Image and Vision Computing 162 (2025) 105689

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Advancements in basketball action recognition: Datasets, methods, explainability, and synthetic data applications

Marco Caruso [a], Lucia Cimmino [a,*], Fabio Narducci [a], Chiara Pero [b], Gianluca Ronga [a]

[a] Department of Computer Science, University of Salerno, via Giovanni Paolo II, 132, Salerno, 84084, Italy
[b] Department of Management & Innovation Systems, University of Salerno, via Giovanni Paolo II, 132, Salerno, 84084, Italy

ARTICLE INFO

Keywords:
Basketball Action Recognition (BAR)
Synthetic data
Explainable AI (XAI)
Survey
Sports analytics

ABSTRACT

Basketball Action Recognition (BAR) has received increasing attention in the fields of computer vision and artificial intelligence, serving as a fundamental component in performance evaluation, automated game annotation, tactical analysis, and referee decision-making support. Despite notable advancements driven by deep learning approaches, BAR remains a challenging task due to the inherent complexity of basketball movements, frequent occlusions, and limited availability of standardized benchmark datasets. This survey provides a comprehensive and structured synthesis of current developments in BAR research, encompassing four principal dimensions: dataset curation, computational methodologies, synthetic data generation, and model explainability. A critical analysis of publicly available basketball-specific datasets is presented, delineating their modalities, annotation strategies, action taxonomies, and representational scope. Furthermore, the survey offers a structured classification of state-of-the-art action recognition methodologies, ranging from video-based and skeleton-based models to sensor-driven and multimodal fusion approaches, emphasizing architectural characteristics, evaluation protocols, and task-specific adaptations. The role of synthetic data is systematically examined as a means to address data scarcity, reduce annotation noise, and enhance model generalization through controlled variability and simulation-based augmentation. In parallel, the integration of explainable artificial intelligence (XAI) techniques is also analyzed, with a focus on post-hoc attribution methods, probabilistic reasoning models, and interpretable neural architectures, aimed at improving the transparency and accountability of decision-making processes. The survey identifies persisting research challenges, including dataset heterogeneity, limitations in cross-domain transferability, and the accuracy-interpretability trade-off in deep models. By delineating current limitations and prospective directions, this work provides a foundational reference to guide the development of robust, generalizable, and explainable BAR systems for deployment in real-world sports intelligence applications.

1. Introduction

Basketball Action Recognition (BAR) has become a critical research area in computer vision and artificial intelligence, with applications spanning performance analysis, automated game annotation, tactical decision-making, and referee assistance. The ability to automatically recognize and classify basketball-specific actions in video sequences is essential for modern sports analytics, enabling more precise player evaluations, injury prevention strategies, and enhanced broadcasting experiences. However, basketball presents unique challenges for action recognition due to its dynamic nature, rapid player movements, frequent occlusions, and the fine-grained distinctions between different types of actions. These complexities necessitate the development of robust and efficient computational models capable of learning both spatial and temporal patterns in video data [1,2]. A fundamental driver of progress in BAR research is the availability of large-scale, high-quality datasets. Datasets serve as the basis for training, validating, and benchmarking action recognition models, yet their quality and diversity greatly influence model performance. Despite the growing interest in sports action recognition, there remains a lack of standardized and comprehensive datasets specifically tailored to basketball, making model generalization across different scenarios particularly challenging. Given the increasing reliance on data-driven deep learning approaches, a structured review of existing basketball action recognition datasets is essential to assess their strengths, limitations, and applicability to different research problems.

* Correspondence to: University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy.
E-mail addresses: carusomarco122@gmail.com (M. Caruso), lcimmino@unisa.it (L. Cimmino), fnarducci@unisa.it (F. Narducci), cpero@unisa.it (C. Pero), g.ronga5@studenti.unisa.it (G. Ronga).

https://doi.org/10.1016/j.imavis.2025.105689
Received 11 March 2025; Received in revised form 30 June 2025; Accepted 29 July 2025
Available online 7 August 2025

From a methodological perspective, BAR has evolved significantly with the advent of deep learning techniques. Traditional approaches relied on handcrafted feature extraction and classical machine learning models, which often struggled to generalize across different game scenarios due to the variability in player movements and court conditions. The introduction of deep convolutional neural networks (CNNs), recurrent neural networks (RNNs), and spatiotemporal architectures such as 3D-CNNs, Long Short-Term Memory (LSTM) networks, and, more recently, transformer-based models, has led to substantial improvements in recognition accuracy. These models leverage both spatial and temporal dependencies within video sequences, enabling them to capture fine-grained details of basketball movements. Additionally, multi-modal learning approaches that integrate skeletal data, optical flow, and pose estimation have shown promise in further enhancing recognition performance.

Despite advancements in model performance, a major challenge remains in the interpretability of deep learning-based action recognition models. The black-box nature of these architectures limits their adoption in high-stakes environments where explainability is crucial, such as in professional sports coaching, referee decision-making, and player scouting. To address this, explainable artificial intelligence (XAI) techniques have gained increasing attention in sports analytics, aiming to provide transparency into model decisions. By employing visualization techniques such as attention maps, feature attribution methods, and model interpretability frameworks, XAI can enhance trust in automated systems and facilitate human-in-the-loop decision-making. While various datasets, methodologies, and techniques have been developed in the broader domain of action recognition, their applicability and effectiveness in basketball remain an area of active exploration. Additionally, emerging approaches such as synthetic data generation and explainable AI are being increasingly investigated, yet their potential impact on BAR is still evolving. This work provides a structured examination of key aspects of BAR research, including datasets, synthetic data applications, methodological advancements, and explainability strategies. By reviewing these areas, we aim to present an overview of recent progress, highlight existing challenges, and discuss potential directions for future studies in BAR.

1.1. Motivations and contributions

The development of BAR models is influenced by multiple factors, including the availability of well-structured datasets, advancements in deep learning methodologies, and the growing need for model interpretability. Despite progress in action recognition, several challenges remain in BAR, particularly concerning dataset standardization, the integration of synthetic data, and the explainability of deep learning models. Addressing these challenges requires a systematic evaluation of existing resources and methods to better understand their limitations and potential improvements.

Although action recognition in sports and general human activities has been the subject of several surveys [3–9], BAR has received limited dedicated attention. Some surveys address sports only marginally within the broader context of human action recognition, while others focus on sports more explicitly but consider basketball merely as one of many examples. In both cases, BAR is often mentioned without a focused analysis of its specific challenges and requirements. In particular, a systematic comparison of BAR-oriented datasets, modeling approaches, and interpretability techniques, tailored to the sport's distinctive spatio-temporal dynamics, is still lacking. Furthermore, prior surveys rarely address vision-based applications that are unique to basketball, such as performance analytics, tactical decision-making, player monitoring, and automated officiating support. This survey fills this gap by providing a comprehensive overview specifically focused on BAR, encompassing datasets, synthetic data generation, learning architectures, explainability strategies, and basketball-specific computer vision tasks. In doing so, it establishes a structured foundation for future research in this domain.

The main contributions of this work are as follows:

- A systematic review of Basketball Action Recognition Datasets: A structured analysis of existing datasets is conducted, focusing on their characteristics, annotation strategies, and applicability to BAR tasks.
- Investigation of Synthetic Data in BAR: The role of synthetic data in addressing dataset limitations is examined, highlighting its benefits and challenges for model generalization and robustness.
- Comparison of Action Recognition Techniques: State-of-the-art deep learning approaches applied to BAR are systematically reviewed, emphasizing key architectural choices, training methodologies, and performance evaluation.
- Examination of Explainability in BAR: The integration of explainable AI (XAI) techniques into BAR models is discussed, assessing their role in improving transparency and trust in AI-driven sports analytics.
- Identification of Research Gaps and Future Directions: Open challenges in BAR research are outlined, including dataset inconsistencies, methodological gaps, and the need for improved benchmarking practices, with a discussion on promising directions for future studies.

This survey provides a structured and detailed examination of key aspects in BAR, offering insights into existing methodologies while outlining critical challenges and research opportunities.

Overview of survey scope. This survey reviews 13 basketball-related datasets, including both public and restricted-access sources. These datasets differ in size, modality (i.e., RGB, depth, skeleton, textual annotations), frame rate, resolution, and the number of annotated action classes, which range from basic movements to complex team interactions. The survey also examines 3 synthetic sports datasets with high-quality annotations and controlled environments: SwimXYZ (swimming, for pose estimation), NBA2K (basketball, for 3D reconstruction), and SoccerSynth-Detection (soccer, for object detection under varying conditions). Of these, only NBA2K is directly focused on basketball, while the others illustrate transferable methodologies and support cross-sport generalization. To classify methodological approaches, the survey analyzes 11 representative human action recognition models applied to basketball, grouped into four categories: video-based, skeleton-based, sensor-based, and multimodal. Finally, 9 explainable action recognition techniques are reviewed, including post-hoc methods (i.e., Grad-CAM, Video-TCAV), probabilistic models (i.e., Bayesian Neural Networks, Probabilistic Graphical Models, Sum–Product Networks), and hybrid interpretable architectures (i.e., Dynamic Conditional Cutoff Networks). While none of these were specifically designed for basketball, a gap noted in current literature, they offer flexible frameworks that can be extended to domain-specific applications.

2. Basketball-related datasets

Datasets are essential for training and evaluating deep learning models, which typically demand large-scale data. In the domain of basketball action recognition, significant efforts have been made to develop specialized datasets. The construction of such datasets generally comprises three main phases: (1) defining categories of basketball actions; (2) collecting video recordings from diverse sources, including online platforms and self-recorded footage; and (3) processing and annotating the collected videos. This section provides a comprehensive review of existing basketball-related datasets, as summarized in Table 1. Furthermore, Fig. 1 illustrates sample images from currently available and accessible datasets.

The **UCF11** dataset [10] comprises 1,160 videos of various sports, organized into 11 categories. The category related to basketball is labeled "basketball shooting". The videos exhibit several characteristics valuable for action recognition: they typically feature the same players, similar backgrounds, and consistent camera positioning. However, the

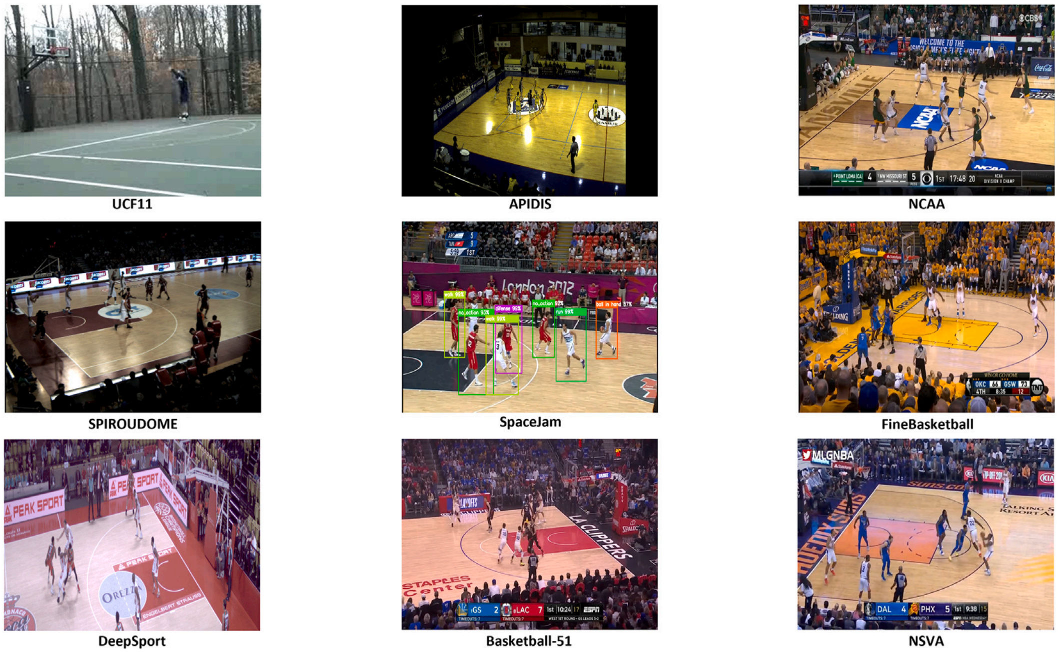

Fig. 1. Representative examples from publicly available basketball-related datasets.

Table 1
Summary of publicly available (P.A.) and non-available (N.A.) basketball-related datasets. The table provides an overview of key characteristics, including dataset type, number of action classes, video count, resolution, data modalities, framerate, available annotations, and additional relevant information.

Dataset	Year	Type	n. Classes of actions	n. of videos	Resolution	Acquired data	Framerate	Annotations	Additional information
UCF11 [10]	2009	P.A.	11	1160	–	RGB	–	Video-level categorical labels	–
APIDIS [11,12]	2008	P.A.	–	7	1600 × 1200	RGB	24	Bounding box, timestamp, ball position	One game, seven camera angles
Basket-1, Basket-2 [13]	2016	N.A.	4	13	–	RGB	25	Timestamp	–
NCAA [14]	2016	P.A.	11	14 548	–	RGB	–	Bounding box	–
SPIROUDOME [15]	2017	P.A.	–	8	1600 × 1200	RGB	25	Bounding box, timestamp	–
SpaceJam [16]	2018	P.A.	10	15	–	RGB, skeleton	–	Bounding box	–
FineBasketball [17]	2020	N.A.	26	3399	–	RGB	–	Frame-level categorical labels	–
NPU RGB+D basketball [18]	2021	P.A.	12	2169	1280 × 720	RGB, Depth maps, skeleton	60	Video-level categorical labels, frame-level depth maps, skeleton coordinates	Amateur players, not real games
DeepSport [19]	2021	P.A.	–	672	–	RGB	–	Ball position, player pose	–
Basketball-51 [20]	2021	N.A.	8	10 311	320 × 240	RGB	–	Video-level categorical labels, timestamp-based annotations	Broadcasted videos, with on-screen info
NSVA [21]	2022	P.A.	124	32 019	1280 × 720	RGB, Text	–	Video-level captions, player identities	–
MultiSubjects [22]	2024	P.A.	3	6144	1920 × 1080	RGB, skeleton	–	Frame-level action labels, player identities, bounding boxes, and skeleton keypoints	Four cameras on one court, amateur players

dataset has limited applicability due to its lack of diversity in elements, which fails to capture the complexity of basketball gameplay.

The **APIDIS** dataset [11,12] is an RGB dataset obtained from seven synchronized video recordings of the same basketball game, captured using Arecont Vision AV2100 cameras at 24 frames per second with a resolution of 1600 × 1200 pixels. A top-view perspective was achieved using Fujinon wide-angle lenses, with each camera positioned differently around and above the court. Frames are synchronized on a central server, which provides bounding boxes for player and ball positions. Each frame is annotated with the action type, time instant, and involved players. Events are classified as "clock" or "non-clock" based on their impact on the 24-second shot clock; for example, a defensive rebound is a clock event, while a pass is a non-clock event. The dataset supports action recognition tasks such as shooting, passing, and rebounding. However, its low video contrast limits its effectiveness for advanced player recognition tasks.

The **Basket-1** and **Basket-2** datasets [13] were generated by synchronizing two groups of cameras. The first group, comprising six cameras, recorded 4,000 frames, while the second group, with seven cameras, recorded 3,000 frames. All cameras captured video at 25 frames per second and were synchronized to enable precise timestamping of each frame. The dataset categorizes actions into four primary classes: possessing the ball, passing, ball in flight, and ball out of bounds. This structure makes it suitable for action recognition and for inferring the ball's status during different phases of the game. It is important to note that the dataset is not publicly available.

The **NCAA** dataset [14] contains recordings from the North American college basketball league, comprising 257 videos, each averaging 1.5 h in length. The dataset includes eleven action categories, resulting in 14,548 clips of varying durations extracted from the original videos. Additionally, 9,000 frames are annotated with bounding boxes for player detection. This dataset is particularly valuable for developing models that analyze interactions between multiple players in the same action. However, the absence of ball position annotations limits its applicability for in-depth action recognition and player-game interaction analysis.

The **SPIROUDOME** dataset [15] is structurally similar to the APIDIS dataset [11,12]. It captures a single basketball game from eight different angles, with four cameras positioned on each side of the court. Each camera records at approximately 25 frames per second (fps) with a resolution of 1600 × 1200 pixels, and frames are stored as .pgm files. The dataset includes timestamps for each frame and provides player positions via bounding boxes. While it can be useful for player detection, it lacks annotations for more complex actions, limiting its suitability for broader action recognition tasks.

The **SpaceJam** dataset [16] consists of 15 videos of NBA and Italian LBA games, each averaging 90 min, as they include full game footage with pauses. The dataset is RGB-based and provides estimated player positions during their movements on the court. It classifies actions into ten categories: step, run, block, passing, catching the ball, shooting, positioning, walking, defensive position, and no action. Additionally, it includes a component called the “joint dataset”, which provides player coordinates on still images, enabling its use for skeleton-based recognition models and time-referenced action analysis. Despite its relatively small size, the dataset offers opportunities for detailed player movement analysis.

The **FineBasketball** dataset [17] is structured on two hierarchical levels. The first level categorizes actions into three primary classes: passing, dribbling, and shooting. The second level provides a more granular classification, recognizing specific actions such as one-hand passing, between-the-legs dribbling, and layup shooting, resulting in a total of 26 action subclasses. The dataset comprises 3,399 clips, averaging 130 clips per subclass. However, the distribution of clips across categories is uneven, reducing its effectiveness for recognizing certain actions. All clips are sourced from broadcasted games, which introduce complex backgrounds filled with spectators and unrelated objects. Additionally, the videos vary in perspective and include different camera movements, making action recognition more challenging.

The **NPU RGB+D Basketball** dataset [18] was created by Northwestern Polytechnical University in China. It contains 2,169 videos of amateur players performing various basketball actions, such as passing and shooting. The videos were recorded using a ZED camera at a resolution of 1280 × 720 pixels and 60 frames per second. The dataset includes twelve action categories: chest pass, over-the-head pass, side pass, squat, one-hand shot, standing dribble, front dribble, running dribble, moving dribble, between-the-legs dribble, behind-the-back dribble, and spin move (turnaround dribble). The dataset offers RGB images, depth maps, and player skeleton data, making it suitable for training different types of models. However, since the actions are performed by amateur players, using this dataset in conjunction with others may be problematic. Datasets featuring professional players often include disturbance elements such as referees and audiences, which could play a crucial role in model training and may not align well with the NPU RGB+D dataset.

The **DeepSport** dataset [19] comprises 672 frames captured from professional-level basketball games. The dataset was curated to represent various lighting conditions encountered in different arenas. Each image, with resolutions ranging from 2 to 5 megapixels, captures half of the basketball court. Due to variations in camera distance from the court, image resolutions range from 65 pixels per meter (px/m) (when taken from low-resolution cameras placed far from the court) to 265 px/m (from high-resolution cameras positioned closer to the action). There is a 40-millisecond delay between successive frames, allowing for temporal analysis. The dataset includes comprehensive annotations: ball position (380 masks) and player poses (5,500 masks). This dataset is highly suitable for detailed player tracking and pose estimation tasks.

The **Basketball-51** dataset [20] comprises 10,311 clips extracted from 51 full NBA broadcasted games. The original broadcasts included highlights, on-screen graphics, interviews, and advertisements, resulting in clips that may contain various disturbance elements. To facilitate experimental use, the video dimensions were reduced from HD to QVGA (320 × 240 pixels), improving computational efficiency. The dataset includes eight action classes: two-point miss, two-point make, three-point miss, three-point make, free throw miss, free throw make, mid-range shot miss, and mid-range shot make. Each clip has a fixed length of six seconds, with the primary action centered in the middle of the clip, providing a suitable time window for action evaluation. This dataset is well-suited for shot classification and performance analysis.

The **NSVA** dataset [21] contains over 32,000 clips from 132 NBA games played during the 2018–2019 season by ten different teams. Each clip is paired with a detailed textual description of the action. The videos are recorded with a full audience background, making the dataset more challenging for model training, particularly for player detection and recognition tasks. The textual annotations provide player involvement, identifying which players participated in the action; action description, describing the specific play (i.e., a blocked shot or a turnover); and action outcome, indicating results (i.e., shot made or missed). To avoid labeling conflicts, such as when a single clip could be interpreted differently depending on the perspective (i.e., a blocked shot counted as a “turnover” for the offensive team or a “steal” for the defensive team), the text annotations are carefully recombined. Non-essential information, such as cumulative statistics (i.e., total assists by a player), is excluded to focus on action-specific details. This dataset is highly versatile, supporting tasks such as action recognition, player detection, and player identification.

The **MultiSubjects** dataset [22] differs from professional game datasets, as it is derived from continuous recordings of a gymnasium where amateur players of various skill levels performed basketball actions. Four 1080p cameras, positioned around the court, recorded gameplay from 7:00 to 21:00, resulting in uninterrupted footage. The dataset contains 6,144 clips, categorized into three primary action classes in a 3:1:2 ratio: dribbling, layup, and shooting. The dataset includes skeleton information, providing 33 three-dimensional (3D) keypoints per player, each with X, Y, and Z coordinates, enabling detailed player pose analysis. Additionally, bounding boxes are provided for player detection. The dataset captures actions performed by 1,000 subjects, contributing to its diversity. The presence of multi-angle views and extensive skeleton annotations makes it valuable for single-player action recognition and pose estimation tasks.

3. Sports-related synthetic datasets

The use of synthetic data in the field of sports science has gained significant attention due to the increasing demand for large-scale, diverse, and well-annotated datasets, especially for AI-driven methodologies in sports analytics applications. The acquisition of a sufficient number of biological data from real-world scenarios and training sessions poses significant challenges and suffers from limitations such as invasive collection, inconsistent annotations, occlusions, and privacy issues. Furthermore, collecting and labeling this data is both a time-consuming and expensive activity. To address these challenges, synthetic data generation has emerged as a valuable solution, providing controlled, scalable, and customizable datasets that can enhance various machine learning and deep learning applications in sports.

One of the most critical areas where synthetic datasets have demonstrated their utility is in image and video-based applications, particularly in Human Pose and Shape (HPS) estimation. Recent studies have shown that synthetic data can supplement or even completely substitute real-world images in training machine learning models for pose estimation and action recognition [23].

In the context of basketball, synthetic datasets have been utilized to model player movements, generate realistic game scenarios, and refine tracking algorithms [24]. Such datasets are particularly valuable for training pose estimation models where real-world annotations are scarce or difficult to obtain. To the best of our knowledge, only a limited number of studies have introduced synthetic datasets specifically designed for sports applications. These studies rely primarily on simulations based on physics, motion capture data, and generative

Table 2
Overview of publicly available synthetic datasets related to sports activities. The table summarizes key characteristics, including data type, target sport, annotations, resolution, and dataset size.

Dataset	Year	Type and source	Sport	Data	Purpose	Annotations	Resolution	# Images/Videos
SwimXYZ [23]	2023	Publicly Available	Swimming	2D/3D Videos	Human motion analysis and pose estimation	3D joints and 2D ground truth	Not specified	11,520 videos
NBA2K [24]	2020	Publicly Available	Basketball	3D Images, 3D Meshes, Textures, 3D Poses, Camera Parameters	3D reconstruction and pose estimation	3D body meshes, joint positions, camera parameters	Not specified High resolution	27,144 images
SoccerSynth-Detection [29]	2025	Publicly Available	Soccer	3D Images	Soccer player detection	Players and balls bounding boxes	1920 × 1080 (clear images) 800 × 600 (motion blurred)	37,621 clear images 17,052 motion-blurred

models to create realistic representations of athletes [25]. Furthermore, advancements in generative adversarial networks (GANs) and neural rendering techniques have significantly improved the photorealism of synthetic images, bridging the gap between synthetic and real-world datasets [26–28].

The study presented in [27] introduces SynthAct, a framework designed to improve human action recognition by leveraging synthetic datasets. SynthAct focuses on generating diverse synthetic human actions to enhance generalization across multiple real-world datasets. The study highlights how domain adaptation techniques, such as adversarial learning and self-supervised learning, can reduce the domain gap between synthetic and real-world human motion data. This approach is particularly relevant in sports analytics, where variations in camera angles, lighting, and occlusions significantly affect action recognition models. Similarly, ElderSim [28] presents a synthetic data generation platform for human action recognition in eldercare applications. While not explicitly designed for sports, ElderSim demonstrates how synthetic human motion datasets can improve the robustness and generalization of machine learning models. The study provides valuable insights into the creation of synthetic datasets with realistic biomechanical constraints, which can be extended to sports-related applications, particularly in injury prevention and rehabilitation analytics.

Based on the growing interest in using synthetic data for sports applications, this study identifies three ready-to-use and publicly available datasets from the current literature: SwimXYZ [23], the NBA2K Dataset [24], and SoccerSynth- Detection [29]. These datasets address specific challenges in different sports disciplines and provide valuable resources for training and evaluating AI models in synthetic data-driven research. Below the key insights and characteristics of each dataset:

- **SwimXYZ** [23]: is a large-scale synthetic dataset specifically designed for swimming-related human motion analysis and pose estimation. It includes a vast collection of synthetic 2D and 3D videos depicting various swimming styles, generated using physics-based simulations and computer graphics techniques. The dataset provides detailed annotations such as 3D joint positions, 2D ground truth keypoints, and segmentation masks. By leveraging realistic fluid dynamics models, SwimXYZ enables the study of human movement in water environments, an area where obtaining real-world labeled data is particularly challenging. This dataset has been widely used for training deep learning models aimed at improving underwater human pose estimation and biomechanical analysis of swimming motions.
- **NBA2K Dataset** [24]: The NBA2K dataset is a synthetic dataset derived from the popular basketball video game NBA2K19, containing high-resolution 3D images, meshes, and annotated player poses. It serves as a crucial resource for research in player tracking, action recognition, and body shape estimation. The dataset consists of detailed representations of basketball players, including skeletal joint positions, motion sequences, and camera parameters. By utilizing the highly detailed graphics engine of NBA2K19, researchers can generate a diverse set of scenarios, covering different player movements, shot attempts, and defensive maneuvers. The dataset has proven particularly useful for training pose estimation algorithms where obtaining precise annotations from real basketball footage is difficult due to occlusions and variations in camera angles.
- **SoccerSynth-Detection** [29]: is a large-scale synthetic dataset tailored for soccer player detection and tracking. This dataset comprises high-resolution 3D images depicting various in-game scenarios, including player formations, ball trajectories, and defensive setups. The dataset provides extensive annotations, including bounding boxes for players and the ball, enabling the development of robust object detection models for soccer analytics. SoccerSynth-Detection also includes motion-blurred images to simulate real-world conditions where players are in rapid movement, offering a more comprehensive training set for deep learning models. By leveraging computer-generated imagery (CGI) and physics-based animations, this dataset allows researchers to train models that can generalize effectively to real-world soccer games, addressing challenges such as occlusion, lighting variations, and viewpoint changes.

A comparative summary of the synthetic datasets for sports applications is provided in Table 2, highlighting key characteristics such as data type, sport, and primary use cases.

Despite the advantages, challenges remain in ensuring that synthetic data accurately captures the complexities of real-world gameplay. Domain adaptation techniques, such as adversarial training and style transfer, have been widely adopted to mitigate the domain gap between synthetic and real images. Additionally, methods such as Synthetic Training for Real Accurate Pose and Shape (STRAPS) [30] and SynthAct [27] have been developed to improve the generalization of synthetic datasets for human action recognition. A comprehensive review of synthetic dataset generation techniques for computer vision applications [31] provides valuable insights into methodologies applicable to sports analytics.

In addition to dataset-specific characteristics, it is essential to consider the broader ethical and methodological implications associated with the use of synthetic data in action recognition. While synthetic datasets provide a scalable and controllable alternative to real-world data collection, their use entails a number of ethical and methodological considerations. A key challenge lies in the realism gap: synthetic environments often fail to replicate the physical variability, contextual unpredictability, and sensor-related noise found in real-world scenarios. As a result, models trained exclusively on synthetic data may exhibit limited generalizability when deployed in unconstrained or heterogeneous settings, unless domain adaptation techniques such as adversarial alignment or feature disentanglement are employed [27,32] Another critical concern is the risk of demographic and behavioral bias. Synthetic data generation pipelines—particularly those based on video game engines or motion capture templates, can inadvertently propagate biases if they rely on narrow archetypes (i.e., male professional athletes, standardized body morphologies), which in turn may lead to systematic performance degradation on underrepresented groups or movement styles [24,31,33]. As highlighted in recent surveys, such representational biases can be amplified by downstream learning algorithms if not explicitly mitigated [34]. These considerations underscore the need to approach synthetic data with a well-informed perspective on its trade-offs—balancing performance gains with ethical and representational concerns.

Table 3
Comparative summary of representative Human Action Recognition methods in Basketball, categorized by data modality.

Method	Type	Dataset	Accuracy	F1-Score	Year
3D CNN + LSTM	Video-Based	NTURGB+D, Basketball-Action-Dataset, B3D Dataset	92%, 88%, 90%	91%, 87%, 89%	2024
C3D + DRAM	Video-Based	BTAD[a]	97.82%	–	2023
LSTA + TSCI	Video-Based	SpaceJam, Basketball-51	89.26%, 92.05%	89.78%, 89.68%	2024
ST-GCN	Skeleton-Based	Kinetics, NTURGB+D	72.4%, 88.3%	–	2018
RMPE + VP3d + ST-GCN	Skeleton-Based	UCF101	88.3%	–	2022
LSTM-DGCN	Skeleton-Based	NTURGB+D	94.6%	–	2021
IMU-HAR	Sensor-Based	IMPACT	96.9%	96.9%	2025
Kinect + PLSTM	Sensor-Based	VRBasketball	93.75%	–	2020
YOLO + Fuzzy LSTM	Multimodal-Based	SpaceJam, Basketball-51	96.45%, 80.23%	96.84%, – %	2023
YOLO-T2FLSTM	Multimodal-Based	BTAD[a]	99.3%	– %	2024
YOLOSS	Multimodal-Based	BPAD[b]	–	–	2024

[a] Basketball Technical Action Dataset.

[b] Basketball Player Action Dataset. The original paper reports only mAP@50 = 82.4%. Accuracy and F1-score are not provided.

4. Action recognition methods

Human action recognition in basketball has emerged as a pivotal research area, driven by the need for real-time game analysis, athlete performance evaluation, and advanced training methodologies. The complexity of basketball movements, including shooting, dribbling, and defensive maneuvers, presents significant challenges due to rapid motion dynamics, occlusions, and varying camera perspectives. While early methods relied on sensor-based approaches, recent advancements in computer vision and deep learning have led to the development of more robust and accurate video-based recognition systems. Among these, 3D convolutional neural networks (3D CNNs) have been widely employed to capture the spatio-temporal features of basketball actions, effectively modeling both spatial appearance and motion dynamics [35]. Further refinements have been introduced through dynamic residual attention mechanisms, which enhance feature extraction across spatial and temporal dimensions [36]. Additionally, adaptive context-aware networks have been proposed to address challenges related to subtle action variations and occlusions by dynamically integrating contextual information, thereby improving recognition accuracy [37].

Beyond video-based methods, skeletal action recognition leverages human joint dynamics for activity classification. Yan et al. [38] introduced a graph-based approach to model spatial and temporal relationships between joints, achieving state-of-the-art performance. Expanding on this concept, the work in [39] integrates additional components to enhance action discrimination in complex scenarios. Similarly, Ma et al. [18] explore optimization strategies and deep learning architectures to improve recognition robustness and accuracy; multimodal approaches further enhance action recognition. In [40], Inertial Measurement Unit (IMU) data is combined with video analysis to improve robustness in scenarios with occlusions and cluttered backgrounds. A novel contribution in [41] integrates virtual reality (VR) simulation, creating an immersive environment that replicates realistic basketball scenarios. This VR-based approach facilitates interactive training while providing real-time feedback, representing a significant advancement in sports performance enhancement. These methodologies illustrate the evolving landscape of human action recognition and the diverse strategies developed to address the specific challenges of basketball action analysis. Recently, hybrid methods that combine multiple modalities, such as RGB, pose, and sensor data, have gained increasing attention due to their capacity to integrate complementary information, thereby improving robustness in real-world scenarios. The following subsections provide a detailed examination of video-based, skeleton-based, sensor-based, and multimodal-based approaches. A comparative overview of representative methods across these categories is reported in Table 3.

4.1. Video-based recognition models

Video-based recognition methods have gained prominence in human action recognition due to their ability to capture spatio-temporal information from sequential frames. In basketball, these methods are particularly effective for analyzing dynamic movements such as shooting, dribbling, and defensive maneuvers by processing entire video sequences rather than isolated frames. A widely adopted approach involves three-dimensional convolutional neural networks (3D CNNs), which extend traditional 2D CNNs by incorporating temporal information. In [35], a 3D CNN architecture specifically designed for basketball action recognition is introduced. The model is trained on a curated dataset of standardized basketball action clips and integrates multiple 3D convolutional layers with pooling, dropout, and batch normalization to extract hierarchical spatio-temporal features. Experimental results show a 9% improvement in accuracy over a baseline 2D CNN, highlighting the effectiveness of 3D convolution in distinguishing subtle action variations under challenging conditions such as lighting changes and complex backgrounds. In [36], the authors refine conventional C3D networks, which often struggle with identifying key frames and extracting discriminative features. The proposed model integrates dynamic residual convolution and an enhanced attention mechanism. A median filter is applied to video frames to reduce noise and background clutter, focusing feature extraction on player motion. The dynamic residual convolution layers enhance spatio-temporal feature learning, while the attention mechanism assigns higher weights to key frames, improving action differentiation. The model achieves an average recognition accuracy exceeding 97%, significantly outperforming standard C3D networks, particularly in scenarios with varying lighting, backgrounds, and camera angles. The study in [37] introduces ACA-Net, an Adaptive Context-Aware Network designed for basketball action recognition. To address challenges such as complex backgrounds, subtle action variations, and varying lighting conditions, ACA-Net integrates a long short-term adaptive (LSTA) module with a triplet spatial-channel interaction (TSCI) module, built on a ResNet50 backbone. The LSTA module captures short- and long-term temporal dependencies, while the TSCI module enhances feature representation by modeling spatial-channel interdependencies. Experiments on benchmark datasets validate ACA-Net's effectiveness. On the SpaceJam dataset (37,085 video clips from NBA and Italian championship games), ACA-Net achieves an accuracy of 89.26%, surpassing state-of-the-art models such as C3D and SlowFast. On the Basketball-51 dataset (10,311 clips from 51 NBA games), ACA-Net reaches an even higher accuracy of 92.05%. Ablation studies demonstrate that integrating both LSTA and TSCI modules yields the best performance. These advancements highlight the increasing sophistication of deep learning models in basketball action recognition. By incorporating spatio-temporal feature learning, attention mechanisms, and adaptive context modeling, these approaches significantly improve classification accuracy and robustness, supporting applications in performance analysis, training, and game strategy optimization.

4.2. Skeleton-based recognition models

Skeleton-based methods utilize pose estimation algorithms (i.e., OpenPose [42,43]) to extract joint coordinates, which are then processed by Graph Convolutional Networks (GCNs) and Long Short-Term Memory (LSTM) networks to model movement dynamics while minimizing background interference. A notable approach is the Spatial–Temporal Graph Convolutional Network (ST-GCN), which represents skeleton sequences as spatial–temporal graphs, where nodes correspond to body joints and edges encode anatomical and temporal relationships [38]. Unlike conventional convolutional networks, ST-GCN applies convolution operations to graph structures, employing adaptive weighting mechanisms to prioritize informative joint connections. Additionally, spatial partitioning strategies categorize joints based on their proximity to the body's center, while temporal convolutions capture motion dependencies across frames. ST-GCN has demonstrated strong performance in action recognition. On the Kinetics dataset (300,000 video clips, 400 action classes), it achieved 30.7% top-1 accuracy and 52.8% top-5 accuracy. On NTU-RGB+D, a 3D joint dataset, it reached 81.5% cross-subject accuracy and 88.3% cross-view accuracy, underscoring its effectiveness for basketball action recognition. In [39], researchers propose a three-dimensional action recognition framework for basketball training, integrating human pose estimation with deep neural networks. The method first employs the Regional Multi-Person Pose Estimator (RMPE) to extract precise two-dimensional (2D) skeleton coordinates from video sequences, ensuring accurate joint localization. These 2D coordinates are then processed using VideoPose3D, a fully convolutional model with time-dilated convolutions that reconstructs three-dimensional (3D) poses, enhancing spatial representation and mitigating issues such as vanishing gradients. Finally, the reconstructed 3D skeletons are input into an ST-GCN, where joints are represented as graph nodes and edges capture spatial–temporal relationships. This model effectively learns motion patterns and classifies basketball-specific actions, outperforming traditional recognition methods while providing enhanced interpretability. Tested within a flipped classroom training model, the framework achieved Top-1 and Top-5 accuracy scores of 42.21% and 88.77%, surpassing the Kinetics-skeleton dataset by 10.61% and 35.09%, respectively. Cross-View and Cross-Subject Accuracy reached 88.3% and 81.5%, demonstrating robustness in controlled settings. The study underscores the benefits of combining 2D-to −3D pose estimation (RMPE, VideoPose3D) with an ST-GCN classifier for improved basketball action recognition. The method proposed in [18], LSTM-DGCN, integrates a Deep Graph Convolutional Network (DGCN) with a Long Short-Term Memory (LSTM) network to recognize basketball actions from three-dimensional skeletal data, effectively capturing both spatial and temporal dependencies in human movement. The DGCN models spatial relationships among 3D skeletal joints, representing key joints as graph nodes and anatomical connections as edges to extract discriminative features for actions such as dribbling, shooting, and passing. The LSTM network captures temporal dynamics, learning long-term dependencies in sequential data to track movement progression. A residual structure mitigates gradient vanishing, improving stability over long sequences. Compared to conventional video-based methods, LSTM-DGCN reduces sensitivity to background noise and better preserves motion details. On the NPU RGB+D dataset (12 basketball actions), it outperforms ST-GCN by approximately 10% across Cross-Validation, Cross-Subject, and Cross-View metrics. However, its performance declines when recognizing two-handed actions (i.e., chest passes, two-handed shooting), indicating the need for further refinement.

4.3. Sensor-based recognition methods

Sensor-based methods employ wearable devices such as inertial measurement units (IMUs), accelerometers, and gyroscopes to capture an athlete's motion signals. These data are processed using machine learning or deep neural networks, making this approach particularly useful in environments where video quality is compromised due to lighting conditions or occlusions. In [40], the authors propose a sensor-based action recognition system for basketball training using an IMU-embedded vest. This cost-effective, minimally intrusive system provides real-time feedback on actions such as dribbling, passing, and shooting. The 3-axis accelerometer and 3-axis gyroscope align with the athlete's body for consistent measurements. Data collected from 21 players were recorded for 30 s per action at 250 Hz to capture fine motion details. IMU signals were preprocessed through filtering, normalization, and a sliding window technique with varying sizes and overlap percentages. Feature extraction included statistical and time–frequency measures such as root mean square and skewness. Several classifiers were evaluated, with XGBoost achieving the best results using a 250-sample window with 75% overlap. Explainable AI techniques identified key motion indicators contributing to classification accuracy. Under optimal conditions, XGBoost achieved 96.9% accuracy, with precision, recall, and F1 scores at similar levels. Leave-one-subject-out (LOSO) cross-validation yielded 87.8% accuracy, slightly lower due to inter-subject variability. Key motion contributors included gyroscope z-axis values and acceleration on the x-axis, demonstrating the system's potential for real-time training. Meanwhile, [41] presents a human action recognition system designed specifically for a virtual reality basketball training application. It employs three Kinect sensors arranged in a circular configuration to capture 3D skeleton data. Each sensor provides 25-joint information at 30 frames per second using both an RGB camera and an infrared system for depth perception. A OpenCV-based marker detection method is used to determine the user's face orientation, ensuring that the system correctly distinguishes between the front and back of the user by swapping the left and right joint data when necessary. When the multiview skeleton data is fused, the system normalizes the data to account for differences in body size by treating the skeleton as a kinematic tree with 25 joints and 24 limbs, using the Spine Base joint as the reference point. The user's position is immobilized by fixing the Spine Base at the origin of the coordinate system, allowing the extraction of features that are invariant to spatial movement. Features extracted from the normalized skeleton include 3D joint positions, joint velocities calculated between consecutive frames, joint angles between adjacent limbs, and angular velocities. For the recognition of actions, the paper introduces a classifier based on a Part-aware Long–Short-Term Memory Network (PLSTM). This network divides the body into five parts — torso, left arm, right arm, left leg, and right leg — and processes the temporal dynamics of each part separately before combining them for the final classification. To improve generalization, the model incorporates dropout with a rate of 0.5 and applies L1 regularization. Implemented with TensorFlow and scikit-learn, the classifier achieves high performance, reaching convergence in fewer than 10 iterations and an overall recognition accuracy of 93.75%.

The experimental evaluation involved data collection from 20 subjects of varying body sizes, genders, and basketball experience. Each subject performed four distinct actions — standing, dribbling, shooting, and passing — each repeated 30 times, resulting in approximately 130,000 frames of labeled data. The hardware setup of the system includes three Kinect sensors connected via USB to individual personal computers, with the collected data transmitted to a central server over TCP/IP for processing. The integrated system is demonstrated in a VR environment developed using Unity3D and HTC Vive, where the recognized actions control a virtual basketball player in real time. Overall, the paper demonstrates that by fusing multiview Kinect sensor data and utilizing a PLSTM-based classifier, it is possible to achieve robust, real-time human action recognition suitable for interactive VR training applications.

4.4. Multimodal fusion methods

Multimodal fusion strategies have emerged as a powerful approach to enhance BAR by integrating complementary information extracted from heterogeneous data sources, such as RGB frames, pose estimations, and motion signals. These methods aim to overcome the limitations of unimodal systems, such as low robustness to occlusions, changes in viewpoint, and lighting variability, by combining the strengths of different modalities through early, late, or hybrid fusion mechanisms.

A representative hybrid model is introduced in [44], where the authors propose a two-stage architecture integrating YOLO for real-time player detection and a deep fuzzy Long Short-Term Memory (LSTM) network for action classification. The YOLO detector isolates players within the frame, allowing the subsequent LSTM network, enhanced with fuzzy logic, to model temporal dynamics while handling uncertainty and enhancing interpretability. Empirical results on the SpaceJam and Basketball-51 datasets demonstrate that this multimodal system significantly outperforms unimodal baselines, achieving high accuracy in classifying basketball-specific actions.

Following this paradigm, a recent work [45] presents a fusion framework combining an improved YOLO variant — augmented with VGG16, VGG19, and ResNet50 backbones for diversified feature extraction — with a Time-aware Two-Stream Fuzzy LSTM (T2FLSTM) classifier. This architecture enables effective feature-level fusion by incorporating visual, spatial, and temporal cues from multiple CNN models, thereby improving robustness in both player detection and action classification. The model attains 99.3% accuracy across eight basketball action classes, confirming the benefits of multi-feature fusion in performance-critical sports applications. In [46], the authors introduce a dedicated Basketball Player Action Dataset (BPAD) and propose YOLOv4 SimSE and SPPFCSPCG (YOLOSS), an action detection algorithm based on YOLOv4 and designed specifically for recognizing basketball movements in video streams. YOLOSS integrates a lightweight GhostNet backbone with a hybrid attention mechanism (SimSE) and a novel spatial pyramid pooling module (SPPFCSPCG), facilitating efficient spatial–temporal feature learning. Evaluated on the BPAD dataset, comprising annotated images of walking, running, and defensive movements, YOLOSS achieves 82.4% mAP@50, confirming its effectiveness in detecting fine-grained actions during real-time gameplay.

4.5. Metrics for performance evaluation in BAR

The effectiveness of a BAR system is typically assessed through a set of well-established evaluation metrics. Among the most widely used are Accuracy, Precision, Recall, and F1-score, which together provide a comprehensive view of the model's classification performance.

Accuracy quantifies the proportion of correct predictions over the total number of predictions, offering a general indication of overall model performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision measures the fraction of correctly predicted positive instances among all instances predicted as positive, thus reflecting the model's ability to minimize false positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall evaluates the proportion of actual positive instances that are correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-score combines Precision and Recall into a single metric by computing their harmonic mean, and is particularly informative in scenarios affected by class imbalance:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Beyond these core metrics, additional measures are often considered to better reflect the real-world applicability of BAR systems. Top-k Accuracy (i.e., Top-1 or Top-5) is commonly used in multi-class classification tasks to indicate whether the correct class appears among the model's top predictions. Mean Average Precision (mAP) provides a more detailed evaluation in multi-label scenarios or when overlapping action segments are present, summarizing precision at various levels of recall.

To assess a model's generalization capabilities, cross-subject and cross-view evaluation protocols are widely employed. These approaches test performance across different individuals and camera perspectives, which are both critical in dynamic sports contexts. Finally, metrics such as inference time and model size are increasingly reported, especially in applications requiring real-time responsiveness, such as live coaching assistance, automated officiating, or player performance monitoring.

5. Explainable Action Recognition (XAR)

The issue of explainability presents a significant challenge in critical applications of video action recognition and machine learning in general [47]. To the best of our knowledge, no dedicated studies specifically address explainability in basketball action recognition. Therefore, in this section, we will provide a general overview of **Explainable Action Recognition (XAR)** methods, discussing their concepts and approaches.

In the domain of sports analytics, AI-driven video analysis systems are increasingly employed to assess athletic performance and detect rule violations. However, the opacity of these models often prevents stakeholders, such as athletes, coaches, and referees, from fully understanding the rationale behind automated decisions. This lack of interpretability can lead to questionable penalties, misjudgments of key events, or inaccuracies in performance evaluation. Moreover, even when the system correctly identifies an infraction or a strategically relevant action, the reasoning behind such classifications may remain obscure, limiting trust and acceptance of AI-generated assessments [48]. To address this limitation, XAR has emerged as a framework that not only identifies high-level activities from video data (or other sensor input) but also provides interpretable justifications for its classifications [32]. Integrating XAR into sports-related AI systems offers significant advantages in multiple domains. For example, an AI-driven coaching assistant could evaluate an athlete's technique while also explaining its assessment, highlighting key factors that differentiate effective from ineffective movements. Similarly, in automated refereeing, an explainable system could justify officiating decisions, such as an offside call in football, an illegal maneuver in gymnastics, or a foul in basketball, by pinpointing the most relevant video frames and motion signals that informed its determination. This transparency improves the credibility of AI-assisted decision-making and supports a more informed evaluation of controversial rulings.

The level of detail provided by XAR-based explanations can vary. A short explanation may highlight the precise moment in a match when a decisive action occurred, such as a goal-scoring opportunity, a rule violation, or a critical pass. In contrast, a detailed explanation might present alternative hypotheses alongside confidence scores for each scenario (i.e., *the system estimates with 85% confidence that the gymnast's foot landed outside the designated boundary due to insufficient landing stability, resulting in a deduction of 0.3 points from the final score*). Beyond improving decision-making, explainability fosters trust and facilitates the broader adoption of AI-driven sports analytics. When the reasoning behind these automated insights is clearly articulated, athletes, coaches, referees, and spectators are more likely to accept and integrate such technologies into training, strategic planning, and officiating.

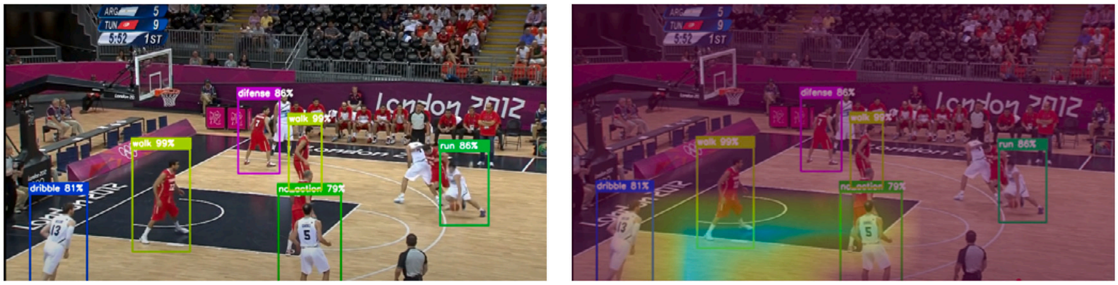

Fig. 2. Correct explainability for detecting the action "walk" using Grad-CAM.

5.1. Current research trends and challenges in XAR

Recent research has increasingly focused on developing techniques to enhance the interpretability of deep learning models in video action recognition. Traditional feature attribution methods, such as Grad-CAM [49], have been adapted for video action recognition (VAR) tasks. However, these methods exhibit notable limitations, particularly in their ability to capture temporal dependencies across video frames. Specifically, they struggle to model sequential relationships effectively, are highly sensitive to background noise, and lack global interpretability across multiple frames. To mitigate these challenges, novel approaches such as Video-TCAV have been introduced, utilizing concept-based explanations instead of pixel-level attributions [50,51]. This paradigm shift aims to provide more human-interpretable insights by focusing on high-level semantic concepts rather than localized gradient-based visual cues (see Fig. 2).

A critical issue in XAR remains the trade-off between accuracy and interpretability. Although deep learning models, particularly CNN-LSTM hybrids and transformer-based architectures, have demonstrated superior classification performance, they largely remain black-box systems. Recently, Mahmoudi et al. [52] systematically examined explainable deep learning models applied to real-time action recognition, revealing that although architectures such as 3D CNNs and two-stream networks achieve high classification precision, their interpretability remains limited due to the complexity of their feature representations. In particular, 3D CNNs effectively capture spatio-temporal features but lack mechanisms for generating human-understandable rationales for their decisions. Similarly, the two-stream convolutional network model, which processes spatial and temporal information separately, improves performance but does not inherently provide an interpretable decision-making process. Consequently, bridging the gap between performance and interpretability remains a key research challenge in explainable video action recognition. To enhance explainability, probabilistic models have been integrated into deep learning pipelines for video action recognition. For instance, Bayesian Neural Networks introduce uncertainty into predictions, providing a measure of confidence; however, they are often computationally expensive, which limits their scalability in real-time applications. Other methods, such as Probabilistic Graphical Models (PGMs) [53] and Sum–Product Networks (SPNs) [54], are notable for their inherent interpretability, as their graphical structures explicitly represent probabilistic dependencies. Despite this advantage, these models usually struggle to maintain high performance when applied to complex, high-dimensional data, such as video sequences [55]. Building on these previous works, Roy et al. [32] propose a hybrid approach that combines deep neural networks with dynamic conditional cut-off networks (DCCNs), a class of tractable probabilistic models designed to efficiently handle explanatory queries. A key distinguishing feature of DCCNs is their ability to provide meaningful explanations by performing inference in polynomial time, thus addressing the limitations of post-hoc interpretability methods. Unlike conventional explainability approaches, which typically analyze model outputs retrospectively, the proposed method integrates DCCNs directly into the deep learning pipeline. This tight integration allows DCCNs to explicitly model the relationships between explanatory variables and outputs, thereby enabling joint inference and producing interpretable results without the need for surrogate models.

Despite these advancements, challenges persist in developing fully interpretable VAR systems capable of operating in real-world conditions. While hybrid architectures that combine deep learning with rule-based reasoning or probabilistic models offer a promising foundation, there is a growing need for methodologies that reconcile performance with transparency in domain-specific contexts such as sports analytics. In particular, future research could explore the integration of spatio-temporal explanation mechanisms directly within deep architectures, for instance, by extending Grad-CAM techniques to 3D CNNs in order to generate temporally consistent saliency maps across key action phases. At the same time, skeletal representations derived from pose estimation could serve as an intermediate semantic space, enabling the identification of motion prototypes that are both compact and interpretable. Embedding logical constraints within attention mechanisms may further enhance model transparency, particularly in scenarios where adherence to domain rules (i.e., possession constraints or movement legality) is essential. Finally, the lack of standardized benchmarks for evaluating explainability in sports-related VAR tasks represents a critical limitation: the construction of annotated datasets linking action phases with expected explanatory evidence would support more rigorous and comparable assessments across models.

6. Applications

Basketball Action Recognition (BAR) datasets have proven essential in advancing research and development across multiple domains, extending beyond traditional sports analytics. The availability of high-quality datasets, encompassing both real-world and synthetic data, has enabled the development of robust models capable of recognizing fine-grained actions in dynamic and complex environments. These datasets support a diverse range of applications, including performance evaluation, tactical analysis, injury prevention, and rehabilitation. Furthermore, the integration of synthetic data has expanded the potential of BAR by addressing data scarcity, class imbalance, and model generalization challenges.

The primary application of BAR datasets lies in sports analytics, where action recognition models facilitate player performance assessment, movement tracking, and automated game annotation. These datasets enable AI-driven systems to generate real-time statistics, detect critical in-game events (passes, shots, rebounds), and classify player actions with high accuracy. Such insights are valuable for coaching staff and players, allowing for data-driven strategy formulation, fatigue management, and personalized training programs. Additionally, automated officiating systems leverage BAR datasets to assist referees in making accurate decisions by detecting rule violations and other game-related events.

Beyond sports performance, BAR datasets have been increasingly applied in healthcare, particularly in biomechanical analysis, injury prevention, and rehabilitation monitoring. Studying basketball movement patterns provides critical insights into injury risk factors, allowing healthcare professionals to develop targeted intervention strategies. In particular, the analysis of jump-landing patterns through computer vision and machine learning techniques has emerged as a promising approach to preventing anterior cruciate ligament (ACL) injuries. These injuries, common in sports involving sudden stops and jumps, often result from incorrect landing mechanics, such as excessive knee valgus or uneven limb loading. BAR datasets, when combined with pose estimation tools, enable the extraction of key kinematic variables from video recordings of jumping movements, helping identify at-risk athletes through automated screening. The adoption of affordable camera systems and scalable processing pipelines supports frequent, remote evaluations, making these tools accessible across a broad spectrum of users, from professional teams to youth sports organizations. Moreover, synthetic data generation and transfer learning further enhance model generalizability across diverse populations and settings [56]. Other approaches have also explored the combination of motion capture sensors with machine learning algorithms to assess posture and technique during dynamic actions like layups or rebound landings. By comparing joint angles to expert-defined references, these systems achieve high consistency with human evaluations, paving the way for real-time feedback and self-correction through mobile platforms [57].

In addition to prevention, BAR datasets are valuable for rehabilitation and return-to-play assessments. For athletes recovering from injuries or concussions, objective movement tracking through action recognition models allows physical therapists to monitor rehabilitation progress more effectively. By analyzing sprinting, pivoting, and lateral movement patterns, these models offer a data-driven approach to assessing functional recovery and readiness for competition, reducing the risk of re-injury.

As BAR datasets continue to evolve, their applications in sports science and healthcare are expected to expand, particularly with advancements in multi-modal learning, synthetic data generation, and explainable AI techniques, further enhancing the precision and interpretability of movement analysis.

7. Conclusions

Basketball action recognition (BAR) has advanced significantly, driven by developments in deep learning and computer vision. Despite this progress, key challenges remain, including dataset limitations, occlusions, and fine-grained action distinctions that affect model generalization. The use of synthetic data has emerged as a promising strategy for augmenting real-world datasets, addressing constraints related to data availability and variability. However, domain adaptation remains a critical challenge, necessitating further advancements in adversarial learning and style transfer to bridge the gap between synthetic and real-world data.

From a methodological perspective, spatiotemporal architectures, multi-modal learning, and skeletal data integration have substantially improved recognition accuracy. Transformer-based models, 3D convolutional networks, and graph-based approaches for pose estimation have further enhanced the ability to capture complex basketball movements. However, ensuring model interpretability remains crucial, particularly in applications requiring transparency, such as automated refereeing and coaching. Explainable AI (XAI) techniques, including probabilistic models and attention-based methods, offer promising solutions, yet the trade-off between accuracy and interpretability persists.

Future research should focus on standardized datasets, improved domain adaptation, and hybrid models that enhance both performance and transparency, while also integrating explainability frameworks to foster trust in AI-driven sports analytics, benefiting coaching, refereeing, and performance evaluation.

CRediT authorship contribution statement

Marco Caruso: Writing – original draft, Investigation, Formal analysis. **Lucia Cimmino:** Writing – original draft, Validation, Investigation. **Fabio Narducci:** Writing – review & editing, Supervision, Conceptualization. **Chiara Pero:** Writing – original draft, Validation, Investigation. **Gianluca Ronga:** Writing – original draft, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the project IDA included in the Spoke 2 - Misinformation and Fakes of the Research and Innovation Program PE00000014, “SEcurity and RIghts in the CyberSpace (SERICS)”, under the National Recovery and Resilience Plan, Mission 4 “Education and Research” - Component 2 “From Research to Enterprise” - Investment 1.3, funded by the European Union - NextGenerationEU. The authors sincerely thank Dr. Alfredo Lamberti for his invaluable guidance throughout this research.

Data availability

No data was used for the research described in the article.

References

[1] C. Bisogni, L. Cimmino, M. Nappi, T. Pannese, C. Pero, Walk as you feel: Privacy preserving emotion recognition from gait patterns, Eng. Appl. Artif. Intell. 128 (2024) 107565.
[2] N. Sikder, M.S. Chowdhury, A.S.M. Arif, A.-A. Nahid, Human activity recognition using multichannel convolutional neural network, in: 2019 5th International Conference on Advances in Electrical Engineering, ICAEE, IEEE, 2019, pp. 560–565.
[3] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comput. Vis. Image Underst. 117 (6) (2013) 633–659.
[4] F. Zhu, L. Shao, J. Xie, Y. Fang, From handcrafted to learned representations for human action recognition: A survey, Image Vis. Comput. 55 (2016) 42–52.
[5] P. Pareek, A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. 54 (3) (2021) 2259–2322.
[6] R. Yue, Z. Tian, S. Du, Action recognition based on RGB and skeleton data sets: A survey, Neurocomputing 512 (2022) 287–306.
[7] K. Host, M. Ivašić Kos, An overview of human action recognition in sports based on computer vision, Heliyon 8 (6) (2022).
[8] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, H. Xiong, A survey on video action recognition in sports: Datasets, methods and applications, IEEE Trans. Multimed. 25 (2022) 7943–7966.
[9] M.G. Morshed, T. Sultana, A. Alam, Y.-K. Lee, Human action recognition: A taxonomy-based survey, updates, and opportunities, Sensors 23 (4) (2023) 2182.
[10] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos “in the wild”, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1996–2003.
[11] C. De Vleeschouwer, F. Chen, D. Delannay, C. Parisot, C. Chaudy, E. Martrou, A. Cavallaro, et al., Distributed video acquisition and annotation for sport-event summarization, NEM Summit 8 (10.1016) (2008).
[12] P. Parisot, C. De Vleeschouwer, Consensus-based trajectory estimation for ball detection in calibrated cameras systems, J. Real Time Image Process. 16 (5) (2019) 1335–1350.
[13] A. Maksai, X. Wang, P. Fua, What players do with the ball: A physically constrained interaction modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 972–981.
[14] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, L. Fei-Fei, Detecting events and key actors in multi-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3043–3053.

[15] P. Parisot, C. De Vleeschouwer, Scene-specific classifier for effective and efficient team sport players detection from a single calibrated camera, Comput. Vis. Image Underst. 159 (2017) 74–88.

[16] S. Francia, Classificazione di Azioni Cestistiche mediante Tecniche di Deep Learning (Ph.D. thesis), 2018.

[17] X. Gu, X. Xue, F. Wang, Fine-grained action recognition on a novel basketball dataset, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 2563–2567.

[18] C. Ma, J. Fan, J. Yao, T. Zhang, NPU rgb+ d dataset and a feature-enhanced LSTM-dgcn method for action recognition of basketball players, Appl. Sci. 11 (10) (2021) 4426.

[19] S.A. Ghasemzadeh, G. Van Zandycke, M. Istasse, N. Sayez, A. Moshtaghpour, C. De Vleeschouwer, Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes, 2021, arXiv preprint arXiv:2112.00627.

[20] S.R. Shakya, C. Zhang, Z. Zhou, Basketball-51: A video dataset for activity recognition in the basketball game, in: CS & IT Conference Proceedings, vol. 11, (7) CS & IT Conference Proceedings, 2021.

[21] D. Wu, H. Zhao, X. Bao, R.P. Wildes, Sports video analysis on large-scale data, in: European Conference on Computer Vision, Springer, 2022, pp. 19–36.

[22] Z. Han, W. Qin, Y. Wang, Q. Wang, Y. Shi, MultiSubjects: A multi-subject video dataset for single-person basketball action recognition from basketball gym, Comput. Vis. Image Underst. 249 (2024) 104193.

[23] G. Fiche, V. Sevestre, C. Gonzalez-Barral, S. Leglaive, R. Séguier, Swimxyz: A large-scale dataset of synthetic swimming motions and videos, in: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games, 2023, pp. 1–7.

[24] L. Zhu, K. Rematas, B. Curless, S.M. Seitz, I. Kemelmacher-Shlizerman, Reconstructing nba players, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 177–194.

[25] X. Suo, W. Tang, Z. Li, Motion capture technology in sports scenarios: a survey, Sensors 24 (9) (2024) 2947.

[26] M. Bicer, A.T. Phillips, A. Melis, A.H. McGregor, L. Modenese, Generative deep learning applied to biomechanics: A new augmentation technique for motion capture datasets, J. Biomech. 144 (2022) 111301.

[27] D. Schneider, M. Keller, Z. Zhong, K. Peng, A. Roitberg, J. Beyerer, R. Stiefelhagen, Synthact: Towards generalizable human action recognition based on synthetic data, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 13038–13045.

[28] H. Hwang, C. Jang, G. Park, J. Cho, I.-J. Kim, Eldersim: A synthetic data generation platform for human action recognition in eldercare applications, IEEE Access 11 (2021) 9279–9294.

[29] H. Qin, C. Yeung, R. Umemoto, K. Fujii, SoccerSynth-detection: A synthetic dataset for soccer player detection, 2025, arXiv preprint arXiv:2501.09281.

[30] A. Sengupta, I. Budvytis, R. Cipolla, Synthetic training for accurate 3d human pose and shape estimation in the wild, 2020, arXiv preprint arXiv:2009.10013.

[31] G. Paulin, M. Ivasic-Kos, Review and analysis of synthetic dataset generation methods and techniques for application in computer vision, Artif. Intell. Rev. 56 (9) (2023) 9221–9265.

[32] C. Roy, M. Nourani, S. Arya, M. Shanbhag, T. Rahman, E.D. Ragan, N. Ruozzi, V. Gogate, Explainable activity recognition in videos using deep learning and tractable probabilistic models, ACM Trans. Interact. Intell. Syst. 13 (4) (2023) 1–32.

[33] Q. Liu, O. Deho, F. Vadiee, M. Khalil, S. Joksimovic, G. Siemens, Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms, in: Proceedings of the 15th International Learning Analytics and Knowledge Conference, 2025, pp. 591–600, http://dx.doi.org/10.1145/3706468.3706546.

[34] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, J. Shen, Digiface-1m: 1 million digital face images for face recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3526–3535.

[35] J. Wang, L. Zuo, C. Cordente Martínez, Basketball technique action recognition using 3D convolutional neural networks, Sci. Rep. 14 (1) (2024) 13156.

[36] J. Xiao, W. Tian, L. Ding, Basketball action recognition method of deep neural network based on dynamic residual attention mechanism, Information 14 (1) (2022) 13.

[37] Y. Zhang, F. Zhang, Y. Zhou, X. Xu, ACA-net: adaptive context-aware network for basketball action recognition, Front. Neurorobotics 18 (2024) 1471327.

[38] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1) 2018.

[39] K. Zuo, X. Su, Three-dimensional action recognition for basketball teaching coupled with deep neural network, Electronics 11 (22) (2022) 3797.

[40] H. Sonalcan, E. Bilen, B. Ateş, A.Ç. Seçkin, Action recognition in basketball with inertial measurement unit-supported vest, Sensors 25 (2) (2025) 563.

[41] B. Yao, H. Gao, X. Su, Human motion recognition by three-view kinect sensors in virtual basketball training, in: 2020 IEEE REGION 10 CONFERENCE, TENCON, IEEE, 2020, pp. 1260–1265.

[42] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[43] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2019) 172–186.

[44] S.B. Khobdeh, M.R. Yamaghani, S.K. Sareshkeh, Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network, J. Supercomput. 80 (3) (2024) 3528–3553.

[45] X. Li, R. Luo, F.U. Islam, Tracking and detection of basketball movements using multi-feature data fusion and hybrid YOLO-T2LSTM network, Soft Comput. 28 (2) (2024) 1653–1667.

[46] Z. Tan, S. Gao, S. Wei, J. Zhang, M. Zhu, Action recognition of basketball players based on hybrid attention module and spatial feature pyramid module, J. Internet Technol. 26 (2) (2025) 211–218.

[47] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023) 1–33.

[48] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[49] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[50] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.

[51] A. Saha, S. Gupta, S.K. Ankireddy, K. Chahine, J. Ghosh, Exploring explainability in video action recognition, 2024, arXiv preprint arXiv:2404.09067.

[52] S.A. Mahmoudi, O. Amel, S. Stassin, M. Liagre, M. Benkedadra, M. Mancas, A review and comparative study of explainable deep learning models applied on action recognition in real time, Electronics 12 (9) (2023) 2027.

[53] M. Vu, M.T. Thai, Pgm-explainer: Probabilistic graphical model explanations for graph neural networks, Adv. Neural Inf. Process. Syst. 33 (2020) 12225–12235.

[54] R. Sánchez-Cauce, I. París, F.J. Díez, Sum-product networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3821–3839.

[55] B.E. Olivas-Padilla, S. Manitsaris, A. Glushkova, Explainable AI in human motion: A comprehensive approach to analysis, modeling, and generation, Pattern Recognit. 151 (2024) 110418.

[56] N. Blanchard, K. Skinner, A. Kemp, W. Scheirer, P. Flynn, " keep me in, coach!": A computer vision perspective on assessing acl injury risk in female athletes, in: 2019 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2019, pp. 1366–1374.

[57] J.R. Balbin, S.P.O. Uy, R.D. Bumagat, Detection of jump landing performance of basketball player using image processing and support vector machine, in: 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM, IEEE, 2023, pp. 1–6.