A Survey on Applications of Deep Learning Algorithms in Basketball and Soccer

Daksh Sangal
IIIT Kota
2023kucp1010@iiitkota.ac.in

Ankit Goyal
IIIT Kota
2023kucp1003@iiitkota.ac.in

Mahesh Waghmare
IIIT Kota
2023kucp1004@iiitkota.ac.in

Abstract

Deep learning and sports analytics represent a rapidly developing field among researchers. However, limited surveys exist in this domain and even fewer focus specifically on sport oriented research applications. To address this gap, we have combined two similar and highly popular sports, basketball and football, as our primary domains of investigation. We are trying to compare different researches in the area of application of deep learning in sports, examining how various neural network architectures and machine learning techniques have been employed to solve complex problems in game analysis and prediction. Our goal would be to compare research on three main topics for now: first, 2D to 3D modeling of games, which involves reconstructing three-dimensional spatial information from two dimensional video footage; second, Game Action Categorization, which focuses on identifying and classifying specific events, movements, and tactical maneuvers during game play; and third, Game Actions predictions or winning predictions, which encompasses forecasting future plays, player decisions, and overall match outcomes. We will compare scores of different models used in different researches across these three application areas, presenting the performance metrics, methodologies, and results in a comprehensive manner, and let the reader decide which is best suited for their specific use case or research direction.

1. Introduction

Sports analytics powered by deep learning represents an area where significant research has been conducted to improve the understanding and analysis of athletic performance. However, despite these efforts, this remains a field with limited comprehensive surveys and is still very much a developing domain. There is considerable scope for advancement in this area, and we aim to help with this survey by comparing different research that has happened since the early applications of deep learning in sports and examining what the future prospects might be for researchers looking to contribute to this field.

For our analysis, we have chosen soccer and basketball as the primary sports to focus on because they are among the most popular sports worldwide. These sports are kind of similar in nature, both are ball oriented team sports with dynamic gameplay, and importantly, much more research has been conducted in these two sports due to their global popularity and high consumer demand for advanced analytics. Therefore, we hope that conducting a survey that focuses primarily on these sports but is not limited

to them alone would be able to generate the greatest impact and provide valuable insight that could potentially be extended to other sports as well.

After studying numerous researches in this domain, we have identified some key areas in sports analytics that researchers are actively trying to solve with deep learning techniques. Some of the major areas we will focus on are: first, 2D to 3D conversion, which involves transforming flat video footage into three dimensional spatial representations; second, Game Prediction and Win Prediction, which uses historical and real time data to forecast match outcomes; and third, Game Action Categorization, which involves identifying and classifying specific events and movements during gameplay.

What our aim is to systematically summarize the researches that different people and research groups have done on these topics. We will list their goods and bads, highlighting the strengths and limitations of each approach detail the models they used, examine the different approaches they took to solve similar problems, and document the datasets they used for training and validation. We will provide this comparison to the reader in a nice tabular and visual way, making it easy to understand and compare multiple studies at a glance. We hope to help fellow researchers with these comparisons so that if in the future someone extends this research or begins new work in these areas, they can take the best path forward by learning from previous successes and avoiding known pitfalls.
Additionally, we will be linking references to all the study material we read and analyzed, providing access to the datasets that were used by the researches that we mention in our survey, and including clear definitions for technical keywords and terminology to ensure our survey is accessible to both newcomers and experienced researchers in the field.

2. KeyWords

This section will act as a reference to the reader providing quick definitions and explanations for terms commonly used in deep learning. this section is indented for readers that are not well versed in the field of deep learning and the explanations will reflect our intent. All the keywords whose definitions are mentioned here will be written in italics.

- ReLu
- Sigmoid
- Negative Log Loss
- ECE
- Spatio Temporal Data
- LSTM
- tanh

3. Researches

This section forms the core of the survey and provides a comprehensive examination of the research studies selected for analysis. To ensure clarity and meaningful comparison, the works are organized into three major thematic categories that reflect distinct methodological goals within sports analytics research: Game Prediction, Game Action Categorization, and 3D Game Simulation. Each thematic category captures a different dimension of how deep learning and computer vision have been applied to soccer and basketball, ranging from forecasting micro-actions and possession outcomes, to identifying complex in-game events, to reconstructing 3D representations of players and gameplay from 2D video.

Within each category, the studies are further subdivided into focused subsections, where each subsection is devoted to a single research paper. This structure enables readers to explore the nuances of each method while maintaining a coherent progression across the broader research landscape.

For every study included in the survey, we present a systematic and critical summary based on five key components:

1. a clear articulation of the problem being addressed and the authors' research objectives;
2. an explanation of the methodological framework and theoretical foundations;
3. a detailed description of the datasets used, including how they were collected, annotated, and prepared;
4. an overview of the machine learning models or algorithms implemented; and
5. a summary of the empirical results, evaluation metrics, and conclusions derived from the findings.

This consistent analytical format provides a unified lens through which to evaluate diverse methodologies, making it easier to compare approaches, understand their contributions, and identify overarching trends across the literature. Additionally, it highlights the strengths and limitations of each study, supporting a holistic understanding of how deep learning is shaping modern sports analytics.

3.1. Game Prediction

This category highlights research that aims to predict the probabilities of micro-actions occurring within a game, using detailed spatio-temporal information such as player positions, ball trajectories, team formations, and lineup identities. Micro-actions—such as passes, dribbles, defensive rotations, screens, or subtle positional adjustments—are fundamental components of both soccer and basketball, yet traditionally remain unquantified in standard analytics. Recent deep learning approaches attempt to model these actions probabilistically by learning from large-scale tracking and event datasets, enabling analysts to estimate expected outcomes, assess decision-making, and understand tactical structures at a granular level. This section presents two representative studies, one focusing on soccer and the other on basketball, each of which develops a deep predictive framework to estimate the likelihood and value of micro-actions based on evolving game context. For both works, we summarize the problem formulations, methodological frameworks, datasets used, model architectures, and key evaluation metrics. Together, these studies demonstrate how modern deep learning pipelines—particularly those combining convolutional, recurrent, and contextual modeling—provide powerful tools for forecasting in-game behavior and quantifying micro-level contributions that were previously invisible in traditional sports analysis.

3.1.1. SoccerMap [LINK]

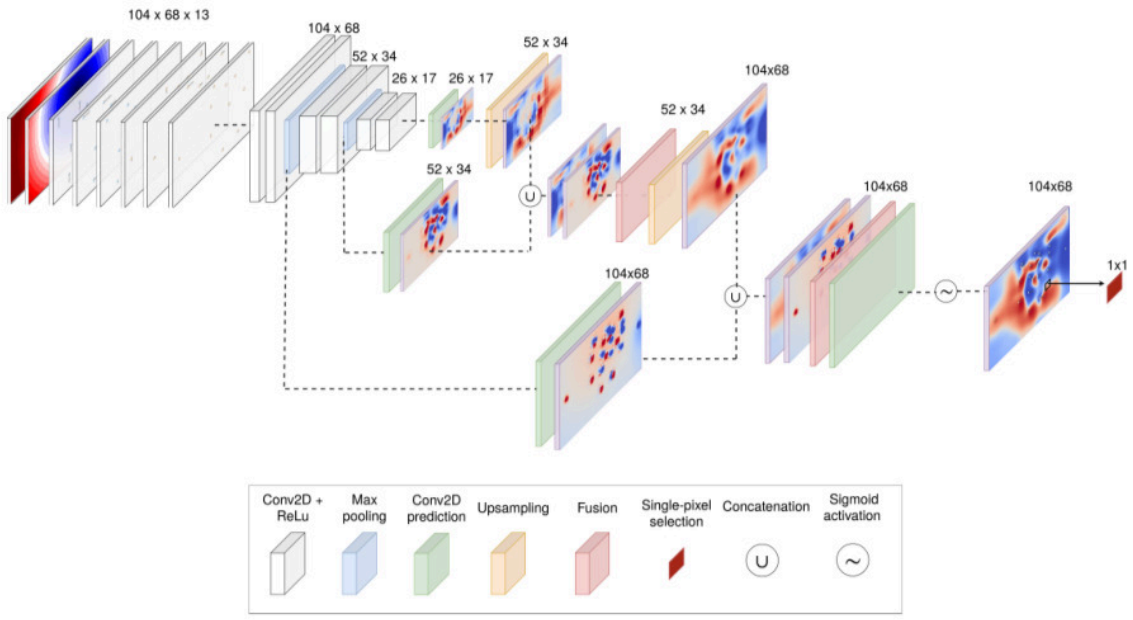

Figure 1: Architecture of the SoccerMap model.

3.1.1.1. Objective

Fernández and Bornn (2020) study the problem of estimating a full probability surface over the soccer pitch that represents, for every possible target location, the likelihood that a pass originating from the current game state will be completed successfully. Traditional soccer analytics tend to focus on discrete, realized events—such as the passes that were actually made, shots taken, or possession transitions—and therefore overlook the rich space of potential decisions available to a player. The authors argue that coaches and analysts require a model that does not merely describe what occurred, but evaluates what could have occurred by quantifying the viability of all possible pass targets. This motivates the creation of SoccerMap, a deep learning architecture capable of producing spatially dense, visually interpretable probability maps. Their research objective is twofold: first, to design an architecture that can learn a meaningful and high-resolution representation of passing opportunities; and second, to demonstrate that such a model can reveal deeper insights into player positioning, team structure, and decision-making quality. Because each training example contains only a single ground-truth output pixel—the actual destination of the pass—the task requires a model that can learn a full dense prediction map from highly sparse supervision, a challenge the authors explicitly position as a novel form of weakly supervised learning within sports analytics.

3.1.1.2. Methodological Framework

To address this problem, the authors represent every game snapshot as a three-dimensional tensor of shape l×h×c l×h×c, where the pitch is discretized into a grid and each channel encodes a different low-level feature derived from high-frequency tracking data. The network is designed as a fully convolutional architecture, allowing it to produce predictions at every spatial location rather than collapsing the data into a single output. At its core, SoccerMap employs a multi-scale convolutional framework, processing the input representation at three levels of resolution $\frac{1}{x}$, $\left(\frac{1}{2}\right)x$, $\left(\frac{1}{4}\right)x$, to extract both fine-grained and

coarse contextual information. At each scale, two symmetric-padded 5×5 convolutional layers followed by ReLU activations capture local spatial patterns such as player proximity, open passing lanes, and defensive pressure, while max-pooling operations reduce resolution in deeper stages to provide a broader contextual perspective.

After learning features at all spatial scales, the architecture produces preliminary predictions at each resolution and then uses learned, non-linear upsampling modules to bring lower-resolution predictions back to the original grid size. These maps are subsequently fused, enabling the model to integrate high-level structural information with detailed local structure. Final predictions are obtained through a 1×1 convolution followed by a sigmoid activation, yielding a probability of pass success at every grid coordinate. The authors train the model end-to-end using a log-loss objective applied only at the single pixel corresponding to the actual pass destination, which forces the network to learn a coherent global probability surface even though supervision is extremely localized. They further emphasize the modularity of the architecture, which can be retrained with alternative labels—such as pass-selection likelihood or expected downstream value—making it applicable to a range of soccer intelligence tasks.

3.1.1.3. Dataset

The model is trained on high frequency spatio temporal tracking data collected from professional soccer matches, consisting of the x,y coordinates of all players and the ball sampled at 10 frames per second. These positions are extracted from match video and aligned with manually annotated event data specifying the start and end points of passes. For each pass, the authors select the frame corresponding to the moment of ball release and label the grid cell that contains the pass destination as the single positive supervision point. All other grid cells in the prediction surface remain unlabeled, which is why the problem is framed as a weakly supervised dense prediction task.

The data comes from 740 English Premier League matches spanning the 2013/14 and 2014/15 seasons, provided by STATS LLC. From the raw tracking data, the authors generate a structured representation with approximately 104×68 spatial resolution and 13 feature channels. These channels encode both raw positional data, such as distances between players—and engineered low level features reflecting defensive pressure, proximity of teammates, and other contextual factors. The result is a large, carefully curated dataset combining fine grained tracking information with high quality event annotations, enabling the model to learn from thousands of real world pass instances across a diverse set of match situations.

3.1.1.4. Models Implemented

The principal model introduced in the paper is the SoccerMap architecture, which merges ideas from fully convolutional networks, multi scale representation learning, and spatial upsampling strategies. After extracting multi-resolution features using stacked convolutional layers, the model creates prediction maps at different scales and fuses them through learned upsampling and integration layers. The reliance on symmetric padding ensures that the spatial resolution remains consistent across layers, preserving the geometry of the pitch and avoiding edge distortions that could bias predictions near the touchlines or goal areas. Training is performed using a log loss function applied only to the actual pass destination pixel, but the dense prediction maps produced by the model enable it to learn spatially coherent representations despite the sparse supervision.

To validate the architecture, the authors compare SoccerMap against simpler baselines: Logistic Net, a single logistic regression unit operating on handcrafted features, and Dense2 Net, a multilayer perceptron with two hidden layers. Both baselines collapse the spatial structure of the problem and produce a single

probability output, in contrast to SoccerMap's pixel-wise surface. The paper also includes an ablation analysis demonstrating that removing multi-scale processing or fusion layers significantly degrades performance, confirming the necessity of the proposed design.

3.1.1.5. Results and Conclusions

The authors split the data into 60% training, 20% validation, and 20% test sets. Their evaluation emphasizes two primary metrics: negative log-loss, which measures probabilistic accuracy, and Expected Calibration Error (ECE), which quantifies how well predicted probabilities align with empirical outcomes. SoccerMap achieves a log-loss of 0.217 and an ECE of 0.0225, outperforming both baselines by large margins. While the model contains more parameters and incurs a slightly higher inference time than the baselines, its substantial gain in predictive quality demonstrates that the added architectural complexity is justified.

Beyond numerical metrics, the authors highlight the value of the visually interpretable probability surfaces generated by the model. These maps reveal which regions of the pitch offer safe or risky passing opportunities, enabling analysts to evaluate player decision-making, identify optimal passing lanes, and understand structural weaknesses in defensive shape. The paper concludes that SoccerMap not only advances the state of the art in predicting pass success but also provides a flexible and general-purpose framework for spatial decision-making analysis in soccer. The authors propose future extensions such as modeling pass value, integrating player orientation or ball velocity, and combining probability surfaces with tactical decision models used in professional clubs.

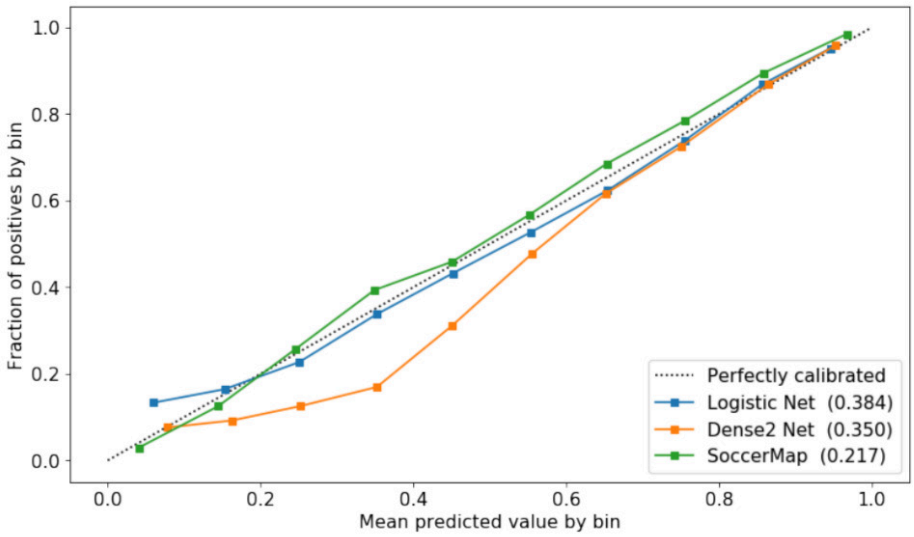

Figure 2: Visual comparison of model outputs on pass-probability surfaces.

Model	Log-loss	ECE	Inference time	Number of parameters
Naive	0.5451	-	-	0
Logistic Net	0.384	0.0210	0.00199s	11
Dense2 Net	0.349	0.0640	0.00231s	231
Soccer Map	0.217	0.0225	0.00457s	401, 259

3.1.2. DeepHoops [LINK]

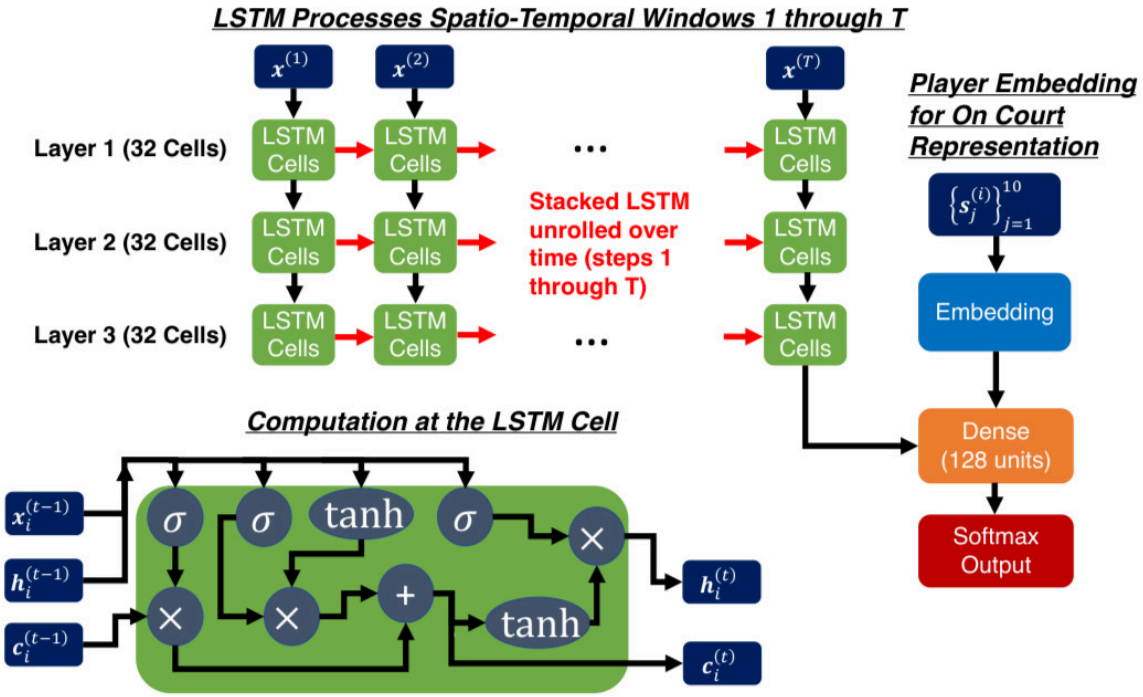

Figure 3: Architecture of the DeepHoops model.

3.1.2.1. Objective

Sicilia, Pelechrinis, and Goldsberry (2019) address the challenge of evaluating the micro-actions that occur within basketball possessions by estimating, at each moment in time, the expected points remaining in a possession. Unlike traditional analytics, which emphasize discrete outcomes such as shots or turnovers, DeepHoops focuses on the evolving sequence of player and ball movements. The central idea is that each possession can lead to a variety of terminal actions—including field-goal attempts, shooting fouls, non-shooting fouls, turnovers, or null endings—and each of these terminal actions is associated with an expected point value. DeepHoops aims to predict the probability distribution over these terminal actions as a function of the spatio-temporal evolution of the play. By doing so, the model can quantify the value of micro-actions (screens, cuts, passes, movements) that traditionally go unmeasured but meaningfully influence possession outcomes. The authors' overarching research objective is therefore to construct a deep learning architecture capable of capturing fine-grained spatio-temporal dynamics and translating

them into a measure of possession value that is interpretable and useful for evaluating player decision-making and team execution.

3.1.2.2. Methodological Framework

The foundational unit of analysis in DeepHoops is the possession, defined as a sequence of n n moments, where each moment is encoded as a 24-dimensional feature vector. The first 20 elements represent the two-dimensional court locations of the 10 players; three additional elements provide the (x, y, z) coordinates of the ball; and the final element encodes the shot-clock value. This formulation transforms raw optical tracking data into a structured representation amenable to sequential modeling. With more than 134,000 such tracked possessions available, the authors define for each possession a temporal window centered at time T, which contains the preceding T moments. This window serves as the input to the model, while its label is determined by the terminal action that concludes the possession.

To model the temporal dynamics embedded in these windows, the authors design the DeepHoops architecture around a stacked LSTM module consisting of three layers with 32 LSTM cells each. This recurrent stack learns a deep temporal representation of player movement, ball trajectory, and evolving spatial configurations. Because the value of a possession is influenced not only by motion patterns but also by the specific players on the court, the architecture includes a second module that encodes lineup identity. This player-identity module allows the model to adjust predictions based on who is involved in the play, thereby capturing the fact that identical movements may yield different expected value depending on personnel. The outputs of both the temporal representation module and the lineup module are combined and fed into a softmax layer that produces the predicted probability distribution over terminal actions. These probabilities, when multiplied by the expected point value associated with each action, yield the model's estimate of expected points at time T, which changes dynamically as the possession progresses.

3.1.2.3. Dataset

DeepHoops is trained on optical tracking data collected from 750 NBA games, a dataset widely known for its precision and high temporal resolution. The optical tracking system records the three-dimensional locations of players and the ball at 25 frames per second, producing tens of millions of spatio-temporal observations. This raw positional data is combined with detailed event annotations—such as fouls, turnovers, shot attempts, and possession boundaries—allowing the authors to segment each game into discrete possessions and label each possession with its terminal action. From this dataset, the authors construct over 134,000 possession sequences that contain rich movement information and span a diverse array of game contexts. Each possession is transformed into a sequence of 24-dimensional moments, preserving both spatial geometry and temporal evolution. The tracking data's high level of annotation makes it particularly well-suited for supervised learning, as each possession can be cleanly aligned with its ground-truth terminal outcome.

3.1.2.4. Models Implemented

The primary model implemented in the study is the DeepHoops stacked LSTM architecture, which serves as the backbone for learning temporal representations of spatio-temporal basketball data. The model consists of three recurrent layers with 32 LSTM cells each, enabling it to capture both short-term movement fluctuations and longer-term structural patterns within the possession. The final hidden state of this LSTM pipeline encodes the evolving game state up to time T. In parallel, a learned embedding is used to model lineup identity, with each player represented via an embedding vector that captures individual influence on possession value. These embedding vectors are aggregated to produce a lineup

representation. The outputs of the LSTM module and the lineup module are concatenated and passed through a fully connected network with a softmax classifier that outputs the probabilities of each terminal event. This probability vector serves as the basis for computing expected possession value, while the architecture as a whole functions as a deep feature extractor that translates raw spatio-temporal data into a meaningful, interpretable metric.

3.1.2.5. Results and Conclusions

To evaluate model performance, the authors employ the Brier Score, a proper scoring rule for assessing the accuracy of probabilistic predictions over categorical outcomes. Training is conducted over five epochs with a minimum improvement threshold of 0.01, ensuring stable convergence. The experiments explore performance as a function of the temporal window size K, with results showing consistent improvements as the model is given larger windows of historical context. For K=1,2,3,4, the Brier Scores, reference scores (BSref), and Brier Skill Scores (BSS) demonstrate meaningful performance gains, with the best accuracy achieved at K=4, corresponding to a Brier Score of 0.2659 and a BSS of 0.2114. These results confirm that incorporating more temporal information leads to more accurate representation of evolving possession states.

The authors also present reliability curves to show that DeepHoops is well-calibrated, meaning its predicted probabilities match observed frequencies of terminal outcomes. Beyond numerical performance, the study concludes that DeepHoops offers a powerful framework for quantifying micro-actions, allowing analysts to attribute value to screens, cuts, passes, and other movements that traditional box-score metrics cannot capture. By modeling expected possession value at every moment, DeepHoops provides a more nuanced understanding of player impact and decision-making, contributing an advanced tool for player evaluation and tactical analysis.

	BS	**BS_ref**	**BSS**	**Epoch Time (s)**
K = 1	0.4569	0.6070	0.2472	2180
K = 2	0.3598	0.4920	0.2686s	2929
K = 3	0.3094	0.4017	0.2299s	3552
K = 4	0.2659	0.3371	0.2114	4200

Table 2: DeepHoops Brier Score (BS), Climatology Model Brier Score (BSref), and DeepHoops Brier Skill Score (BSS). DeepHoops outperforms the climatology (baseline) model in all cases. Performance is best for K = 2 (among the values examined). Epoch Time (in seconds) is lowest over all epochs

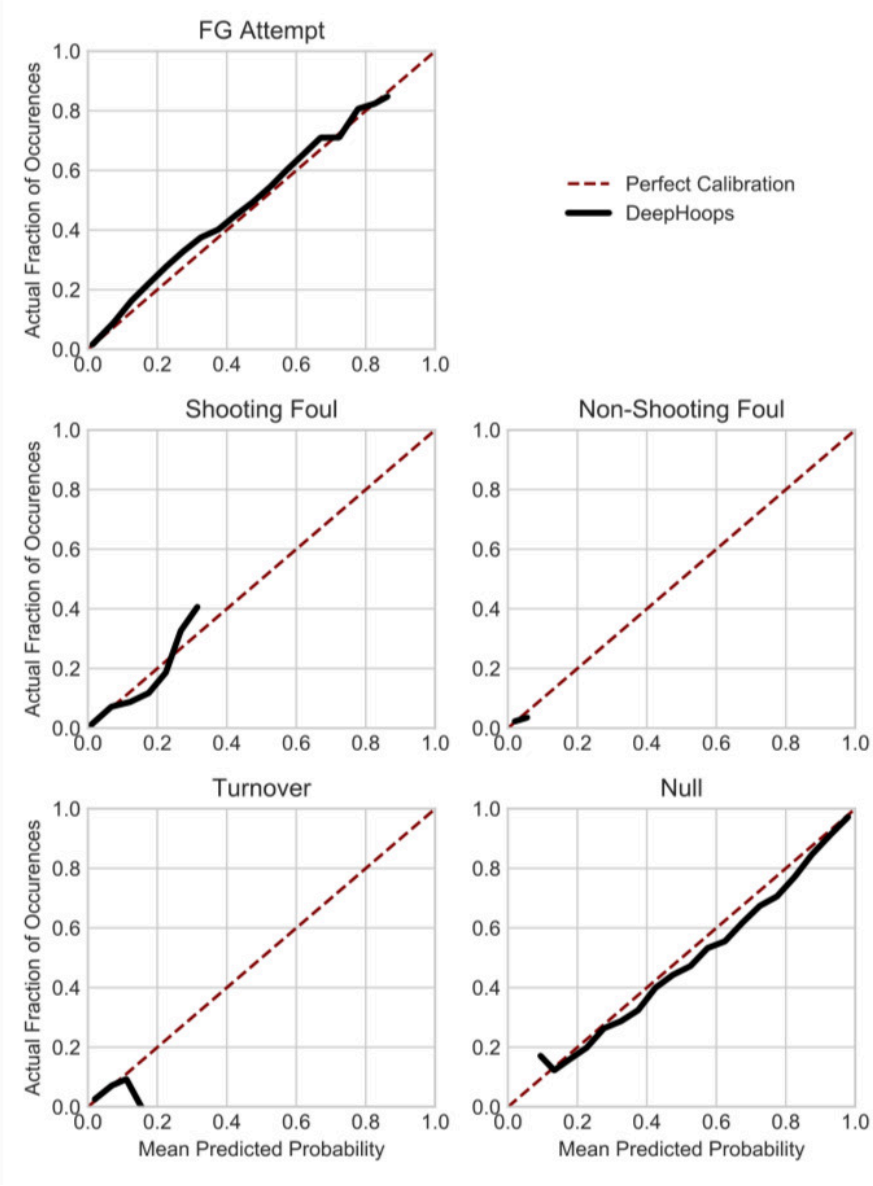

Figure 4: reliability Curves for DeepHoops' probability estimates. The dashed line y = x represents perfect calibration

3.2. 3D Game Simulation

This category focuses on research efforts aimed at transforming traditional 2D sports imagery into enhanced 3D visual representations, enabling a richer and more immersive understanding of game dynamics. Soccer and basketball, in particular, pose significant challenges for such reconstruction due to rapid player motion, camera panning, occlusion, and the inherently limited depth cues of monocular broadcast footage. Researchers in this domain attempt either to simulate 3D structure from 2D images—for example, estimating depth maps or player geometry—or to reconstruct full 3D scenes that can be interacted with using augmented or virtual reality systems. These methods typically rely on advanced deep learning architectures capable of inferring spatial depth, body shape, motion, and camera pose directly from raw video input. In this section, we explore two representative studies: one focused on soccer, where

monocular video is converted into detailed 3D player models for AR visualization, and one centered on basketball, where spatio-temporal trajectories are used to generate predictive or reconstructed 3D representations. Together, these works highlight how deep learning has enabled a new generation of 3D modeling tools that enhance tactical analysis, fan engagement, and simulation capabilities in modern sports analytics.

3.2.1. Soccer On Your TableTop [LINK]

3.2.1.1. Objective

Rematas, Kemelmacher-Shlizerman, Curless, and Seitz (2018) tackle the problem of transforming an ordinary monocular broadcast video of a soccer match into a full 3D reconstruction that can be viewed interactively using augmented reality (AR) devices. Traditional soccer broadcasts present the game from a fixed vantage point, significantly limiting the audience's spatial understanding of player motion, positioning, and tactical geometry. The authors aim to overcome these constraints by reconstructing both the players and the playing field in three dimensions, enabling users to examine the game from arbitrary viewpoints—including a tabletop AR environment. Their overarching research objective is therefore to design a system that uses deep learning, geometric reasoning, and camera pose estimation to recover per-player depth and geometry from a single 2D video, ultimately producing an immersive 3D visualization of real soccer footage that blends computer vision with interactive AR rendering.

3.2.1.2. Methodological Framework

A central methodological challenge is estimating the depth map of each player from a single 2D image, since monocular depth recovery is inherently ill-posed. To address this, the authors develop a dedicated depth-estimation neural network that takes as input a 256×256 cropped RGB image of a player and predicts a 64×64×50 volumetric representation encoding quantized depth. The model architecture is built around eight hourglass modules, a design that captures multi-scale contextual information while preserving fine-level spatial detail. The 50-channel output volume corresponds to 49 quantized depth bins plus one background class. The quantization scheme is centered around a virtual vertical plane passing through the player's midline. Distances are discretized into 49 bins spanning 24 bins in front of the plane, 24 behind, and one bin at the plane itself, each at 0.02-meter spacing. This binning approximates a one-meter depth range centered on the player and provides a manageable discretized representation for learning.

The network is trained using an entropy-based loss with a batch size of 6 over 300 epochs. Once depth is inferred, the authors integrate this information into a complete 3D reconstruction pipeline. The first stage, camera pose estimation, determines the extrinsic parameters of the broadcast camera relative to the pitch. Next, player detection and tracking identify individual athletes throughout the video frames. With camera pose known and each player's depth map estimated, mesh generation is performed by projecting the depth values into 3D space and creating fine-grained player models. This multi-stage pipeline ultimately reconstructs the full 3D geometry of the game, including players and field alignment, enabling the system to render the match inside an AR environment with dynamically selectable viewpoints.

3.2.1.3. Dataset

Because no publicly available dataset contains high-quality depth maps for real soccer players, the authors generate their own dataset using an innovative acquisition method. They leverage RenderDoc to intercept GPU calls between the rendering engine of the video game FIFA and the graphics hardware. By extracting Normalized Device Coordinates (NDC) from the FIFA rendering pipeline, they obtain highly accurate 3D

point clouds corresponding to players' meshes and per-pixel depth buffers. After isolating each player and removing irrelevant scene elements, the authors compile a dataset of approximately 12,000 paired RGB and depth-map samples. These pairs form the supervised training data for their depth-estimation network. The resulting dataset is high-quality, densely annotated, and captures realistic soccer poses and appearances thanks to FIFA's detailed animation models. The use of synthetic yet photorealistic data allows the authors to train a deep network that generalizes effectively to real broadcast footage.

3.2.1.4. Models Implemented

The core model implemented in the study is a convolutional neural network constructed around the hourglass architecture, which is well suited for dense prediction problems such as keypoint detection and depth estimation. The network processes the input player crop through sequential encoder–decoder hourglass modules, repeatedly compressing and expanding spatial resolution to extract global context while maintaining fine discriminative features. The design enables the network to infer subtle depth cues from shading, limb configuration, and uniform texture, despite the inherent ambiguity of monocular input. The final output is a multi-channel depth-probability volume, where each channel corresponds to one of the 49 depth bins or the background. Depth assignment is computed by selecting the most likely bin per pixel. Beyond the depth-estimation model, the authors also implement algorithms for camera pose estimation, player tracking, and 3D mesh construction to form a complete end-to-end 3D reconstruction pipeline.

3.2.1.5. Results and Conclusions

The authors evaluate their approach on a held-out test set composed of 32 RGB-depth pairs extracted from additional FIFA game captures using the same RenderDoc-based method as the training data. Their primary evaluation metric is the scale-invariant Root Mean Squared Error (st-RMSE), a measure designed to assess relative depth accuracy independent of global scale differences. They benchmark their method against three alternatives: a non-human-specific depth-estimation model, a human-specific but non-soccer-trained model, and a parametric human-shape-model fitting approach based on 2D pose estimation. Their model achieves an st-RMSE of 0.06, outperforming the parametric shape model (0.14) and dramatically surpassing both non-human and non-soccer-specific baselines. In terms of Intersection-over-Union (IoU) for foreground segmentation, their method reaches 0.86, indicating highly accurate localization of player geometry. The results demonstrate that the proposed approach can reliably recover per-player depth from single images, and that the synthetic FIFA-derived dataset provides robust supervision.

In conclusion, the authors show that their depth-estimation model and reconstruction pipeline enable high-fidelity 3D visualization of soccer matches from ordinary broadcast video. By combining deep learning with geometric reasoning and synthetic data acquisition, they introduce a new direction for sports analytics, immersive visualization, and interactive AR applications. The work demonstrates that with appropriate training data and architectural design, monocular sports footage can be transformed into compelling 3D reconstructions suitable for analysis, fan engagement, and future research in augmented sports environments.

	st-RMSE	IoU
Non-human training	0.92	-
Non-soccer training	0.16	0.41
Parametric Shape	0.14	0.61
Their Model	0.06	0.86

Table 3: Results of Depth Estimation Network

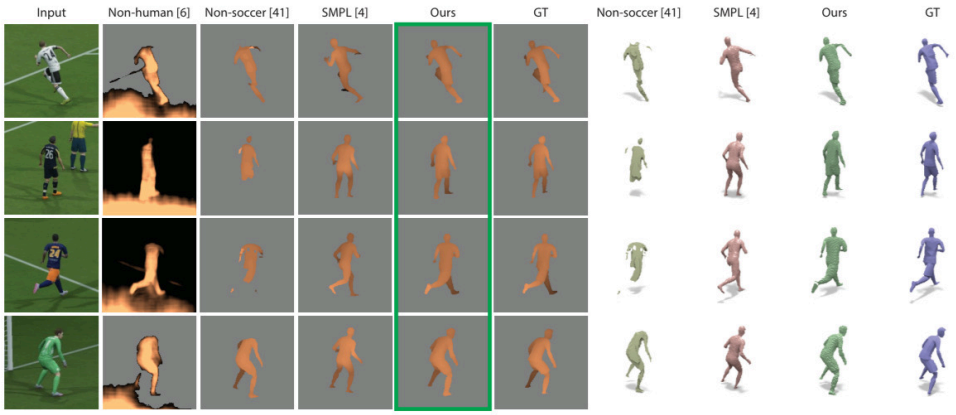

Figure 5: results of the experiment

3.2.2. Basketball Trajectories [LINK]

3.2.2.1. Objective

Shah and Romijnders (2016) investigate whether deep learning can be used to accurately predict the success of three-point shots in basketball, based solely on the spatio-temporal trajectory of the ball. Prior to this work, shot prediction models largely depended on contextual variables such as defender proximity, shooter identity, or game situation, rather than the actual physical arc of the ball. The authors frame the problem as a sequence-learning task, in which the flight path of the ball is treated as a time-series and the goal is to estimate, at each timestep, the probability that the shot will ultimately be made. Their main research objective is to determine whether a Recurrent Neural Network (RNN) equipped with Long Short-Term Memory (LSTM) units can outperform traditional machine-learning models in predicting shot success by learning patterns directly from raw trajectory data. In doing so, the study aims to explore the viability of deep learning as a tool for physics-based sports analytics.

3.2.2.2. Methodological Framework

To capture the temporal dependencies inherent in ball flight, the authors employ a two-layer LSTM architecture using peephole connections, a variant of LSTM that allows each memory cell to inspect its internal state when computing gate activations. This design enables the model to learn fine-grained physical cues from the evolving trajectory. The input at each timestep consists of the ball's X, Y, and Z coordinates, along with the game clock value, providing both spatial and temporal context. As the time series progresses, the LSTM processes successive ball positions and outputs, at each timestep, the probability that the shot

will be successful. These probabilities are computed by a softmax output layer and trained using cross-entropy loss, allowing the network to learn to distinguish the subtle characteristics of successful versus unsuccessful trajectories.

The model is optimized using the Adam optimizer, which accelerates convergence through adaptive learning-rate adjustments. To examine the impact of engineered physics-based features, the authors also create a second input formulation that augments raw XYZ coordinates with additional variables derived from ball mechanics. These include per-timestep velocity components, distance to the rim, change in distance over time, and the shot angle relative to the rim. The expectation is that these physics-informed variables may help the model better infer the underlying ballistic dynamics of successful shots. The overall framework thus combines time-series modeling, physical intuition, and deep recurrent architectures to learn discriminative trajectory patterns.

3.2.2.3. Dataset

The study draws from the publicly available SportVU optical tracking system, which captures the positions of all players and the ball at 25 frames per second across all NBA arenas. The dataset used for this research consists of over 20,000 three-point shot attempts collected from 631 games during the early portion of the 2015–2016 NBA season. The system records precise spatial coordinates of the ball in three dimensions—X along the court length, Y along the width, and Z representing height—making it ideally suited for trajectory-based modeling. Of the collected shots, 35.7% were made, providing a realistic class distribution for predictive modeling.

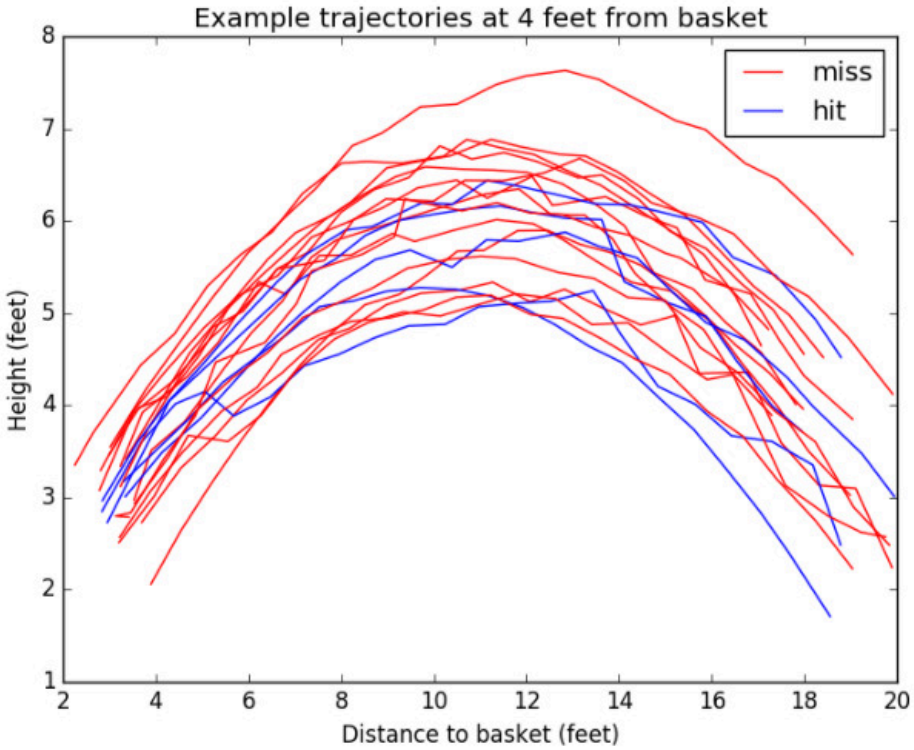

Figure 6: basketball data examples

To construct the training data, each three-point attempt is represented as a sequence of XYZ coordinates indexed over time, along with the game clock. A secondary dataset is derived by augmenting these

trajectories with engineered features inspired by projectile dynamics. These include the ball's per-axis movement differences between frames, distance from the ball to the center of the rim, temporal differences in that distance, and the angle of approach. By combining raw positional data with physics-based attributes, the authors create two distinct model inputs to evaluate whether deep networks can learn these dynamics implicitly or whether engineered features provide measurable predictive improvements.

3.2.2.4. Models Implemented

The primary model used in the study is a Recurrent Neural Network with Long Short-Term Memory (LSTM) units designed to capture the temporal evolution of the ball's flight path. The network consists of two stacked LSTM layers incorporating peephole connections, enabling it to learn temporal patterns and internal cell-state dynamics more effectively than standard LSTMs. The final output is a probability estimate of shot success produced at each timestep. The learning process utilizes cross-entropy loss and Adam optimization, making the model well suited for binary classification tasks involving sequential data. For comparison, the authors also implement two baseline machine-learning models—a Generalized Linear Model (GLM) and Gradient Boosted Machines (GBM)—to evaluate the advantages of deep sequence modeling over traditional approaches.

3.2.2.5. Results and Conclusions

The data is split into an 80-20 train-test partition, and model performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), a standard metric for binary classification that measures the ability of the model to separate successful from unsuccessful shots. The baseline GLM performs poorly, achieving an AUC of 0.53, only slightly above random guessing. The GBM model performs significantly better with an AUC of 0.80, indicating that engineered features and non-linear decision boundaries are effective. However, the LSTM-based RNN achieves the highest performance with an AUC of 0.843, demonstrating that deep recurrent architectures are particularly well suited for capturing the temporal and physical nuances of ball trajectories.

The authors conclude that RNNs outperform traditional models in trajectory-based prediction tasks by learning the dynamics of ball motion directly from raw spatio-temporal sequences. Their results suggest that deep learning can detect subtle cues such as arc height, release angle, velocity decay, and rim-approach angle—factors strongly associated with shot success. More broadly, the study highlights the potential for deep sequence models to advance sports analytics by leveraging high-resolution trajectory data, offering a framework that could be extended beyond basketball to other domains where physics-driven motion plays a central role.

	GLM	**GBM**	**RNN**
AUC	0.53	0.80	0.843

Table 4: Results of Experiment

3.3. Game Action Categorization

This category focuses on the problem of automatically recognizing and classifying actions within soccer video sequences, an area that has attracted substantial research attention due to the fast-paced and visually complex nature of the sport. Action recognition aims to identify meaningful soccer-specific events—such as free kicks, fouls, goals, offsides, corner kicks, or player interactions—directly from video footage, without requiring manual annotation or external metadata. These actions often involve subtle

temporal dynamics, rapid player movement, and significant camera motion, making them challenging to detect using traditional vision methods. Modern approaches therefore rely on advanced machine learning techniques capable of modeling both spatial and temporal information present in video streams. In this section, we examine two representative research works that address this challenge using deep learning, each proposing a different methodology for extracting frame-level features, modeling temporal dependencies, and generating accurate action classifications. Together, these studies demonstrate how recurrent architectures, motion descriptors, and visual-content representations can be combined to effectively interpret and categorize soccer actions from raw video sequences.

3.3.1. Action Classification using LSTM RNN

3.3.1.1. Objective

Baccouche et al. (2010) investigate the challenge of classifying human actions directly from video sequences in the context of soccer, using only visual content without relying on handcrafted priors or domain-specific handcrafted rules. At the time, many action-recognition approaches depended on additional cues such as temporal segmentation rules, motion templates, or domain-specific heuristics, which limited their adaptability. The authors instead propose a method that learns action representations directly from raw video content, treating the video as a sequential signal composed of frames that evolve over time. Their research objective is to develop a learning system capable of recognizing complex, visually diverse soccer actions—such as shots, goals, or dribbling—using only the extracted frame descriptors and a recurrent architecture capable of modeling temporal evolution. By doing so, the authors aim to demonstrate that recurrent neural networks, particularly LSTM-based models, can effectively capture temporal patterns in sports videos and outperform classical machine-learning baselines that rely on static image descriptors.

3.3.1.2. Methodological Framework

The authors follow a sequential learning framework in which each video is divided into its constituent frames, and each frame is transformed into a visual descriptor that captures its appearance and motion characteristics. These descriptors are then fed into a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) that processes them timestep by timestep, learning how frame content evolves as the action unfolds. The final classification decision for the video is produced by aggregating the network's frame-level predictions across the entire sequence.

fig. 7 show's the approach.

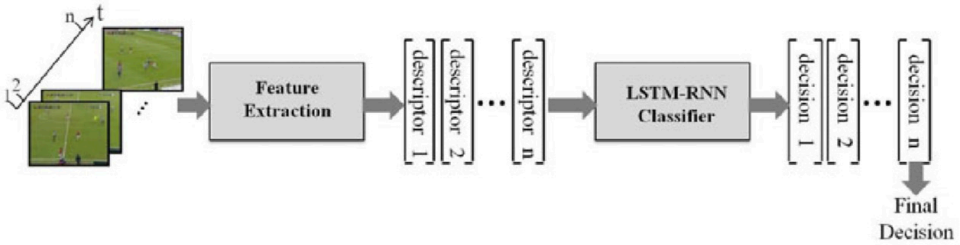

Figure 7: approach used by researchers

To represent visual content, the authors extract a Bag-of-Words (BoW) descriptor for every frame. The BoW representation is built by identifying local interest points across the image and mapping them to a fixed vocabulary of visual words, forming a histogram that summarizes recurrent visual patterns.

However, visual appearance alone cannot fully capture motion dynamics. Therefore, the authors also compute dominant motion features using a SIFT-based approach. SIFT keypoints are extracted from consecutive frames and matched using a KD-tree structure to identify correspondences. Since raw SIFT matches often contain both player-induced and camera-induced motion, the authors apply RANSAC to eliminate outliers and isolate the global camera motion, which is highly informative in soccer broadcasts. Static elements such as TV channel logos are automatically removed, ensuring that the extracted motion descriptor reflects meaningful video dynamics. These combined features—appearance from BoW and motion from SIFT-based estimation—serve as the input representation for the LSTM model.

3.3.1.3. Dataset

All experiments are conducted using the MICC-Soccer-Actions-4 dataset, a benchmark specifically designed for soccer action classification. The dataset contains annotated video clips corresponding to four soccer-specific action categories, recorded from real broadcast footage and manually segmented into distinct events. Each clip is labeled with the action type and contains consistent camera angles typical of televised soccer matches. Although relatively small by modern standards, the dataset offers a challenging test bed due to variations in lighting, occlusions, player density, and rapid camera motion. Its structure makes it suitable for evaluating temporal models, as each clip provides a clean, fixed-length video sequence from which frame descriptors and motion features can be extracted. Because the dataset is carefully curated to minimize noise while maintaining realistic broadcast conditions, it provides a reliable foundation for assessing the model's action classification abilities.

3.3.1.4. Models Implemented

The primary model used is an LSTM-RNN designed to capture long-term dependencies across video frames. Classical RNNs struggle with vanishing gradients, making them unsuitable for long sequences such as sports footage; the LSTM architecture overcomes this through its Constant Error Carousel (CEC) and gating mechanisms that selectively retain or discard information. The network consists of a single hidden recurrent layer whose size depends on the dimensionality of the input descriptor, and a SoftMax output layer that predicts action probabilities at each timestep. The authors employ 150 LSTM cells, noting that larger networks risk overfitting while smaller ones may fail to capture relevant temporal patterns.

Training is conducted using Online Backpropagation Through Time (BPTT) with a learning rate of $10^{-}4$ and momentum of 0.9. This training setup enables the model to update weights incrementally as it processes each temporal window, making it well suited for sequential video data. The combination of BoW, dominant-motion estimation, and recurrent modeling results in a system capable of learning richer temporal dynamics than conventional frame-based classifiers such as k-Nearest Neighbors or Support Vector Machines.

3.3.1.5. Results and Conclusions

All experiments are carried out using a 3-fold cross-validation scheme on the MICC-Soccer-Actions-4 dataset, ensuring robust performance evaluation. The authors test a variety of configurations, beginning with classical machine-learning baselines. Using BoW features alone, a k-NN classifier achieves a classification rate of 52.75%, while an SVM improves performance to 73.25%, indicating that appearance information contains meaningful but incomplete discriminative power. Incorporating temporal modeling through an LSTM-RNN increases accuracy to 76%, demonstrating that recurrent networks can exploit sequential dependencies absent in static classifiers.

When dominant motion features are used as input to the LSTM, performance increases slightly to 77%, affirming that motion cues contribute additional useful information. The most significant performance gain occurs when both BoW appearance features and dominant motion descriptors are combined and fed into the LSTM model. This integrated approach achieves a classification accuracy of 92%, substantially outperforming all baselines. The confusion matrices presented in the paper further reveal that the combined model reduces misclassification between visually similar actions, confirming that modeling both appearance and motion over time is crucial for robust soccer action recognition.

The authors conclude that LSTM-RNNs provide a powerful framework for action classification in sports video, particularly when complemented with domain-appropriate visual and motion features. Their results highlight the importance of temporal modeling and demonstrate that even simple handcrafted descriptors, when paired with recurrent architectures, can achieve high accuracy on challenging sports datasets.

Fig 8 showcases the confusion matrices of different approaches

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0.92**	0.08	0	0
Placed-kick	0.08	**0.8**	0	0.12
Shot-on-goal	0	0.2	**0.72**	0.08
Throw-in	0.12	0.12	0.16	**0.6**

(a)

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0.64**	0.28	0.08	0
Placed-kick	0.08	**0.68**	0.08	0.16
Shot-on-goal	0.08	0	**0.88**	0.04
Throw-in	0.08	0	0.04	**0.88**

(b)

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**1**	0	0	0
Placed-kick	0.04	**0.84**	0.08	0.04
Shot-on-goal	0	0.12	**0.88**	0
Throw-in	0.04	0	0	**0.96**

(c)

Figure 8: Confusion matrices : (a) - BoW-based approach (b) - Dominant motion-based approach (c) - Combination of the BoW and the dominant motion

below is the table that showcase the results

	Classification Rate
BoW + k-NN [LINK]	73.25%
BoW + SVM [LINK]	73.25%
BoW + LSTM-RNN [LINK]	76%
Dominant motion + LSTM-RNN	77%
BoW + dominant motion + LSTM-RNN	92%

Table 5: Results of Experiment

3.3.2. Soccer Video Summarization using Deep Learning

3.3.2.1. Objective

The authors of "Soccer video summarization using deep learning" observe that full-length soccer matches are long and watching entire videos to find important events (goals, shots, key actions) is time-consuming

and laborious, especially for coaches, analysts, or fans seeking highlights. They argue that automatic summarization — extracting the most relevant segments of a match — can save time and enable easier review of highlights. Their problem is therefore: given a long soccer video, how to automatically identify and extract the important "highlight" segments (clips corresponding to significant soccer actions) that best represent the match, using learned video content features rather than handcrafted heuristics. The research objective is to build a deep-learning based summarization system that can detect relevant soccer actions (from broadcast or recorded video), decide which video segments are worth including, and produce a concise summary (highlight reel) of the match.

3.3.2.2. Methodological Framework

To realize this objective, the authors design a two-stage deep learning pipeline combining a 3D convolutional neural network (3D-CNN) and a recurrent neural network (LSTM-RNN). First, they build a 3D-CNN — based on a residual network (ResNet) backbone — to learn spatio-temporal features from video clips: the 3D convolution captures motion and temporal dynamics as well as spatial appearance, making it suitable for recognizing actions in video. They manually annotate a set of soccer video clips (described below) corresponding to different action classes, and train the 3D-CNN to classify those actions. Once the 3D-CNN is trained, the features it extracts from video segments are fed into an LSTM network. The LSTM processes these features over time — treating the match as a sequence of segments — to capture longer-term dependencies and context across the game. Based on the LSTM's output, the system scores each segment for "highlight relevance." Finally, to produce a match summary, the video is modeled as a sequential concatenation of these segments, and those segments with highest relevance scores are selected (or pruned) to generate the final summary video.

3.3.2.3. Dataset

For training the system, the authors manually annotate 744 soccer video clips, drawn from broadcast or recorded matches, and assign each clip to one of five "soccer action classes." These annotated clips serve as the labeled training set for the 3D-CNN's action recognition module. The five action classes represent different kinds of soccer events relevant for highlights (though the paper does not treat all possible match events — only a subset of actions the authors consider important). Once trained, the network uses these classes to detect similar events in full match videos. For evaluation, the authors apply their summarization pipeline to ten full soccer match videos (not part of the training set) to test how well the generated summaries align with expected highlights.

This dataset creation strategy (manual annotation of action-level clips + full-match video summarization) enables the system to learn domain-specific features relevant to soccer, while remaining simple enough to be annotated and processed.

3.3.2.4. Models Implemented

The core components of the system are a 3D-CNN (Residual-Network-based) and an LSTM-RNN. The 3D-CNN is designed to ingest short video clips (a few frames or short segments) and output a feature representation encoding both spatial and temporal information — i.e. appearance, motion, and short-term dynamics. This representation serves as a learned alternative to handcrafted features or heuristics for recognizing soccer actions. The LSTM-RNN then processes these feature vectors across successive segments of a match, capturing longer-term temporal context: this is important because soccer events and highlights often depend not only on isolated clips but their sequence and buildup (e.g., possession leading to a goal). By combining 3D convolution for short-term motion and LSTM for long-term context,

the method aims to leverage the strengths of both architectures. Finally, the system uses a selection mechanism (based on LSTM output) to pick which segments are "highlight-worthy" and discard non-essential or dull parts, producing a concise summary video.

3.3.2.5. Results and Conclusions

To evaluate their summarization system, the authors generated summaries for ten soccer match videos and evaluated the outputs using user studies: 48 participants from eight different countries watched the generated highlight videos and rated them using a Mean Opinion Score (MOS) scale. On average, the summarizations received a 4 out of 5 MOS, indicating users judged the summaries as high-quality and acceptable. This suggests that the system was successful at capturing relevant highlights and producing summaries that human viewers find satisfactory. The results support the authors' claim that a deep-learning based approach — combining 3D-CNN for action recognition with LSTM for temporal context — can effectively automate soccer video summarization, reducing the manual labor involved in analyzing entire match footage while preserving user-perceived quality of highlight reels.

Additionally, the authors argue that their framework is general enough to be adapted to other sports or domains, as long as a suitable annotated clip dataset is available, making the approach a promising direction for automated video summarization beyond just soccer.

Bibliography

[1] J. Fernández and L. Bornn, "SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer," *arXiv preprint arXiv:2010.10202*, 2020.

[2] A. Sicilia, K. Pelechrinis, and K. Goldsberry, "DeepHoops: Evaluating Micro-Actions in Basketball Using Deep Feature Representations of Spatio-Temporal Data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2096–2104.

[3] K. Rematas, I. Kemelmacher-Shlizerman, B. Curless, and S. Seitz, "Soccer on Your Tabletop," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4738–4747.

[4] R. Shah and R. Romijnders, "Applying Deep Learning to Basketball Trajectories," *arXiv preprint arXiv:1608.03793*, 2016.

[5] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks," in *International Conference on Artificial Neural Networks*, Springer, 2010, pp. 154–159.

[6] R. Agyeman, R. Muhammad, and G. S. Choi, "Soccer Video Summarization Using Deep Learning," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 270–273.

[7] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, 2000.

[10] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2015.

[11] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.

[12] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Chapman, Hall, 1989.

[13] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[14] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[15] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[16] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] R. J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," *Neural Computation*, vol. 1, pp. 270–280, 1989.

[18] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic Estimation of 3D Human Shape and Pose from a Single Image," in *European Conference on Computer Vision*, 2016, pp. 561–578.

[19] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network," in *Advances in Neural Information Processing Systems*, 2014.

[20] P. Ghosh, D. Tzionas, and M. J. Black, "Learning Human Motion Models for Long-Term Predictions," in *CVPR Workshops*, 2017.