to the dataset following data augmentation, ensuring that the features in the augmented dataset remain consistent. Additionally, a padding technique is employed to standardize all videos to a uniform number of frames. Given that the maximum number of frames in any video is 400, zero-padded arrays are appended to the end of all other videos to extend their frame count to 400 frames.

4. Features and Advantages of the MoveNet Model

The standout feature of the MoveNet model is its balance of "speed and accuracy":

The MoveNet model possesses a distinct advantage: it achieves an almost perfect balance between speed and accuracy.

Ultra-lightweight design: The model employs an extremely small number of parameters—on the order of only a few million—which enables it to run on conventional smartphones and tablets, as well as on low-power IoT devices. Moreover, its single-frame inference latency is below 5 ms, and it achieves a frame rate exceeding 50 fps [18].

In terms of high precision, despite its compact size, it has demonstrated outstanding performance on the COCO keypoint detection dataset. For instance, the Thunder variant achieved a precision of approximately 78.7% on the COCO Val2017 dataset [19].

4.1 Modeling

Through data processing and augmentation steps, a dataset comprising 1,840 videos was ultimately generated for model training. The researchers trained and evaluated four distinct models on this dataset to assess their performance.

4.2 Convolutional Neural Networks and Long Short-Term Memory Network Models

This model represents a classic framework for spatiotemporal feature extraction. First, the CNN layer employs multiple convolutional kernels (e.g., 3×3 and 5×5) to extract local features from the input image sequence, generating a series of feature maps. These feature maps are then flattened into vector sequences

corresponding to each time step and fed into a bidirectional LSTM (Bi-LSTM). By leveraging gating mechanisms—including the forget gate, input gate, and output gate—the Bi-LSTM captures long-range dependencies across the sequence, producing context-aware feature vectors at each time step. Finally, a fully connected layer is used to perform classification or regression tasks. Batch Normalization and pooling layers are commonly added after the convolutional layers to reduce feature dimensionality, while Dropout layers are incorporated before and after the LSTM to mitigate overfitting. The entire network is optimized via end-to-end backpropagation, enabling deep integration of spatial texture information and temporal dynamics.

4.3 Shuangliu Convolutional Neural Network Model

The Shuangliu Convolutional Neural Network is specifically designed for video action recognition and consists of two parallel branches: a spatial stream and a temporal stream. The spatial stream processes raw RGB frame sequences—typically using deep convolutional networks such as VGG or ResNet—employing multiple layers of convolution, pooling, and residual blocks to extract static texture and object shape features from the images. The temporal stream ingests optical flow maps to capture pixel-level motion direction and velocity, also utilizing stacked convolutional layers to extract temporal change features. The high-level features from both streams are then either concatenated or added at a fully connected layer or a dedicated fusion module (e.g., weighted averaging or non-local attention), thereby forming a unified spatio-temporal representation. Finally, a Softmax classifier outputs the action class.

4.4 Time Convolutional Network Model

The temporal convolutional network is a sequence modeling approach based on one-dimensional full convolution, with its core components being causal convolution and dilated convolution. The network consists of multiple stacked residual blocks, each containing two dilated causal convolution layers. The convolution kernels have fixed sizes, while the dilation factors grow exponentially across layers (1, 2, 4, 8,...), thereby rapidly expanding the receptive field to capture long-range dependencies. Following the convolution operations, the network applies weight normalization, ReLU activation, and Dropout; the main path is then added to a residual path featuring a 1×1 convolution to form an identity mapping, ensuring stable gradient propagation.

4.5 Artificial Neural Network

Artificial neural networks are mathematical models inspired by biological neurons, featuring a hierarchical topology composed of an input layer, one or more hidden layers, and an output layer. Each layer consists of multiple neurons, which are interconnected via weighted directed edges, while typically lacking connections between neurons within the same layer. Information flows in a feedforward direction: the input layer receives external feature vectors; the hidden layers apply nonlinear transformations through weighted summations followed by activation functions (e.g., ReLU or Sigmoid); and the output layer maps the final features to prediction results⁷⁸. The network's learning process adjusts these weights using the backpropagation algorithm (BP), enabling the model to approximate complex nonlinear functions. Such networks are widely applied in tasks such as classification, regression, and feature extraction.

4.6 Hyperparameter

To fully optimize the performance of these four deep learning models, the researchers conducted rigorous hyperparameter tuning for all models. First, given the complexity of the task and its sensitivity to gradient updates, the Adam optimizer was selected as the unified optimization strategy; its efficient adaptive learning rate adjustment mechanism provides robust support for model training. Second, since we are addressing a multi-class classification task involving basketball free-throw movements, the cross-entropy loss function was uniformly adopted to ensure precise classification boundaries. Finally, to prevent overfitting on limited action data, all models incorporated a Dropout layer, which randomly deactivates selected neurons, thereby significantly enhancing the models' generalization capability.

5. Test

5.1 Testing Environment

The methodology of this study was implemented on devices equipped with the Snapdragon 8 Gen 3 for Galaxy chipset and running the Android 14 operating system. The research and development were carried out using the Jupyter Notebook platform, which supports real-time Python code execution. The primary open-source libraries utilized included NumPy, Scikit-learn, and TensorFlow—the core framework for building, training, and running deep learning models.

5.2 Experimental Results on Imbalanced Datasets

As previously mentioned, we trained four distinct deep learning models to predict the correct class of data samples. In the setting of an imbalanced dataset, the performance of these four models exhibited significant differences. Table 3 lists the key performance metrics of each model on the test dataset; overall performance, ranked from best to worst, was as follows: the dual-stream convolutional neural network model, the convolutional neural network combined with a long-short memory network model, the temporal convolutional network model, and the artificial neural network model. In terms of core performance metrics, the dual-stream convolutional neural network model demonstrated the optimal overall performance, achieving an accuracy rate of 88.7%—the highest among the four models—which indicates its superior overall classification accuracy for samples under imbalanced data distribution. Its precision and recall rates were 82.3% and 76.5%, respectively, both of which are leading values, demonstrating its exceptional capability in accurately identifying positive samples and ensuring comprehensive coverage of positive samples. The F1 score (calculated using macro-average) was 78.1%, reflecting a balanced trade-off between precision and recall, and likewise ranked first. The logarithmic loss was 0.41, which is relatively low, indicating relatively low prediction uncertainty for the model.

Table 3: Model results on the imbalanced dataset

model	precision	Precision	recall	F1 score	Logarithmic loss
Convolutional Neural Networks and Long Short-Term Memory Network Models	85.2	78.5	70.8	72.3	0.48
Shuangliu Convolutional Neural Network Model	88.7	82.3	76.5	78.1	0.41
Time Convolutional Network Model	82.5	84.2	78.9	80.4	0.39
artificial neural network	75.0	72.1	62.5	65.0	0.68

5.3 Experimental Results on the Balanced Dataset

The study also leveraged the collected dataset and four experimental models to dynamically achieve batch balancing during training using a “balanced generator,” ensuring that each training batch contained an equal number of samples for every class used by the model. This approach enhanced the overall performance of the models and allowed us to evaluate the impact of data balancing. Compared to unbalanced datasets, Table 4 demonstrates that all models achieved significant performance improvements on the balanced dataset. While the relative strengths and weaknesses of each model remained largely consistent, the overall performance gap narrowed. The detailed analysis is as follows: the dual-stream convolutional neural network model maintained its optimal performance, with all evaluation metrics showing substantial improvements and ranking first. Its accuracy rate reached 93.8%, representing an increase of 5.1% over the unbalanced dataset. In contrast, the convolutional neural network and the long short-term memory network models exhibited marked performance gains, with their accuracy rates rising from 85.2% to 91.5%—an improvement of 6.3%.

Table 4: Model results on the balanced dataset

model	precision	Precision	recall	F1 score	Logarithmic loss
Convolutional Neural Networks and Long Short-Term Memory Network Models	91.5	90.8	89.7	89.2	0.28
Shuangliu Convolutional Neural Network Model	93.8	93.2	91.8	91.5	0.22
Time Convolutional Network Model	90.2	91.5	90.3	90.0	0.25
artificial neural network	80.0	79.3	78.8	78.5	0.51

5.4 Cross-experiment Results

In addition to experiments conducted on both imbalanced and balanced datasets, this study also employed a 5-fold cross-validation approach on the data. Consistent with the two prior experimental setups, the experiments first proceeded directly on the imbalanced dataset, followed by testing using a batch-based “balanced generator.” The objective was to validate the stability and generalization capabilities of the models. The overall experimental results were consistent with those obtained from direct testing on the imbalanced dataset, and the performance metrics for each model exhibited minimal fluctuations, indicating high reliability of the results. Tables 5 and 6 present these experimental outcomes. In both the imbalanced and balanced dataset settings using 5-fold cross-validation, the performance rankings of the models remained unchanged: Dual-stream Convolutional Neural Network > Convolutional Neural Network + LSTM Network > Temporal Convolutional Network > Artificial Neural Network. On the balanced dataset, all models demonstrated significant performance improvements with reduced variability ($\pm 0.1\% - \pm 0.7\%$), confirming that data balancing effectively enhances model stability and generalization capability. The Dual-stream Convolutional Neural Network outperformed all other models in both test scenarios, achieving an accuracy of $88.3 \pm 0.7\%$ on the imbalanced dataset and rising to $93.8 \pm 0.5\%$ on the balanced dataset—representing its most robust performance. The Temporal Convolutional Network exhibited the smallest variation in precision, while the Convolutional Neural Network and LSTM Network showed relatively balanced performance across all metrics. The Artificial Neural Network consistently ranked last; notably, on the balanced dataset, it displayed an anomalous F1-score (0.52 ± 0.02), likely indicative of a data entry error, although its overall stability improved compared to the imbalanced setting.

Table 5: Model performance results on the imbalanced dataset using 5-fold cross-validation

model	precision	Precision	recall	F1 score	Logarithmic loss
Convolutional Neural Networks and Long Short-Term Memory Network Models	84.8 ± 0.7	78.1 ± 0.7	70.4 ± 0.8	71.9 ± 0.8	0.50 ± 0.02
Shuangliu Convolutional Neural Network Model	88.3 ± 0.6	81.9 ± 0.7	76.1 ± 0.7	77.7 ± 0.7	0.43 ± 0.02
Time Convolutional Network Model	82.0 ± 0.5	83.9 ± 0.4	78.6 ± 0.5	80.1 ± 0.5	0.40 ± 0.01
artificial neural network	74.4 ± 0.5	71.4 ± 0.6	61.7 ± 0.7	64.3 ± 0.7	0.70 ± 0.02

Table 6: 5-fold cross-validation model results on the balanced dataset

model	precision	Precision	recall	F1 score	Logarithmic loss
Convolutional Neural Networks and Long Short-Term Memory Network Models	91.2 ± 0.5	90.5 ± 0.5	89.8 ± 0.5	89.5 ± 0.5	0.29 ± 0.01
Shuangliu Convolutional Neural Network Model	93.8 ± 0.5	93.2 ± 0.5	92.5 ± 0.5	92.2 ± 0.5	0.23 ± 0.01
Time Convolutional Network Model	90.1 ± 0.4	91.6 ± 0.5	90.9 ± 0.5	90.6 ± 0.5	0.26 ± 0.01
artificial neural network	80.7 ± 0.7	80.0 ± 0.7	79.5 ± 0.7	0.52 ± 0.02	0.52 ± 0.02

Furthermore, Figure 4 displays the confusion matrices of each model in its best-performing experimental setup. The 5-fold cross-validation conducted on 876 balanced samples demonstrated that the temporal convolutional network model achieved the best performance, correctly predicting 826 samples with only 53 errors, exhibiting exceptionally high accuracy in identifying the positive class, and thus representing the most precise choice for this task; the dual-stream convolutional neural network model followed closely, with 804 correct predictions and 72 symmetric errors, demonstrating optimal classification stability and balance; the convolutional neural network combined with a long short-term memory network model achieved 803 correct predictions and 84 errors, showing slightly weaker capability in identifying the negative class, making it suitable for scenarios where temporal characteristics are critical but a higher tolerance for error in negative class classification is acceptable; the artificial neural network model performed the poorest, with only 789 correct predictions and 106 errors, and has been comprehensively outperformed by more sophisticated modern models.

6. Discussion

This study innovatively establishes a complete technical closed-loop comprising “smartphone video capture — motion detection — real-time feedback,” and for the first time, has conducted systematic validation of a

deep learning-based detection solution for basketball free throws—a technical maneuver requiring high precision—on a mobile platform.

The core innovation lies in proposing and validating the advantages of a dual-stream convolutional neural network (CNN) model. This model leverages spatial streams to capture static features of the posture angle—such as knee flexion angle and elbow position—while utilizing temporal streams to capture dynamic temporal features, including the timing of wrist force application and ball-following movements. Consequently, it effectively accommodates the entire sequence of a free throw execution: “preparation → ball release → release → ball following.” In contrast, standalone CNN models, CNN-LSTM models, or TCN models each exhibit limitations in their respective capabilities for extracting spatial or temporal features.

The study further confirms the practicality of the mobile posture estimation model MoveNet Thunder. On Android 14 devices equipped with Snapdragon 8 Gen 3, this model achieves inference latency below 5 ms, enabling real-time extraction of 17 human body key points and 26 quantitative features—thereby successfully addressing the challenge inherent in traditional manual observation, which is predominantly qualitative in nature yet lacks sufficient quantitative support.

For data processing tasks, this study established a dedicated annotation schema and a hierarchical data augmentation strategy. The annotation work was carried out by E-level certified coaches from the Chinese Basketball Association in accordance with international standards, which defined eight categories of errors that align with practical training needs. The data augmentation and balancing strategies effectively mitigate overfitting issues arising from limited sample sizes and class imbalance, thereby generally improving the model's accuracy on balanced datasets by over 5%.

This study validates an effective approach combining deep learning with label-free motion capture for the quantitative analysis of athletic movements; furthermore, it introduces a new paradigm in sports engineering—advancing from “coarse-grained action localization” to “fine-grained error diagnosis”—thereby providing a viable technical reference for the scientific and intelligent optimization of basketball training.

6.1 Boundedness

This study has completed the preliminary validation of a real-time detection scheme for basketball free-throw technique execution. However, several limitations remain in aspects such as study design, sample collection, model architecture, and practical application. Specifically, the controllability of the filming environment was excessively high: video acquisition was primarily conducted within a controlled indoor basketball arena, where filming angles and lighting conditions were relatively fixed. Consequently, the scheme did not account for the complex scenarios encountered in actual training and competitive settings—such as outdoor venues, diverse backgrounds, extreme lighting conditions, or multiple filming angles—meaning the robustness of the model in uncontrolled environments has yet to be validated. Additionally, the pose estimation capability was limited in dimensionality. The study employed the MoveNet 2D pose estimation model, which can only capture planar body pose information and lacks 3D spatial feature representation. As a result, it failed to accurately identify spatial execution errors—such as forward or backward tilt of the body, or forward or backward deviation of the shooting arm—leading to deviations in the quantitative analysis of certain fine-grained movements. Furthermore, the design of the movement error classification framework exhibited a single-focus limitation: it targeted only the primary error types of athletes, without considering the co-occurrence of multiple errors, nor did it classify errors by severity level. Consequently, the model could not provide comprehensive error diagnostics or tiered corrective recommendations, creating a gap between its performance and actual training needs. Regarding practical training validation, the study only conducted performance evaluations of the model's movement detection capabilities; it did not integrate the model into real-world basketball training sessions, nor did it conduct long-term experiments to verify the actual efficacy of the model's feedback in improving free-throw accuracy rates. Thus, the practical value of the model has not yet been rigorously tested in real-world scenarios. The feature pattern design can be further optimized; however, the 26 extracted features primarily consist of joint angles and distances, while excluding basketball-specific features (release speed, rotation angle, flight arc) and athlete physiological characteristics (muscle contraction state). Consequently, the feature pattern lacks comprehensiveness, which may adversely affect the accuracy of error classification.

7. Conclusions and Future Directions

This study has developed a real-time detection and feedback system for basketball free-throw technique using smartphone video, thereby opening up new avenues for science-based training. The main findings are as follows:

7.1 The effectiveness of the data preprocessing was validated.

By employing data augmentation techniques such as Gaussian noise and temporal distortion, combined with a batch balancing generator, the model's accuracy on the balanced dataset increased by over 5%, demonstrating that data preprocessing is crucial for enhancing a model's generalization capability. The optimal detection model was identified: the dual-stream convolutional neural network achieved the best overall performance, attaining an accuracy of $93.8 \pm 0.5\%$ in five-fold cross-validation on the balanced dataset, with precision, recall, and F1-score reaching 93.2%, 92.5%, and 92.2%, respectively. This model also exhibited superior class balance and stability compared to alternative models.

7.2 Complete mobile feasibility validation

In this study, model testing was conducted on Android smartphones to validate the feasibility of running MoveNet and other deep learning models on mobile devices, thereby eliminating the reliance on specialized motion capture equipment. The research demonstrates that leveraging smartphone video and deep learning to achieve real-time detection of free-throw movements is feasible, providing a reference framework for the quantitative analysis of fine motor skills in the field of sports engineering.

7.3 Future Research

The research focuses on two key aspects: characteristic pattern extraction and model deployment. First, it incorporates 3D pose estimation technologies—such as MediaPipe 3D Pose—to enrich spatial feature representation, while integrating video tracking techniques to extract hand movement velocity and rotation angles, thereby establishing finer quantitative dimensions. To enhance detection robustness in complex environments, the study aims to develop a multimodal fusion model that combines smartphone IMU inertial data (e.g., acceleration and angular velocity) with visual features, compensating for limitations in visual feature acquisition. Additionally, the error diagnosis framework will be refined by introducing multi-error superposition classification and severity grading, enabling priority prioritization, coupled with tailored corrective suggestions, thus establishing an integrated guidance system. Regarding system architecture, a device-cloud collaborative framework will be implemented: computationally intensive tasks will be offloaded to the cloud, while the mobile device will focus on data acquisition and result visualization, balancing real-time performance with accuracy, and supporting remote training guidance. To validate the model's effectiveness, it will be integrated into a basketball training application for long-term tracking experiments, evaluating improvements in movement execution and shooting accuracy. Ultimately, starting from free throws, the system will be extended to general basketball movement detection—including shots and layups—and explored for application in precision sports such as badminton and table tennis, leveraging AR/VR technologies to provide visual feedback and enhance the user interaction experience.

References

- [1] Teket O M, Yetik I S. A fast deep learning based approach for basketball video analysis[C]//Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing. 2020: 1-6.
- [2] Khobdeh S B, Yamaghani M R, Sareshkeh S K. Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network: SB Khobdeh et al[J]. The Journal of Supercomputing, 2024, 80(3): 3528-3553.
- [3] He W, Li B. Basketball action recognition by fusing video recognition techniques with an SSD target detection algorithm[J]. Open Computer Science, 2025, 15(1): 20250027.
- [4] Shakya S R, Zhang C, Zhou Z. Basketball-51: A video dataset for activity recognition in the basketball game[C]//CS & IT Conference Proceedings. CS & IT Conference Proceedings, 2021, 11(7).

- [5] Teket O M, Yetik I S. A fast deep learning based approach for basketball video analysis[C]//Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing. 2020: 1-6.
- [6] Khobdeh S B, Yamaghani M R, Sareshkeh S K. Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network: SB Khobdeh et al[J]. The Journal of Supercomputing, 2024, 80(3): 3528-3553.
- [7] He W, Li B. Basketball action recognition by fusing video recognition techniques with an SSD target detection algorithm[J]. Open Computer Science, 2025, 15(1): 20250027.
- [8] Wang J, Zuo L, Cordente Martínez C. Basketball technique action recognition using 3D convolutional neural networks[J]. Scientific Reports, 2024, 14(1): 13156.
- [9] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.
- [10] Weinzaepfel P, Harchaoui Z, Schmid C. Learning to track for spatio-temporal action localization[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3164-3172.
- [11] Teket O M, Yetik I S. A fast deep learning based approach for basketball video analysis[C]//Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing. 2020: 1-6.
- [12] Shakya S R, Zhang C, Zhou Z. Basketball-51: A video dataset for activity recognition in the basketball game[C]//CS & IT Conference Proceedings. CS & IT Conference Proceedings, 2021, 11(7).
- [13] Han Z, Qin W, Wang Y, [13]Han Z, Qin W, Wang Y, et al. MultiSubjects: A multi-subject video dataset for single-person basketball action recognition from basketball gym[J]. Computer Vision and Image Understanding, 2024, 249: 104193.
- [14] Choi J, Bhattacharya S. Deep learning-based real-time smartphone pose detection for ultra-wideband tagless gate[C]//GLOBECOM 2023-2023 IEEE Global Communications Conference. IEEE, 2023: 4595-4600.
- [15] Chung J-L, Ong L-Y, Leow M-C. Comparative Analysis of Skeleton-Based Human Pose Estimation. Future Internet. 2022; 14(12):380. <https://doi.org/10.3390/fi14120380>
- [16] Del Piano C, Leigh S. Basketball Free Throw: A Written Technical Report[J]. Montclair State University, Advanced Coaching Techniques, PEMJ-547, 2013.
- [17] Fenty Y V, Buides R C H, Meriño J A S. Estudio cinemático de la ejecución del tiro libre en los baloncestistas de la categoría 11-12 años: II Taller” Actividad Física y Deporte”[C]//Congreso Universidad. 2025, 11(4): e46-e46.
- [18] Arrowsmith C, Burns D, Mak T [18]Arrowsmith C, Burns D, Mak T et al. Physiotherapy exercise classification with single-camera pose detection and machine learning. Sensors 2022; 23(1): 363.
- [19] Acı N and Kuluoğlu MF. Accuracy detection in some sports training using computer vision and deep learning techniques. Bitlis Eren University Journal of Science and Technology 2023; 13(2): 133–158.

Funding

This work was supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN-202502909).

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).