Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37118

Data-Driven Analysis of Basketball Player Performance Using Statistical and Machine Learning Methods

Qingcheng Fu

Shanghai Hongwen School, Shanghai, China
fuqingcheng777@163.com

Abstract. With the rapid accumulation of sports tracking and match record data, data-driven methods have become vital tools for evaluating basketball players. Traditional box-score statistics can only reflect partial on-court output, while hidden factors such as game pace, playing time, and team context interfere with fair player comparison. This paper systematically reviews common data types, frequently used indicators, mainstream statistical and machine-learning methods, evaluation criteria, and recent representative research cases, and further distinguishes tasks including retrospective performance description, player-type identification, and future performance prediction. Existing studies are still restricted by inconsistent statistical standards, limited access to high-level tracking data, selection bias caused by playing time, and poor interpretability of complex models. Future work may focus on multi-source data fusion, explainable machine learning, and unified evaluation benchmarks. Overall, this review explains how data science can support more comprehensive, objective, interpretable, and reproducible basketball player evaluation across different analytical tasks and data settings.

Keywords: basketball player performance, statistical analysis, machine learning, sports data science, player evaluation

1. Introduction

Basketball is a highly collaborative team sport, and objective player evaluation is important for talent scouting, lineup adjustment, training planning, and game-day decision-making.

Classic statistics such as points, rebounds and assists have long served as the primary basis for judging player quality. Nevertheless, raw box-score numbers cannot fully reflect defensive contribution, off-ball movement or tactical value. Players with different roles, playing time and positions cannot be directly compared through simple counting statistics, which creates obstacles for coaches and analysts seeking objective judgments. In recent years, large public datasets containing box scores and play-by-play logs have become increasingly available, while high-resolution player-tracking data remain more restricted.

Researchers can extract multidimensional features from complex match records, and statistical and machine-learning techniques are increasingly used to build broader player-evaluation systems. Therefore, basketball player performance analysis has gradually grown into a typical intersection of sports science and data science [1-3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37118

Early basketball evaluation mainly depended on manual observation and simple descriptive statistics. As sports analytics developed, researchers introduced regression analysis to adjust for selected contextual factors and dimensionality-reduction methods to summarize correlated performance indicators.

Afterwards, supervised and unsupervised machine-learning models including clustering, random forest, support vector machine and gradient boosting were introduced to solve problems such as player grouping and performance prediction. Previous papers have explored datasets from the National Basketball Association, college basketball and international tournaments. Many works try to control confounding variables such as position, minutes played and seasonal differences. A wide range of metrics are adopted, covering true shooting percentage, usage rate, plus-minus value and model-derived composite scores. Because basketball studies differ in data availability, model complexity, and evaluation goals, review research should compare methods in terms of data requirements, interpretability, computational cost, and reproducibility [3, 4].

This review focuses on applications of statistical and machine-learning methods in basketball player performance analysis. The core purpose is to summarize frequently-used data types, evaluation indicators, modeling pipelines and assessment criteria, and compare key outcomes from representative existing studies. The rest of this paper is organized as follows. Section 2 introduces basic concepts of basketball performance data and player evaluation. Section 3 describes mainstream statistical and machine-learning algorithms used in this domain. Section 4 reviews recent real-world research cases and compares different analytical pipelines. Section 5 discusses major limitations of current research and puts forward prospects for future exploration. Section 6 concludes the whole paper.

2. Basic concepts of basketball performance data and player evaluation

To carry out data-driven player assessment, it is necessary to clarify core analytical units, conventional basketball metrics and basic data-science terminology, which provide shared foundations for subsequent method comparison. Fundamental analytical objects include matches, seasons, individual athletes, teams, offensive and defensive possessions, player positions and playing minutes. Traditional box-score technical statistics contain points, rebounds, assists, steals, blocks, turnovers and personal fouls, which record countable events within each game. It should be noted that statistical outputs can be calculated on per-game, per-minute or per-possession bases. Per-minute and per-possession normalization can reduce the influence of unequal playing time and is widely used in modern sports analytics [4].

Advanced derived indicators further enrich the evaluation system. Common examples consist of effective field-goal percentage, true shooting percentage, usage rate, assist rate, rebound rate and on-court plus-minus value. Composite efficiency indicators integrate multiple box-score variables into one single score for quick player ranking. Even so, these derived statistics still rely on observable on-court events and cannot fully capture invisible tactical sacrifice or off-ball defensive pressure.

Several key data-science concepts should also be clarified for basketball research. A sample usually corresponds to a player-season record or single-game observation. Features refer to all measurable variables such as shooting efficiency and rebound rate, while target variables are the outcomes researchers intend to predict or interpret. Analysts should inspect outliers carefully and handle missing observations using transparent, task-appropriate procedures. Standardization unifies feature scales for model input. Separating data into training, validation, and test sets helps estimate generalization performance and reduces optimistic evaluation bias.

DOI: 10.54254/2753-8818/2026.37118

3. Statistical and machine learning methods for basketball player performance analysis

A variety of statistical tools and machine-learning algorithms have been applied to identify patterns in basketball datasets, with different methods serving different analytical purposes. Classical statistical approaches start with descriptive statistics to summarize the distribution of player-level indicators. Correlation analysis quantifies linear relationships among different performance variables. Multiple linear regression can estimate associations between contextual factors such as position and game pace and player output while statistically adjusting for selected covariates. Principal component analysis transforms correlated indicators into orthogonal principal components and can reduce dimensionality while retaining a chosen proportion of the total variance [5].

Machine-learning models expand research boundaries for basketball analytics. Unsupervised clustering algorithms divide athletes into groups according to performance feature similarity, so analysts can identify archetypal player roles without manual labels. Decision-tree-based ensemble models including random forest and gradient boosting have gained popularity for their strong fitting ability on tabular sports data. Support vector machines can perform classification tasks such as distinguishing high-level and ordinary athletes. These models are used for player-role recognition, performance-grade classification and continuous-indicator estimation tasks [3, 6-8].

Standard data processing pipelines cannot be ignored. Data cleaning removes abnormal records, and feature selection filters redundant variables. Feature standardization places variables on comparable scales and is particularly important for scale-sensitive algorithms. Cross-validation helps estimate out-of-sample performance and identify overfitting risk during model selection [4].

Different evaluation metrics are selected according to task types. For regression tasks, mean absolute error and root mean squared error measure prediction error, while R^2 summarizes the proportion of outcome variation explained by the fitted model under the stated evaluation setting.

For classification tasks, accuracy, precision, and recall reflect model classification performance. Accordingly, method comparison should focus on assumptions, data requirements, interpretability, computational cost, and suitability for different basketball-analysis tasks.

Method comparisons should emphasize algorithm assumptions, applicability, strengths, and limitations across basketball-analysis tasks.

4. Recent studies of statistical and machine learning methods for basketball player performance analysis

Recent studies have applied statistical and machine-learning methods to basketball player evaluation, role identification, and performance prediction across different datasets and analytical settings. Most recent studies use public datasets from the National Basketball Association or university basketball. Sample scales range from hundreds to thousands of player-season observations. Input features combine classic box-score variables, advanced derived metrics and partial play-by-play information. Some studies take composite player quality scores as target variables, while others focus on predicting future win contribution [4, 8].

For comprehensive player evaluation research, regression and dimensionality-reduction methods show good interpretability, yet they are limited when dealing with complex non-linear relations among sports variables. Tree-based ensemble models can provide strong predictive performance on tabular sports data, although their advantage depends on the dataset, target variable, and baseline configuration [7, 8].

When researchers identify player roles through clustering algorithms, final grouping results are heavily affected by feature selection and normalization strategies. Confounding factors must be

DOI: 10.54254/2753-8818/2026.37118

seriously considered. Playing-time thresholds can bias sample composition. Changes in game pace can also alter per-possession statistics [6, 9].

Team tactical arrangement, opponent defensive intensity and physical conditions caused by injury will all reshape individual performance. Results obtained from one league or single season cannot be directly generalized to other environments.

Researchers should distinguish three different kinds of analytical tasks: describing historical performance, discovering hidden player types, and estimating future performance. This paper introduces a general workflow based on public basketball datasets. First, analysts confirm concrete research objectives. Next, suitable player-performance indicators are chosen. Researchers then handle missing values and reduce bias associated with uneven playing time through appropriate normalization or sample-design choices. After that, statistical or machine-learning models are implemented, and outcomes are assessed by predefined metrics. This whole procedure demonstrates the essential difference between simple single-statistic judgment and multi-dimensional data-driven evaluation. Even with identical raw data, feature engineering choices and model settings will produce divergent conclusions. From the perspective of data science, feature construction, normalization, dimensionality reduction and model interpretability jointly determine the reliability of basketball-analysis outputs [4, 8].

5. Limitations and prospects

Despite promising progress, current data-driven basketball-player-evaluation research still faces multiple obvious constraints. First of all, statistical standards are inconsistent across different leagues and historical seasons. Data recording rules vary, which obstruct cross-league comparison of research conclusions. Traditional box-score statistics cannot quantify off-ball defense, screen quality and tactical sacrifice, so many critical parts of player value remain invisible to mainstream datasets. High-precision optical tracking data which record every athlete's movement trajectory are normally non-public for ordinary researchers [2].

Selection bias brought by playing time and team role creates another major obstacle. Coaches tend to assign more minutes to reliable athletes. Players who receive limited opportunities may have unstable or unrepresentative statistical profiles because of small sample sizes and role constraints. Injuries and frequent roster turnover disturb model stability. Under high-dimensional and relatively small-sample conditions, machine-learning models are prone to overfitting [4, 8].

Furthermore, different research teams adopt diverse target variables, train-test-split schemes and evaluation metrics. Without unified experimental standards, results from separate papers cannot be directly compared. Complex black-box ensemble models often achieve high predictive accuracy, yet analysts struggle to explain why a certain player receives high or low scores, which reduces practical value for coaching staff who need actionable explanations [4, 10].

Several promising directions deserve attention in follow-up exploration. Multi-source data fusion combines box-score logs, play-by-play events and partial tracking information to capture more comprehensive athlete behavior. Context-aware performance indicators should be developed, so that outputs can reflect game situations when each stat is generated. Cross-season validation helps test whether established models remain valid under changing rosters and tactical styles. Explainable machine-learning techniques can unpack the decision logic of complicated models, making computational outcomes understandable for basketball practitioners. Time-series modeling is suitable for tracking performance fluctuations throughout one season. Building shared evaluation benchmarks will promote communication among different research groups. Through these

DOI: 10.54254/2753-8818/2026.37118

improvements, data-driven player analysis can gain greater reliability, comparability, and practical value for real-world basketball scenarios [2, 3, 10].

6. Conclusion

This paper reviews the applications of statistical and machine-learning techniques in basketball-player-performance analysis. First, this review summarizes basic basketball analytical units, conventional box-score statistics, advanced derived indicators, and fundamental data-science terminology used in sports research. Afterwards, this paper introduces frequently-used statistical approaches including correlation analysis, multiple regression and principal-component analysis, alongside mainstream machine-learning algorithms such as clustering, random forest and gradient boosting. Recent representative studies are summarized to compare model performance on retrospective description, player-type recognition and future-performance estimation tasks. Current research faces several bottlenecks, including inconsistent statistical standards, poorly measured off-ball and tactical contributions, sample-selection bias, data-access barriers, overfitting risk, and limited model interpretability.

In the future, multi-source data fusion, context-oriented indicators, cross-season validation, explainable artificial intelligence and unified evaluation benchmarks may address these existing difficulties. Overall, this review shows that data-driven basketball evaluation is most reliable when descriptive statistics, contextual normalization, and predictive models are interpreted together rather than used as isolated rankings. Future research should test models across multiple seasons, leagues, and player roles, report playing-time and pace adjustments explicitly, compare interpretable statistical baselines with more complex machine-learning methods, and evaluate whether performance gains remain stable under alternative data splits. Studies should also distinguish retrospective description, player-type identification, and future-performance prediction because these tasks require different targets and validation designs. Greater access to standardized play-by-play and tracking data would improve reproducibility and allow stronger external validation. These steps can support player assessments that are more comparable, transparent, and useful for scouting, coaching, and long-term performance analysis.

References

[1] Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big ideas in sports analytics and statistical tools for their investigation. *WIREs Computational Statistics, 15*(6), e1612. https://doi.org/10.1002/wics.1612

[2] Kovalchik, S. A. (2023). Player tracking data in sports. *Annual Review of Statistics and Its Application, 10*, 677–697. https://doi.org/10.1146/annurev-statistics-033021-110117

[3] Ghosh, I., Ramamurthy, S. R., Chakma, A., & Roy, N. (2023). Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective. *WIREs Data Mining and Knowledge Discovery, 13*(5), e1496. https://doi.org/10.1002/widm.1496

[4] Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J., & Van Roy, M. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning, 113*, 6977–7010. https://doi.org/10.1007/s10994-024-06585-0

[5] Stone, J. D., Merrigan, J. J., Ramadan, J., Brown, R. S., Cheng, G. T., Hornsby, W. G., Smith, H., Galster, S. M., & Hagen, J. A. (2022). Simplifying external load data in NCAA Division-I men's basketball competitions: A principal component analysis. *Frontiers in Sports and Active Living, 4*, 795897. https://doi.org/10.3389/fspor.2022.795897

[6] Muniz, M., & Flamand, T. (2022). A weighted network clustering approach in the NBA. *Journal of Sports Analytics, 8*(4), 251–275. https://doi.org/10.3233/JSA-220584

[7] Zuccolotto, P., Sandri, M., & Manisera, M. (2023). Spatial performance analysis in basketball with CART, random forest and extremely randomized trees. *Annals of Operations Research, 325*(1), 495–519. https://doi.org/10.1007/s10479-022-04784-3

DOI: 10.54254/2753-8818/2026.37118

[8] Papageorgiou, G., Sarlis, V., & Tjortjis, C. (2024). Evaluating the effectiveness of machine learning models for performance forecasting in basketball: A comparative study. *Knowledge and Information Systems, 66*, 4333–4375. https://doi.org/10.1007/s10115-024-02092-9

[9] Chen, R., Zhang, M., & Xu, X. (2023). Modeling the influence of basketball players' offense roles on team performance. *Frontiers in Psychology, 14*, 1256796. https://doi.org/10.3389/fpsyg.2023.1256796

[10] Kranzinger, S., Halmich, C., Hofer, D., & Kranzinger, C. (2026). A scoping review of explainable artificial intelligence in sports science. *Discover Artificial Intelligence, 6*, Article 5, 1–17. https://doi.org/10.1007/s44163-025-00709-8