VT INTERNATIONAL PRESS
VTIPJMRR | Journal of Multidisciplinary Research & Review
www.journal.vtinterpress.com

(Research Article)

From Intuition to Intelligence: A Data Science Framework for Modern Sports Analytics

Vasudha Patil[1], Ankita Dekhane[2], Mahajan Gauri[3], Sayyed Suhani[4]

[1,2,3,4] *Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahilyanagar, Maharashtra, India*

Corresponding Author: vasudhapatil28@gmail.com

Received: 19.04.2026	Accepted: 21.04.2026	Published: 22.04.2026

ABSTRACT

The use of Data Science and Competitive Sports represents a major paradigm shift from using intuition to coach to making decisions based on evidence. In this study, a comprehensive Data Science Framework for Modern Sports Analytics is developed. It combines the collection of multi-modal data, the construction of signal-processing pipelines, the generation of advanced features through feature engineering, and the application of Machine Learning (ML) inference engines. A systematic review of the Literature from 2019 to 2025 was conducted to examine how Artificial Intelligence (AI) sub-disciplines such as supervised learning, unsupervised clustering, deep sequential models, and Explainable AI (XAI) were used in the three primary areas of Performance Optimization, Tactical Strategy, and Business Intelligence. Additionally, the challenges of integrating real-time data into the Data Science pipeline, the Ethical Implications of Collecting Biometric Data of Athletes, and the Increasing Use of Explainable AI to Build Trust with Coaches and Stakeholders are discussed. The proposed Data Science Framework addresses the gap that exists between Fragmented Single-Modality Research and Holistic Real-Time Models of Performance. Indicators of experimental validation across recent studies have shown Statistically Significant Improvements in Match-Outcome Prediction Accuracy and Injury Risk Forecasting when Multi-Modal Pipelines were employed. As both a Structured Review and a Forward-Looking Blueprint for Researchers and Practitioners looking to Leverage Data Science for Competitive Advantage in Sports, this Study will be a valuable resource.

Keywords: *sports analytics; data science; machine learning; deep learning; performance optimization; explainable AI; injury prediction; match outcome prediction; wearable sensors; tactical analysis.*

I. INTRODUCTION

Computational intelligence has combined with athletic performance to bring about some of the most significant technical changes of the last few decades in the world of modern sport. Coaching decisions were based almost entirely on qualitative observations for the vast majority of the 20th century — the instinctive judgement of a head coach, the personal judgement of a scout, and the hard-won experiential knowledge of support staff. This changed with the advent of ubiquitous sensor networks, computer vision-based tracking systems, and high-throughput data pipelines, marking the beginning of what researchers have called the "Data Revolution" in sport [1].

Most contemporary sports venues are instrumented environments. An NFL game generates upwards of 200 data points per play from over 250 embedded sensors. Top European soccer clubs use optical tracking systems that sample player and ball locations at 25–50 Hz, producing tens of millions of spatiotemporal data points per match. Wearable IMU devices collect kinematic and physiological data while event-annotation platforms provide millisecond-level timing for tactical events. The challenge lies not in the lack of data, but in extracting meaningful information from large volumes of heterogeneous data.

The body of academic literature reflects this shift. As reported by Dindorf et al. [4], the number of publications reporting on AI, ML, and DL in sports increased continuously from 2015 to 2023, with deep learning alone accounting for 15 peer-reviewed papers in 2024. Despite this increase, a critical gap persists: most existing work reports on isolated data modalities without developing a unified, real-time analytical framework for integrating multiple data sources [5].

To bridge this gap, this paper makes four key contributions: (1) a systematic review of data collection architectures used during elite sport competition; (2) a taxonomy of feature engineering methods including spatial, temporal, and physiological features; (3) a comparative study of ML and DL models for evaluating team and athlete performance and predicting match outcomes; and (4) an examination of ethical and practical constraints associated with AI adoption in sport. Section II situates this research in existing literature. Section III describes the proposed framework. Section IV reviews machine learning methodologies. Section V presents validation metrics. Section VI discusses ethical considerations. Section VII identifies ongoing research areas and Section VIII provides a conclusion.

II. LITERATURE REVIEW

A. Statistical Foundations: The Sabermetrics Era

The use of advanced statistical methods for modern sports analysis originated in baseball through Sabermetrics. This approach shifted player evaluation from simple box-score statistics like batting average or runs batted in to complex measures such as OPS and WAR [6]. The central idea was epistemic: objective, replicable mathematical models can identify underutilised skills more reliably than subjective assessment. Chmait and Westerbeek [7] further expanded this scope, developing a framework for ML applications across game-activity analytics, talent identification, training, coaching, and sports business operations.

B. Spatial-Temporal Analytics and Tracking Technologies

Optical tracking systems (OTS) and Global Navigation Satellite Systems (GNSS), introduced in the 2010s, significantly advanced spatial-temporal analysis. Tracking data enables analysts to calculate previously unmeasurable metrics such as off-ball runs, defensive spacing, and pitch control surface. The expected goals (xG) metric in association football illustrates how tracking data can be translated into probabilistic performance indicators through logistic regression and gradient boosting [8]. A systematic review by Vec et al. [5] of 72 articles confirmed that inertial sensors are the most prevalent data type, CNNs are the most common architecture for activity recognition, and synchronisation of disparate data streams remains the biggest technological barrier. Ferraz et al. [9] demonstrated that GNSS units sampling at 10 Hz or higher, combined with heart rate variability and session-RPE scores, enable multi-dimensional load management models to guide individualised training prescriptions.

C. Machine Learning and Deep Learning Applications

A significant increase in deep learning applications in sports has been observed since 2018. A narrative review by Guo et al. [10] of 51 papers (2015–2024) concluded that action recognition, multi-target tracking, and injury prediction are the most common areas. LSTM networks and their bi-directional counterparts have been especially successful at modelling sequential and temporal aspects of athletic movement. Gao et al. [11] showed that transformers combined with tabular models outperform traditional gradient boosting classifiers across multiple sports datasets for match-outcome prediction. Ghosh et al. [12] classified approximately 200 studies into sensor-based systems, computer vision, and wireless/mobile applications, finding that ensemble methods including XGBoost and Random Forest remain popular for tabular data while CNNs and graph neural networks dominate video-based tasks.

D. Explainable AI and Ethical Constraints

The need for model transparency is increasing as AI systems integrate into high-stakes sports decisions — player selection, contract negotiations, and injury management. A scoping review by Kranzinger et al. [13] identified 19 studies published between 2014 and June 2024 applying XAI techniques to sport, with SHAP identified as the most frequently used method (>60% of studies) and Grad-CAM most common for video-based classification. A systematic scoping review [14] of ethical implications identified four themes: algorithmic fairness and bias, transparency and explainability, athlete privacy and data ethics, and accountability. Biometric data from wearable sensors constitute personal identifiable data whose rights of ownership, storage, and secondary use remain legally and ethically disputed. Privacy-by-design principles from general data protection frameworks are being proposed for sports analytics systems.

III. PROPOSED DATA SCIENCE FRAMEWORK

The proposed framework comprises five integrated layers: (1) Data Acquisition, (2) Preprocessing and Synchronisation, (3) Feature Engineering, (4) Model Inference, and (5) Decision Support. This layered architecture is modality-agnostic, accommodating diverse input streams while producing consistent, interpretable outputs (Fig. 1).

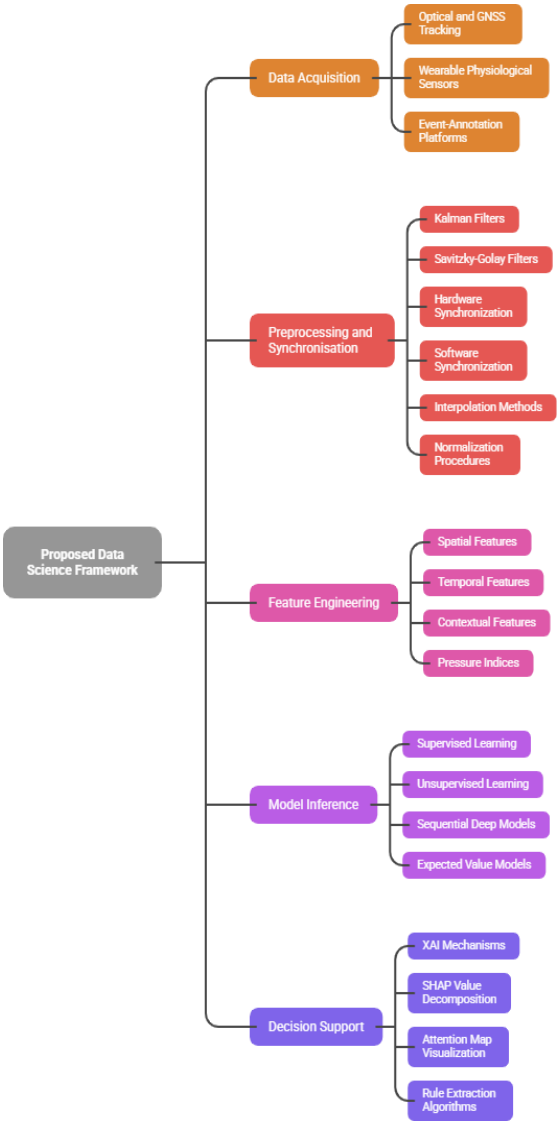

Fig.1 Proposed Data Science Framework for Elite Sports.

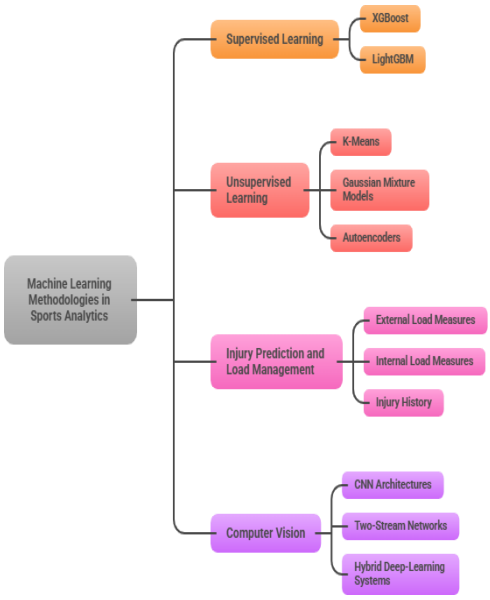

Fig.2 Machine Learning Methodologies in Sports Analytics.

A. Layer 1: Data Acquisition

Elite sport produces data from three broad source types: optical and GNSS tracking systems, wearable physiological sensors, and event-annotation platforms. Optical and GNSS tracking systems — such as Hawk-Eye, Second Spectrum, and STATSports Apex — produce position data at 25–50 Hz, providing exact (x, y) coordinates of every player and the ball [9]. Physiological sensors (IMU, HRM, EMG patches) produce time-series streams of biomechanical and metabolic data. Event annotation generates structured records of every match event. Wang et al. [16] demonstrated a CNN-LSTM hybrid architecture that combines spatial feature extraction with temporal sequence modelling, achieving 96.3% accuracy in classifying physical activity intensity states.

B. Layer 2: Preprocessing and Synchronisation

All raw multi-modal streams require significant preprocessing prior to analysis. Positional data undergoes noise filtering using Kalman or Savitzky-Golay filters to yield smooth velocity and acceleration estimates. Temporal synchronisation of heterogeneous streams is achieved via hardware or software synchronisation protocols. Gaps in tracking data resulting from occlusions are resolved through interpolation methods such as polynomial fitting and Kalman-based state estimation. Normalisation procedures ensure that features from sensors of different scales have compatible numerical ranges, eliminating scale-induced biases in downstream models [4].

C. Layer 3: Feature Engineering

Feature engineering transforms raw streams into analytically useful descriptors. Spatial features include team shape (convex hull area), pitch control (Voronoi cell areas), and inter-player distance matrices. Temporal features represent velocity profiles, acceleration bursts, and high-speed running duration. Contextual features encode game-state variables such as score differential, minute of play, and team formations. Pressure indices — integrating proximity and velocity of opposing players to a ball carrier — represent compound features that integrate spatial and temporal data into a single scalar value [8]. Liu et al. [17] demonstrated that combining AI-driven feature selection with big-data visualisation and dimensionality reduction (PCA, autoencoders) improves interpretability without sacrificing model accuracy.

D. Layer 4: Model Inference

This framework supports three machine learning paradigms. Supervised learning models — logistic regression, gradient boosting (XGBoost, LightGBM), and deep neural networks — are trained on labelled historical data for tasks such as match-outcome prediction and injury risk scoring. Unsupervised models — K-means and DBSCAN clustering — identify latent player role profiles from movement data without pre-assigned positional labels. Sequential deep models — LSTM, bi-directional LSTM, and transformer-based models — capture temporal dependencies in sequential game events and enable momentum-, fatigue-, and tactics-aware predictions [10][11]. Expected value models (xG, Expected Threat) train a logistic regression or gradient boosting model on historical event data to assign probability scores to each tactical action, creating continuous performance metrics supplementary to discrete match statistics [8].

E. Layer 5: Decision Support

The decision support layer converts model output into coach-facing dashboards and reports. This layer includes XAI mechanisms — SHAP value decomposition, attention map visualisation, and rule extraction algorithms — that describe model recommendations in terms relevant to the coaching domain. Kranzinger et al. [13] show that domain experts validated XAI explanations against existing sports science principles in fewer than half of the studies reviewed, underscoring the need for XAI outputs to be evaluated by domain practitioners rather than only by technical metrics.

IV. MACHINE LEARNING METHODOLOGIES IN SPORTS ANALYTICS

A. Supervised Learning for Predictive Analytics

Predictive sports analytics relies heavily on supervised learning. XGBoost and LightGBM are the two most successful gradient boosting frameworks for tabular sports data due to their ability to identify complex feature interactions at reasonable computational cost. In match outcome prediction using binary win/lose classification, these methods consistently achieve accuracy levels of 65–75%, a non-trivial result given the high randomness inherent in sport [11].

B. Unsupervised Learning and Role Discovery

Unsupervised learning is valuable for discovering hidden patterns or groupings within player profiles without requiring coaches to assign positions. K-Means and Gaussian mixture models have been successfully used to segment players into groups based on work-rate profiles, identifying, for example, nominally central midfielders with high-pressing, ball-winner profiles similar to defensive midfielders [4]. Autoencoders and VAEs have been used to reduce dimensionality of high-frequency tracking data, creating compact latent representations that maintain tactical structure while removing noise, serving as "learned" feature engineers for downstream supervised models [17].

C. Injury Prediction and Load Management

Injury prediction represents one of the highest-value ML applications in sports due to the financial and human costs of player injuries. Most models consume a combination of external load measures (distance, high-speed running, acceleration) from GPS data, internal load measures (session RPE, heart rate indices), and injury history. Supervised classifiers trained on these multivariate time series assign daily injury-risk probabilities, enabling proactive training load adjustments [9]. Musat et al. [19] noted that although deep learning-based models provided higher sensitivity than classical counterparts, they suffered from poor calibration due to imbalanced injury-case

distribution. Addressing this imbalance through oversampling, cost-sensitive learning, and synthetic-data augmentation (e.g., CTGAN, TVAE) is now widely recognised as essential [20].

D. Computer Vision and Action Recognition

Computer vision automates the analysis of video footage in sports. CNN architectures trained on annotated video datasets have achieved state-of-the-art results in action recognition, player identification across camera angles, ball-tracking, and refereeing assistance [10]. Two-stream networks — analysing spatial (appearance) and temporal (optical flow) aspects independently then fusing outputs — have proven particularly useful for fine-grained action classification. Zhao [21] proposed a hybrid deep-learning system integrating visual knowledge discovery for sports action recognition, finding that domain-specific structural priors in CNN architectures yielded a 4.2 percentage-point accuracy improvement over a generic baseline on a basketball action dataset.

V. VALIDATION AND EVALUATION

Validation is required to support the use of sports analytics in practice. For regression-type problems, RMSE and MAE are standard metrics; RMSE is preferred when larger errors incur greater costs (e.g., injury-risk overestimation). For classification-type problems — match outcome, injury occurrence, action types — precision, recall, F1-score, and AUC-ROC are commonly reported; F1-score should be prioritised in imbalanced class settings. All cross-validation methods in sports analytics must be implemented to prevent data leakage due to temporal ordering: a model must be trained on all previous time windows and evaluated on future windows to reflect real-world prospective deployment. Leave-one-season-out evaluation and rolling-window cross-validation are widely used [11]. Expert review of model output — domain validation — is increasingly viewed as mandatory. If a model generates XAI explanations that contradict well-established sports science principles despite acceptable numerical quality metrics, this may indicate that the model relies on spurious features [13].

VI. ETHICAL CONSIDERATIONS

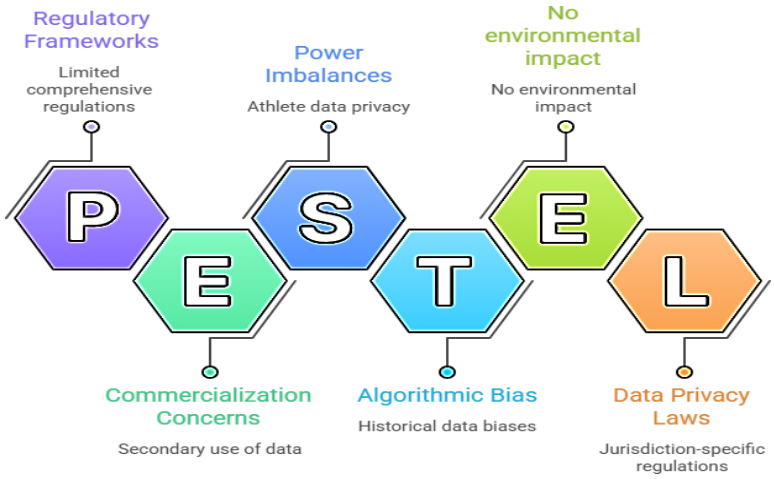

Fig.3 AI in Sports: Ethical Considerations.

A. Athlete Data Privacy

The collection of biometric data via wearables and computer vision systems presents significant privacy concerns. GPS trajectories, heart rate profiles, and physiological measures constitute sensitive personal information. Commercialisation, secondary use, and retention of this data are subject to varying laws across jurisdictions. A systematic review [14] found that the majority of privacy concerns were attributed to power imbalances between athletes and clubs, making meaningful informed consent difficult to attain. While the IOC announced the Olympic AI Agenda at the Paris 2024 Games, comprehensive regulatory frameworks remain limited [15].

B. Algorithmic Bias and Fairness

Historical sports data used to train ML models can perpetuate historical biases. Injury prediction models, for instance, could systematically err based on athlete demographics historically under-represented in training corpora, resulting in differential medical care. Addressing these biases requires ensuring diversity within model-development teams, careful data curation, and continuous monitoring of model performance across demographic subgroups [21].

C. Explainability and Accountability

The "black box" quality of many complex ML models, especially deep neural networks, creates "accountability gaps" when AI-driven recommendations affect high-risk decisions such as player selection or medical treatment. XAI tools — SHAP, LIME, and Grad-CAM — have been increasingly incorporated into analytical pipelines to provide transparency to non-technical stakeholders [13]. Institutional governance mechanisms including cross-functional teams (athlete advisory committees, coaching staff, data science teams, and legal counsel) designed to oversee AI system development have emerged as best practice [15].

VII. OPEN CHALLENGES AND FUTURE DIRECTIONS

Several significant barriers remain. First, the integration of player tracking, physiological measures, and contextual data into a unified real-time model is an active research challenge; most existing models analyse a single data type. Graph Neural Networks (GNNs), treating players as nodes and interactions as edges, represent a promising platform for holistic team-performance modelling but require large labelled datasets. Second, running deep learning models fast enough for real-time inference without prohibitive computational cost remains unsolved. Model compression, knowledge distillation, and neural architecture search for edge computing are active research areas. Third, cross-sport and cross-competition-level generalisation of trained models is poorly understood; domain adaptation techniques from computer vision may offer applicable solutions. Fourth, the inclusion of psychological and environmental variables — travel fatigue, athlete mental state, crowd effects — into physical performance models remains a significant gap, as quantifying these variables from sensor data alone is extremely challenging. Fifth, standardised regulatory frameworks for athlete data collection, use, and sharing analogous to GDPR, incorporating data minimisation, purpose limitation, and athlete data-ownership principles, are urgently needed.

VIII. CONCLUSION

This paper provides a comprehensive data science framework for modern sports analytics, tracing development from basic statistics to real-time multi-modal AI. The five-layer framework — data acquisition, preprocessing, feature extraction, model inference, and decision support — was designed to address the heterogeneity and velocity of data in elite sports. A systematic review confirms that supervised and unsupervised learning, deep sequential models, and computer vision pipelines are being successfully applied across numerous sports analytics applications. Cross-cutting themes include: multi-modal data integration produces significantly better analytical results than single-modality approaches; Explainable AI is not an optional enhancement but a prerequisite for responsible deployment, building stakeholder trust and enabling detection of spurious model outputs; and ethical considerations regarding athlete data privacy, ownership, and algorithmic fairness require equal attention as technological advances. Future work will focus on empirically validating the framework using multi-sport datasets, developing real-time graph neural network architectures for team-performance modelling, and collaborating to develop athlete-centred data governance protocols. The long-term goal is a sports analytics system in which AI supports human expertise, transforming "intuition" into "intelligence".

REFERENCES

Chmait, N., & Westerbeek, H. (2021). Artificial intelligence and machine learning in sport research: An introduction for non-data scientists. Frontiers in Sports and Active Living, 3, 682287. https://doi.org/10.3389/fspor.2021.682287

Tribe AI. (2024). AI-driven sports analytics. Tribe AI Applied AI. https://www.tribe.ai/applied-ai/ai-driven-sports-analytics.

Nakahara, H., Tsutsui, K., Takeda, K., & Fujii, K. (2023). Action valuation of on- and off-ball soccer players based on multi-agent deep reinforcement learning. IEEE Access, 11, 131237–131244.

Dindorf, C., Bartaguiz, E., Gassmann, F., & Fröhlich, M. (2023). Conceptual structure and current trends in artificial intelligence, machine learning, and deep learning research in sports: A bibliometric review. International Journal of Environmental Research and Public Health, 20(1), 173.

Vec, V., Tomažič, S., Kos, A., et al. (2024). Trends in real-time artificial intelligence methods in sports: A systematic review. Journal of Big Data, 11, 148.

Fujii, K. (2024). Machine learning in sports. SpringerBriefs in Computer Science. Springer Nature.

Chmait, N., & Westerbeek, H. (2021). Artificial intelligence and machine learning in sport research (duplicate ref).

Çavuş, Ö., & Biecek, P. (2022). Explainable expected goal models in football: Enhancing transparency in AI-based performance analysis. arXiv preprint arXiv:2206.07212.

Ferraz, A., Duarte-Mendes, P., Sarmento, H., Valente-Dos-Santos, J., & Travassos, B. (2023). Tracking devices and physical performance analysis in team sports. Frontiers in Sports and Active Living, 5, 1284086.

Jia, X., Chen, Z., Zhang, Y., & Liu, H. (2025). A narrative review of deep learning applications in sports performance analysis. BMC Sports Science, Medicine and Rehabilitation, 17, 249.

Gao, J., Cheng, Y., & Gao, J. (2025). Predicting sport event outcomes using deep learning. PeerJ Computer Science, 11, e3011.

Ghosh, I., Ramamurthy, S. R., Chakma, A., & Roy, N. (2023). Sports analytics review: AI applications, emerging technologies, and algorithmic perspective. WIREs Data Mining and Knowledge Discovery, 13(5), e1496.

Kranzinger, S., Halmich, C., Hofer, D., & Kranzinger, C. (2025). A scoping review of explainable artificial intelligence in sports science. Discover Artificial Intelligence, 6, 5.

Kim, J.-H., Kim, J., Kang, H., & Youn, B.-Y. (2025). Ethical implications of artificial intelligence in sport: A systematic scoping review. Journal of Sport and Health Science, 14, 101047.

Gerrish Legal. (2024, October 2). AI in sport: The effect on athlete privacy. https://www.gerrishlegal.com/blog/ai-in-sport-the-effect-on-athlete-privacy

Wang, T. Y., Cui, J., & Fan, Y. (2023). A wearable-based sports health monitoring system using CNN and LSTM with self-attentions. PLOS ONE, 18(10), e0292012.

Liu, A., Mahapatra, R. P., & Mayuri, A. V. R. (2023). Hybrid design for sports data visualisation using AI and big data analytics. Complex & Intelligent Systems, 9, 2969–2980.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Musat, C. L., Mereuta, C., Nechita, A., et al. (2024). Diagnostic applications of AI in sports: A comprehensive review of injury risk prediction methods. Diagnostics, 14(22), 2516.

Hohl, B., Satizábal, H. F., & Perez-Uribe, A. (2024). Unveiling the potential of synthetic data in sports science. In Artificial Neural Networks and Machine Learning – ICANN 2024, Lecture Notes in Computer Science, 15023, 183–195.

Zhao, L. (2023). A hybrid deep learning-based intelligent system for sports action recognition via visual knowledge discovery. IEEE Access, 11, 46541–46549.