AIMS Mathematics, 11(9): 29636–29671.
DOI: 10.3934/math.20261176
Received: 11 July 2026
Revised: 23 August 2026
Accepted: 03 September 2026
Published: 14 September 2026

https://www.aimspress.com/journal/Math

Research article

An interpretable unsupervised ranking system for auditing subjective ranking decisions

Zheng Fang[1], Wei Liang[2,*], Xinyue Weng[2], Dongding Zhang[2], and Junjie Zhou[1]

[1] School of Aerospace Engineering, Xiamen University, Xiamen 361102, China
[2] School of Mathematical Sciences, Xiamen University, Xiamen 361005, China

* **Correspondence:** Email: wliang@xmu.edu.cn.

Abstract: Subjective ranking decisions are widely used in practice, yet their outcomes may reflect voter preferences, reputation, and contextual considerations that are not fully captured by quantitative indicators, making it difficult to assess how closely final decisions align with measurable evidence. To address this challenge, we developed an interpretable intelligent system with an unsupervised ensemble ranking module for data-driven auditing of subjective ranking decisions. The system consists of two components: (1) an unsupervised feature-weighted ensemble ranking module that integrates heterogeneous feature-relevance estimators to construct a within-season performance ranking of the evaluated players without using historical voting outcomes in feature-importance estimation, ensemble-weight selection, or player scoring; and (2) an interpretable diagnostic comparison module that applies the same rule-based model classes under performance-based and voting-based labels to examine agreement and divergence between the two evaluation perspectives. We used the National Basketball Association (NBA) regular-season Most Valuable Player (MVP) award as a representative case study, with the official MVP winner serving as an external voting-based reference for the performance-based ranking. Experiments covering the 1980–2025 award years showed that the framework produces a stable performance-based benchmark and identifies notable seasons, including 2001, 2005, 2006, and 2008, in which the official voting-based outcome diverges markedly from the data-driven ranking. Robustness analyses further showed that feature-attribution changes caused by correlated variables do not substantially propagate to the final ranking structure, while the integrated ranking remains comparatively stable under feature perturbation and correlated-feature ablation. These results demonstrate the value of integrating label-free ranking construction, external winner-based evaluation, and interpretable diagnostic comparison within a unified auditing framework for subjective ranking decisions.

Keywords: unsupervised learning; ranking framework; interpretable machine learning; subjective evaluation; sports analysis
Mathematics Subject Classification: 62H30, 62H25, 62P25, 68T05, 68T10

1. Introduction

Subjective ranking decisions are widely used in practice. Many team-sport awards, including the NBA regular-season MVP award, are determined through voting or scoring procedures that necessarily involve human judgment. Such decisions often generate debate because different evaluators may place different emphasis on individual production, team success, leadership, or broader contextual value. More generally, this reflects a common challenge in subjective evaluation systems: final outcomes may incorporate preferences and qualitative considerations that are not fully observable from quantitative indicators alone.

In many real-world ranking problems, the final decision is easier to observe than the evaluation logic that produced it. This lack of transparency motivates data-driven tools that can provide a performance-based reference and make the relationship between measurable evidence and subjective outcomes more explicit. Unsupervised learning is particularly suitable for this purpose because it allows a ranking benchmark to be constructed without using the historical decision outcome as a supervisory label. In the setting considered here, the resulting framework produces a within-season performance ranking of the evaluated players, while the official winner is used as an external voting-based reference for evaluation and interpretation.

The NBA regular-season Most Valuable Player (MVP) award provides a useful case for studying subjective ranking decisions because it combines rich public performance statistics with a voting-based award process. The award is commonly interpreted through multiple considerations, including individual statistical production, team success, leadership, and contextual value. Because the relative importance assigned to these considerations is not explicitly specified, it is difficult to determine the extent to which a final voting outcome is supported by measurable player-performance evidence alone.

The NBA also provides detailed player-season statistics over a long historical period, making it suitable for reproducible data-driven evaluation. This setting allows a label-free performance ranking to be constructed without using historical MVP outcomes in feature-importance estimation, ensemble-weight selection, or player scoring. The official MVP winner can be used as an external voting-based reference to examine where the voting-based outcome appears within the performance-based ranking.

The NBA MVP award was first given in the 1955–56 season. Through the 1979–80 season, the winner was selected by NBA players. Beginning with the 1980–81 season, the award has been decided by a panel of sportswriters and broadcasters, whose ballots assign points to first- through fifth-place selections [1, 2]. Although the voting procedure itself is explicit, the relative importance attached to statistical production, team context, leadership, and other qualitative considerations is not directly observed. The MVP setting therefore provides a natural test bed for an auditing framework that compares a label-free performance benchmark with the voting-based outcome.

Existing data-driven studies of sports awards have largely focused on predicting historical voting outcomes or reproducing observed voting-outcome patterns. Such approaches are useful when the objective is prediction, but they address a different question from the one considered here. When historical MVP outcomes are used as supervisory labels, the resulting model learns directly from the subjective decision process that an auditing framework is intended to examine.

This study instead focuses on constructing a within-season performance ranking of the evaluated players without using historical MVP outcomes in feature-importance estimation, ensemble-weight selection, or player scoring, and then evaluating how the official winner appears within that

performance-based ranking. The framework further connects the ranking stage to an interpretable diagnostic comparison in which the same rule-based model classes are applied under performance-based and voting-based labels. This matched design allows differences in the resulting diagnostic profiles to be examined primarily in relation to the alternative evaluation perspectives rather than differences in the diagnostic model class.

Accordingly, we propose the unsupervised feature-weighted ensemble ranking (UFER) framework for auditing subjective ranking decisions. Its novelty is framework-level and problem-driven: it combines performance-based ranking construction, external comparison with voting outcomes, and interpretable diagnostic analysis within a single framework. The main contributions are summarized as follows:

(1) We formulate subjective award evaluation as an auditing problem rather than an outcome-prediction problem, and construct a within-season performance ranking of the evaluated players independently from historical voting outcomes.

(2) We develop an unsupervised ensemble ranking strategy that integrates complementary feature-relevance signals from the URF, Genie3, and URelief estimators at the feature-importance level before player scoring, rather than aggregating historical award labels or final voting rankings.

(3) We introduce a matched interpretable diagnostic comparison based on the RuleFit and Skope-Rules algorithms, in which the same diagnostic model classes are applied under performance-based and voting-based labels to identify where the two evaluation perspectives align or diverge.

The proposed framework therefore provides a transparent approach for comparing measurable performance evidence with subjective recognition, using NBA MVP selection as an empirical case study.

2. Related work

2.1. Subjective evaluation and decision auditing

Subjective and objective evaluations represent two complementary ways of assessing performance, but they do not necessarily lead to the same conclusions. In organizational settings, studies of compensation and performance appraisal have shown that subjective evaluation can be affected by institutional design and evaluator discretion, creating the scope for systematic bias or inconsistency [3–5]. Similar differences between subjective judgments and observable indicators have also been documented in community and policy evaluation, where perceived outcomes may depart from objective welfare measures [6, 7]. These findings have motivated evaluation frameworks that compare subjective assessments with independently observable evidence rather than treating the final judgment as self-explanatory [8–10].

Sports provide a particularly useful setting for studying such discrepancies because performance is extensively recorded while final evaluations may still depend on human judgment. Previous research has shown that sports-related decisions can be associated with social and contextual factors, including racial mismatch between decision-makers and athletes [11], institutional or social ties [12], home-crowd pressure [13], and conformity to prevailing group assessments in judged sports [14]. These studies demonstrate that subjective decisions may differ from what would be expected from directly

observable performance information. However, most focus on identifying particular sources of bias or explaining specific decision mechanisms. The present study addresses a related but different problem: rather than estimating the causal effect of any specific contextual factor, we construct a performance-based benchmark without learning from the voting outcome being examined and use that benchmark to identify where the voting-based outcome agrees with or departs from measurable evidence.

2.2. Machine learning for sports ranking and award evaluation

Machine learning has become increasingly important in sports analytics because it provides tools for summarizing high-dimensional performance information and identifying patterns beyond traditional statistics. In basketball, player evaluation has evolved from summary measures such as the player efficiency rating (PER) [15] and adjusted plus-minus [16] toward broader machine-learning approaches for performance analysis and prediction [17, 18]. Explainable modeling techniques have also been increasingly applied to reveal the performance characteristics associated with complex predictive models and to improve model transparency [19, 20].

However, prediction and evaluation auditing address different objectives. Supervised models trained on historical MVP outcomes can reproduce voting patterns or predict future winners, but they directly learn from the subjective decisions that an auditing framework aims to examine. Therefore, constructing a performance-based benchmark requires methods that can extract information from observed statistics without relying on historical award outcomes.

Unsupervised feature-relevance methods provide one basis for estimating variable contributions from data structure without outcome labels [21], while interpretable modeling methods offer rule-based or feature-based explanations of complex relationships [19, 20]. These approaches provide complementary methodological components: feature-relevance estimation supports data-driven weighting, whereas interpretable models support the analysis of resulting evaluation patterns. However, integrating these components with multi-criteria player-ranking and external comparison against subjective recognition remains insufficiently explored.

2.3. Objective weighting and consensus ranking

The construction of a multi-indicator ranking also requires a mechanism for determining the relative contribution of different criteria. This problem is closely related to criterion weighting in multi-criteria decision-making (MCDM), where weighting approaches are commonly classified as subjective, objective, or hybrid according to whether criterion importance is derived from expert judgment, observed data, or a combination of the two [22, 23]. In this literature, "objective weighting" refers to weights derived from the observed decision matrix rather than from externally supplied preference judgments, and does not imply that the resulting evaluation represents a value-free ground truth. Such methods are particularly relevant to the present study because they can generate criterion weights without using historical MVP voting outcomes as supervisory information. The Entropy Weight method derives criterion importance from the information dispersion of observed values, CRITIC combines criterion contrast with inter-criterion dependence [24], MEREC evaluates importance through the effect of removing individual criteria [25], and LOPCOW derives weights from logarithmic percentage-change information [26]. These methods provide representative rather than exhaustive comparison mechanisms for deriving criterion importance from unlabeled data.

A different strategy is to combine rankings after they have already been produced. Borda Count is a classical consensus-ranking procedure that assigns points according to rank position and aggregates several input rankings into a single ordering [27]. This provides a useful conceptual contrast with UFER: Borda Count combines separately generated component rankings at the ranking level, whereas UFER integrates feature-importance information before player scores and season-specific rankings are constructed. Supervised learning-to-rank methods generally require relevance labels, preference pairs, or historical rankings, while subjective MCDM procedures require externally specified expert judgments or preference information. These assumptions are less suitable when the purpose is to construct a benchmark for examining subjective recognition. Therefore, existing weighting and consensus-ranking methods provide important components for data-driven evaluation, but the integration of criterion weighting, ranking aggregation, and interpretable analysis for subjective award evaluation remains an open methodological issue.

3. Methods

3.1. System overview and workflow

As shown in Figure 1, the proposed system constructs a data-driven performance benchmark for NBA MVP evaluation. The system takes NBA player-season statistics as input and produces two main outputs: a label-free performance ranking and an interpretable diagnostic comparison with official voting outcomes. First, the unsupervised feature-weighted ranking module integrates unsupervised random forest, Genie3, and URelief to estimate feature importance and generate season-specific player rankings. The top-ranked players are then treated as data-driven MVP candidates. Second, the diagnostic comparison module uses RuleFit and Skope-Rules to summarize the player profiles associated with different labeling schemes. The resulting ranking provides a reference for comparing performance-based evaluation with voting-based outcomes.

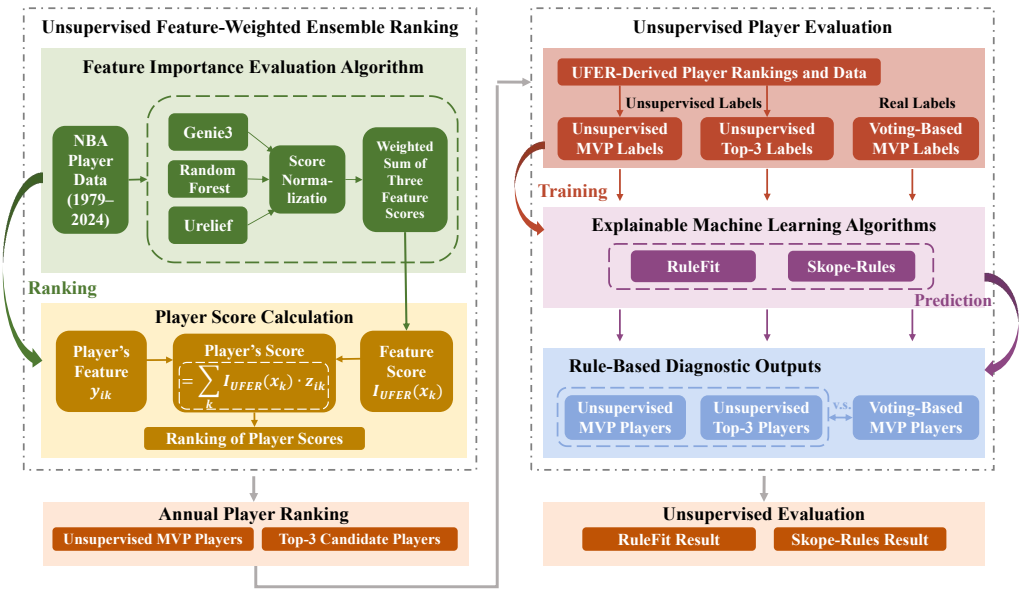

Figure 1. Workflow of the proposed UFER-based auditing system.

3.2. Data and feature description

To more accurately assess the performance differences among elite players, this study uses a precompiled candidate file of high-performing player-season records covering the 1979–80 through 2024–25 NBA seasons. The candidate file was assembled before model estimation from public NBA statistical records and includes the seasonal MVP winner in the candidate sample. The analysis is therefore restricted to the evaluated candidate sample rather than the full league population. Here, "label-free" refers specifically to model estimation and ranking construction: historical MVP labels are not used to estimate feature weights, calculate player scores, or select ensemble coefficients. The ranking pipeline does not further alter candidate membership, and no additional minutes-played, games-played, or per-season top-N filtering rule is imposed inside the pipeline reported in this manuscript. The analysis file stores these seasons using season-start labels from 1979 to 2024; throughout the manuscript, seasons are reported by the corresponding award year from 1980 to 2025. The data were obtained from the public databases `https://www.nba-stat.com/` and `https://www.nba.com/stats`.

All ranking and robustness experiments use the same analytical sample. The initial candidate file contained 1000 player-season records. After records containing missing values in any of the 43 variables used for analysis were removed, the final analytical dataset contained 934 player-season observations and 43 performance features across 46 seasons.

The feature set combines basic box-score statistics, advanced contribution metrics, and team wins to capture multiple dimensions of player performance and team context. For the purpose of constructing the final player-scoring matrix, criterion directions are defined consistently across all feature-weighting methods. Turnovers (TOV), turnover percentage (TOV%), and personal fouls (PF) are treated as cost criteria, for which lower values are preferable. The cost designation is restricted to variables whose undesirable direction is unambiguous; the remaining 40 variables are treated as performance or production indicators under the monotonic scoring convention used in the present benchmark.

Let $C = \{\text{TOV}, \text{TOV\%}, \text{PF}\}$ denote the set of cost criteria. The direction-aligned feature value is defined as

$$x_{ij}^{\text{dir}} = \begin{cases} -x_{ij}, & j \in C, \\ x_{ij}, & j \notin C. \end{cases}$$

After criterion directions are aligned, each feature is standardized over the cleaned analytical sample using

$$z_{ij} = \frac{x_{ij}^{\text{dir}} - \mu_j}{\sigma_j},$$

where μ_j and σ_j denote the mean and standard deviation of the direction-aligned values of feature j, respectively. The resulting matrix $\mathbf{Z} = [z_{ij}]$ is used as the common downstream player-scoring matrix for UFER and the feature-weighting baselines. Method-specific normalization used internally for estimating criterion weights is described separately below.

The player characteristics used in the empirical analysis are listed in Table 1.

Table 1. Player characteristics used in the empirical analysis.

	Meaning		Meaning
FG	Field Goal	FGA	Field Goal Attempts
FG%	Field Goal Percentage	3P	3-Point Field Goal
3PA	3-Point Field Goal Attempts	3P%	3-Point Field Goal Percentage
2P	2-Point Field Goal	2PA	2-Point Field Goal Attempts
2P%	2-Point Field Goal Percentage	eFG%	Effective Field Goal Percentage
FT	Free Throw	FTA	Free Throw Attempts
FT%	Free Throw Percentage	ORB	Offensive Rebounds
DRB	Defensive Rebounds	TRB	Total Rebounds
AST	Assists	STL	Steals
BLK	Blocks	TOV	Turnovers
PF	Personal Foul	PTS	Points
PER	Player Efficiency Rating	TS%	True Shooting Percentage
3PAr	3-Point Attempt Rate	FTr	Free Throw Rate
ORB%	Offensive Rebound Percentage	DRB%	Defensive Rebounds Percentage
TRB%	Total Rebounds Percentage	AST%	Assist Percentage
STL%	Steal Percentage	BLK%	Block Percentage
TOV%	Turnover Percentage	USG%	Usage Rate
OWS	Offensive Win Shares	DWS	Defensive Win Shares
WS	Win Shares	WS/48	Win Shares Per 48 Minutes
OBPM	Offensive Box Plus/Minus	DBPM	Defensive Box Plus/Minus
BPM	Box Plus/Minus	VORP	Value Over Replacement Player
WIN	The Team's Regular Season Win Total		

3.3. Unsupervised feature-weighted ranking module

Because player evaluation involves multiple performance dimensions without a predefined objective importance structure, UFER employs unsupervised feature relevance methods to estimate criterion importance from the data itself. This section first introduces three core unsupervised algorithms, unsupervised random forest (URF), Genie3, and URelief, focusing on their respective mechanisms for feature selection and importance estimation [21]. Subsequently, these methods are integrated into a unified player-ranking module.

3.3.1. Unsupervised feature-ranking components

The unsupervised feature-weighted ensemble ranking module combines three complementary estimators of unsupervised feature relevance. URF measures the deterioration in out-of-bag predictive performance caused by feature permutation, Genie3 accumulates split-based impurity reduction, and URelief contrasts feature differences between nearby and less-similar observations. Their respective formulations are summarized below.

Unsupervised random forest (URF). We begin by introducing the predictive clustering tree (PCT), a tree-based method that recursively partitions the data into homogeneous clusters. A training subset $\mathscr{D}_t$ is used to grow the tree. At each node, a subset $E \subseteq \mathscr{D}_t$ is evaluated for splitting. The algorithm

examines each feature x_k to determine the optimal binary split. The impurity of a node E is defined as the average normalized variance across all features:

$$\text{impu}(E) = \frac{1}{p}\sum_{k=1}^{p}\frac{\text{Var}(E, x_k)}{\text{Var}(\mathscr{D}_t, x_k)}.$$

The improvement achieved by a split x_k is quantified as

$$h(E, x_k) = |E| \cdot \text{impu}(E) - \left(|E_1| \cdot \text{impu}(E_1) + |E_2| \cdot \text{impu}(E_2)\right).$$

The maximum value of h obtained among all x_k is denoted as h_{max}. Here, node impurity is measured by the average normalized variance of the features in the node, so lower values indicate more homogeneous partitions. In intuitive terms, $h(E, x_k)$ compares heterogeneity before and after a candidate split, and a larger value indicates a clearer partition.

The URF method extends the PCT principle to an ensemble setting and uses permutation-based out-of-bag evaluation to estimate feature importance. For each clustering tree $\mathcal{T}$ in the forest $\mathcal{E}$, the out-of-bag sample $OOB_{\mathcal{T}}$ is used as an independent evaluation set. The perturbed set $OOB_{\mathcal{T}}^{(k)}$ is obtained by randomly permuting feature x_k within the out-of-bag observations while leaving the fitted tree unchanged.

Let $\hat{x}_{ij}^{(\mathcal{T})}$ denote the prediction produced by tree $\mathcal{T}$ for feature x_j of observation i. Consistent with the numerical implementation, the prediction error of tree $\mathcal{T}$ on an evaluation set S is defined as the average root-mean-square error across the p numerical features:

$$e_{\mathcal{T}}(S) = \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{1}{|S|}\sum_{i\in S}\left(x_{ij} - \hat{x}_{ij}^{(\mathcal{T})}\right)^2}.$$

The permutation importance of feature x_k is then calculated as the average relative increase in out-of-bag prediction error across all trees:

$$\text{importance}_{\text{URF}}(x_k) = \frac{1}{|\mathcal{E}|}\sum_{\mathcal{T}\in\mathcal{E}}\frac{e_{\mathcal{T}}\left(OOB_{\mathcal{T}}^{(k)}\right) - e_{\mathcal{T}}(OOB_{\mathcal{T}})}{e_{\mathcal{T}}(OOB_{\mathcal{T}})}.$$

Thus, URF importance measures the deterioration in out-of-bag predictive performance after feature permutation, rather than a change in the marginal variance of the permuted feature. A larger importance score indicates that permuting x_k causes a greater loss of predictive accuracy and that the feature contributes more strongly to the unsupervised partitioning structure.

Genie3. Genie3 is a random-forest-based feature-importance ranking algorithm initially designed for supervised feature selection. Based on the unsupervised random forest ensemble $\mathcal{E}$ comprising M cluster trees, for each feature x_k, we identify the set of nodes $T(x_k)$ across all trees where x_k was used as the splitting attribute. For each feature node N in $T(x_k)$, E_N is the sample set that reaches node N. Therefore, we use the optimal impurity reduction value $h_{\text{best}}(E_N, x_k)$ computed during tree construction.

The importance score for feature x_k is then calculated as the average cumulative impurity reduction attributable to splits on x_k across the entire forest:

$$\text{importance}_{\text{Genie3}}(x_k) = \frac{1}{|\mathcal{E}|}\sum_{\mathcal{T}\in\mathcal{E}}\sum_{N\in T(x_k)} h_{\text{best}}(E_N, x_k).$$

This formulation effectively measures how frequently and effectively each feature contributes to creating a cluster tree throughout the entire forest. Thus, Genie3 records the cumulative split contribution, whereas URF measures the increase in out-of-bag prediction error caused by feature permutation.

URelief. Following the tree-based approaches, we introduce URelief, an unsupervised extension of the Relief algorithm. URelief evaluates feature importance by comparing the feature-wise differences between similar and less-similar observations. Features that remain stable among nearby observations but vary more strongly across less-similar observations are regarded as more informative for capturing the underlying data structure.

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ denote two player-season observations. The normalized distance on feature x_k is defined as

$$d_k(x_i, x_j) = \frac{|x_{ik} - x_{jk}|}{\max_s x_{sk} - \min_s x_{sk} + \epsilon},$$

where $\epsilon > 0$ is a small constant used to avoid division by zero. The overall distance between x_i and x_j is then computed as the average feature-wise distance:

$$d(x_i, x_j) = \frac{1}{p} \sum_{k=1}^{p} d_k(x_i, x_j).$$

For a randomly selected observation x_i, let $\mathcal{N}_i^{\text{near}}$ denote the set of its m nearest neighbors according to $d(x_i, x_j)$, and let $\mathcal{N}_i^{\text{far}}$ denote a comparison set of m less-similar observations. The average feature-wise difference of feature x_k within each set is defined as

$$\overline{d}_k(x_i, \mathcal{N}_i^{\text{near}}) = \frac{1}{m} \sum_{x_j \in \mathcal{N}_i^{\text{near}}} d_k(x_i, x_j)$$

and

$$\overline{d}_k(x_i, \mathcal{N}_i^{\text{far}}) = \frac{1}{m} \sum_{x_j \in \mathcal{N}_i^{\text{far}}} d_k(x_i, x_j).$$

This process is repeated for r randomly selected observations. The URelief importance score of feature x_k is then computed as

$$\text{importance}_{\text{URelief}}(x_k) = \frac{1}{r} \sum_{i=1}^{r} \left[\overline{d}_k(x_i, \mathcal{N}_i^{\text{far}}) - \overline{d}_k(x_i, \mathcal{N}_i^{\text{near}}) \right].$$

A larger value of $\text{importance}_{\text{URelief}}(x_k)$ indicates that feature x_k better distinguishes less-similar observations while remaining relatively stable among nearby observations. Therefore, it is assigned a higher feature-ranking score in the URelief module. This near–far contrast provides a non-tree-based complement to the two forest-derived scores.

3.3.2. Unsupervised feature-weighted ensemble ranking algorithm

The unsupervised feature-weighted ensemble ranking (UFER) framework integrates three feature-ranking methods: unsupervised random forest, Genie3, and URelief. These methods characterize the data structure through different mechanisms, including permutation-based partitioning, cumulative tree-split improvement, and neighborhood-based contrast. The aggregated weights combine different perspectives of feature relevance and provide a comprehensive representation of criterion importance. The sensitivity of the component importance estimates, integrated UFER weights, and final rankings to correlated and redundant features is examined separately in Section 4.4.3.

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the statistical profile of player-season observation i, where p is the number of performance indicators. For the final player-scoring stage, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$ denotes the corresponding direction-aligned and standardized profile defined in the Data and Feature Description section. Let $I_m(x_k)$ denote the importance score of feature x_k estimated by method m, where

$$m \in \{\text{URF}, \text{Genie3}, \text{URelief}\}.$$

The UFER framework consists of the following steps:

(1) **Standardization.** Since different feature-ranking methods may generate importance scores on different scales, the scores from each method are first normalized to a common range:

$$\tilde{I}_m(x_k) = \frac{I_m(x_k)}{\max_j I_m(x_j)}.$$

Here, $\tilde{I}_m(x_k)$ denotes the standardized importance score of feature x_k under method m.

(2) **Ensemble weighting.** The standardized scores are then aggregated to obtain the UFER feature weight:

$$I_{\text{UFER}}(x_k) = \omega_{\text{URF}}\tilde{I}_{\text{URF}}(x_k) + \omega_{\text{Genie3}}\tilde{I}_{\text{Genie3}}(x_k) + \omega_{\text{URelief}}\tilde{I}_{\text{URelief}}(x_k),$$

where

$$\omega_{\text{URF}} + \omega_{\text{Genie3}} + \omega_{\text{URelief}} = 1, \quad \omega_{\text{URF}}, \omega_{\text{Genie3}}, \omega_{\text{URelief}} \geq 0.$$

Because the ensemble coefficients are nonnegative and sum to one, $I_{\text{UFER}}(x_k)$ is a convex combination of the normalized component importance scores. Consequently, for every feature x_k,

$$\min_m \widetilde{I}_m(x_k) \leq I_{\text{UFER}}(x_k) \leq \max_m \widetilde{I}_m(x_k).$$

The aggregation therefore combines their feature-relevance assessments while controlling the contribution of each component through weights $\boldsymbol{\omega}$.

In the absence of reliable prior evidence for assigning greater importance to any one of the three component methods, we adopt the equal-weight specification as the default ensemble scheme. Over the feasible simplex of nonnegative coefficients summing to one, the equal-weight vector $(1/3, 1/3, 1/3)$ uniquely maximizes Shannon entropy and minimizes coefficient concentration. It therefore provides the most balanced specification and avoids introducing an additional preference for URF, Genie3, or URelief.

(3) **Feature ranking.** Features are ranked in descending order according to $I_{\mathrm{UFER}}(x_k)$. A larger value indicates that the corresponding performance indicator contributes more strongly to the integrated ranking benchmark.

(4) **Sample scoring.** For each player-season observation i, the UFER ranking score is calculated using the common direction-aligned and standardized scoring matrix:

$$S_i^{\mathrm{UFER}} = \sum_{k=1}^{p} I_{\mathrm{UFER}}(x_k) z_{ik},$$

where z_{ik} denotes the standardized value of feature x_k after criterion-direction alignment. Using the same scoring matrix across weighting methods separates differences in feature-weight estimation from differences caused by feature scale or criterion direction. The score S_i^{UFER} represents the player's overall performance level under the integrated feature-weighting scheme.

(5) **Ranking and pseudo-label construction.** Within each season, players are ranked in descending order according to their UFER scores. The top-ranked player is treated as the data-driven MVP candidate of that season, and the top three players are extracted as the UFER-based candidate set. For the subsequent RuleFit-based diagnostic analysis, players in the UFER top three are assigned the pseudo-label $y_i = +1$, while the remaining players are assigned $y_i = -1$. These pseudo-labels are used only in the post-ranking interpretability module and are not used in the unsupervised feature-weighting stage.

The UFER framework therefore provides a transparent performance-based ranking benchmark by combining multiple unsupervised feature-ranking mechanisms. The ranking results are subsequently converted into pseudo-labels for rule-based diagnostic analysis.

3.4. Interpretable diagnostic comparison module

After ranking construction, RuleFit and Skope-Rules are employed to extract interpretable patterns associated with highly ranked players. Within the proposed system, these methods form a diagnostic layer rather than the primary ranking engine. They are used to examine whether different labeling schemes identify similar MVP-like player profiles, and their separate outputs preserve information about agreement and disagreement across distinct rule-induction mechanisms.

3.4.1. Labeling schemes

The diagnostic module uses three labeling schemes for MVP identification:

(1) **Historical MVP labels**: the official voting-based outcomes of the MVP award process.

(2) **Data-driven MVP labels**: pseudo-labels generated by selecting each season's top-ranked player according to the UFER ranking.

(3) **Data-driven top-3 labels**: a broader data-driven benchmark generated by selecting the top-three players in each season.

This triple-label design enables RuleFit and Skope-Rules to be trained under identical feature sets while varying the supervisory signals. By comparing diagnostic coverage across labeling schemes, we assess the consistency between historical voting-based MVP outcomes and data-driven performance assessment. The complementary strengths of RuleFit's probabilistic scoring and Skope-Rules' binary classification provide multidimensional evidence on whether voting outcomes align with data-driven MVP-like profiles.

3.4.2. RuleFit and Skope-Rules algorithms

RuleFit integrates tree-based ensemble learning with sparse linear modeling to create an interpretable diagnostic classifier [28]. The algorithm operates through three key components.

(1) Rule Generation. Each root-to-leaf path in the ensemble defines a binary, threshold-based rule:

$$\mathcal{R}_r(x) = \prod_{k=1}^{p} \mathbf{1}\{x_k \in s_{kr}\} \in \{0, 1\}, \qquad r = 1, 2, \cdots, R,$$

where s_{kr} defines the value interval or range that the k-th feature x_k must satisfy within rule r.

(2) Predictor Construction. The RuleFit model combines these rules with optional linear terms from the original features:

$$F(x) = a_0 + \sum_{r=1}^{R} a_r \mathcal{R}_r(x) + \sum_{k=1}^{p} b_k \mathcal{L}_k(x_k),$$

with $\mathcal{L}_k(\cdot)$ denoting truncated features and (a_r, b_k) the contribution of each rule and feature to the prediction.

(3) Sparse Learning. For binary MVP classification ($y_i \in \{-1, +1\}$), coefficients are estimated by minimizing an L_1-regularized empirical loss:

$$\min_{a_0, \{a_r\}, \{b_k\}} \sum_{i=1}^{n} L(y_i, F(x_i)) + \lambda \left(\sum_{r=1}^{R} |a_r| + \sum_{k=1}^{p} |b_k| \right),$$

where $L(\cdot, \cdot)$ is a standard logistic loss and $\lambda > 0$ controls sparsity.

The predicted probability of being an MVP is obtained via logistic transformation:

$$p(x) = \sigma(F(x)) = \frac{1}{1 + e^{-F(x)}} \in [0, 1].$$

A threshold $\tau \in (0, 1)$ is applied to assign final MVP/non-MVP labels.

This combination of rule-based feature engineering and sparse linear modeling makes RuleFit suitable for diagnostic comparison because it yields probabilistic MVP-like scores under different labeling schemes.

Skope-Rules is used as a complementary rule-based diagnostic procedure. In the implementation used in this study, it operates on candidate rules and applies precision-based rule selection together with support-similarity pruning, providing a discrete rule-based complement to RuleFit's probabilistic scoring.

(1) Rule Generation. The same candidate rules $\mathcal{R}_r$ used in the RuleFit-based diagnostic stage are used as the initial rule set.

(2) Rule Selection via Precision Optimization. For each candidate rule $\mathcal{R}_r$, we compute its precision score:

$$\text{precision}(\mathcal{R}_r) = \frac{\text{TP}(\mathcal{R}_r)}{\text{TP}(\mathcal{R}_r) + \text{FP}(\mathcal{R}_r)},$$

where $\text{TP}(\mathcal{R}_r)$ and $\text{FP}(\mathcal{R}_r)$ denote true and false positives. Rules are retained only if their precision exceeds a predefined threshold $\tau_{\text{precision}}$.

(3) Diversity-Preserving Rule Pruning. The Jaccard index measures the similarity between two rules, $\mathcal{R}_{r1}$ and $\mathcal{R}_{r2}$:

$$\text{sim}(\mathcal{R}_{r1}, \mathcal{R}_{r2}) = \frac{|supp(\mathcal{R}_{r1}) \cap supp(\mathcal{R}_{r2})|}{|supp(\mathcal{R}_{r1}) \cup supp(\mathcal{R}_{r2})|},$$

where

$$\text{supp}(\mathcal{R}_r) = \{x_i : \mathcal{R}_r(x_i) = 1\}$$

represents the support of rule $\mathcal{R}_r$. Rules with similarity exceeding threshold τ_{sim} are consolidated to avoid redundancy. A player is designated as an MVP candidate if they satisfy at least one retained rule, i.e., $\text{MVP}(x_i) = 1$ when $\max_r \mathcal{R}_r(x_i) = 1$, and $\text{MVP}(x_i) = 0$ otherwise. Compared to RuleFit's continuous scoring mechanism, Skope-Rules offers crisp classification boundaries, making it useful for comparing which players are covered by rule-induced MVP-like profiles under different labeling schemes.

In the present study, RuleFit and Skope-Rules are used for within-sample diagnostic characterization rather than as out-of-sample predictive benchmarks. The same diagnostic model classes and feature representation are used across the three labeling schemes, so that the comparison focuses on differences associated with the target definitions.

4. Experimental results

4.1. Evaluation setup

The experimental evaluation examines three aspects of the proposed framework: (1) the robustness of the UFER ranking under different perturbation settings, (2) the agreement and divergence between the resulting performance-based benchmark and historical voting-based outcomes, and (3) the stability of the main conclusions under alternative ensemble-weight configurations, label-free comparison methods, and correlated-feature perturbations.

The analytical dataset contains 934 player-season observations and 43 performance-related features covering the 1979–80 through 2024–25 NBA seasons, corresponding to award years 1980–2025. All experiments use the same cleaned analytical sample and the criterion-direction specification defined in Section 3. TOV, TOV%, and PF are treated as cost criteria, while the remaining variables are treated as benefit criteria. For methods based on feature weighting, player scores are evaluated on the common direction-aligned and standardized scoring matrix described in the Methods section, while method-specific normalization is used only where required for estimating criterion weights.

Ranking robustness is evaluated using Spearman's rank correlation, top-k agreement, NDCG@3, and mean absolute rank shift (MARS), depending on the corresponding experiment. Repeated perturbation analyses are conducted under feature removal, statistical-group deletion, irrelevant-feature injection, and correlated-feature ablation. For the diagnostic analysis, RuleFit and

Skope-Rules are applied after ranking construction to characterize the player profiles associated with performance-based and voting-based labels.

The proposed UFER ranking is evaluated against its three component methods, URF, Genie3, and URelief, and against additional weighting and rank-aggregation baselines. Historical MVP outcomes are considered only in the subsequent external agreement and diagnostic analyses.

4.2. Performance of the proposed UFER ranking framework

4.2.1. Robustness and component complementarity analysis

Robustness is evaluated by comparing perturbed rankings with the corresponding full-information ranking rather than with external award outcomes. Let $\pi^0_{q,s}$ denote the ranking produced by method q for season s using the complete original feature set, and let $\pi^b_{q,s}$ denote the ranking obtained under perturbation setting b. The ranking $\pi^0_{q,s}$ is used as the reference for method q in season s.

We use three perturbation settings. In feature perturbation, a proportion $\alpha \in \{0.4, 0.5, 0.6, 0.7\}$ of the original features is randomly retained. In group deletion, one predefined statistical group, such as scoring, shooting efficiency, rebounding, play-making, defense, advanced impact, or team context, is removed at a time. In irrelevant-feature injection, $c \in \{20, 40, 60\}$ additional nuisance feature columns are appended to the original feature matrix. To reflect the correlated and redundant structure of NBA statistics, the injected nuisance variable for player-season observation i and injected feature j is generated as $\nu_{ij} = u_{g(j),i} + \epsilon_{ij}, u_{g(j),i} \sim \mathcal{N}(0, 1), \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, where $g(j)$ maps injected feature j to a latent nuisance block. In our experiments, two latent blocks are used and $\sigma = 0.5$. These nuisance variables are generated independently of player performance, voting outcomes, and model rankings.

For each season, we evaluate ranking preservation using three metrics. Let $T^0_{q,s}(3)$ and $T^b_{q,s}(3)$ be the top-3 player sets under the reference and perturbed rankings, respectively. The top-3 agreement is defined as

$$A^b_{q,s}(3) = \frac{|T^0_{q,s}(3) \cap T^b_{q,s}(3)|}{3}.$$

To measure whether the ordering of the most relevant candidates is preserved, we also compute NDCG@3. The relevance score is assigned from the reference ranking as $\mathrm{rel}_{q,s}(y) = 4 - r^0_{q,s}(y)$ if player y is ranked in the reference top-3, and $\mathrm{rel}_{q,s}(y) = 0$ otherwise, where $r^0_{q,s}(y)$ is the reference rank of player y. Thus,

$$\mathrm{NDCG}^b_{q,s}@3 = \frac{\sum_{\ell=1}^{3} \frac{\mathrm{rel}_{q,s}(y^b_\ell)}{\log_2(\ell+1)}}{\sum_{\ell=1}^{3} \frac{4-\ell}{\log_2(\ell+1)}},$$

where y^b_ℓ is the player ranked at position ℓ in the perturbed ranking. Finally, overall rank-order consistency is measured by Spearman's rank correlation,

$$\rho^b_{q,s} = \mathrm{corr}\left(r^0_{q,s}, r^b_{q,s}\right).$$

All perturbation experiments use a paired design. Within each repetition, the same feature-retention mask, deleted statistical group, or nuisance-feature matrix is shared by the compared methods. After each perturbation, the corresponding feature weights and season-specific player rankings are recomputed, and the perturbed ranking is compared with the same method's full-information ranking.

Top-1 agreement is retained only as an auxiliary diagnostic in the detailed outputs and is not used to calculate the reported average robustness ranks.

In the component analysis, UFER is compared with URF, Genie3, and URelief under the three perturbation regimes. The feature-perturbation experiment uses 40%, 50%, 60%, and 70% feature-retention levels with $B = 200$ repetitions per level. The group-deletion experiment removes one coherent statistical group at a time, and the irrelevant-feature injection experiment appends 20, 40, or 60 correlated nuisance features with $B = 30$ repetitions per level. For each regime, competing methods are ranked by their mean top-3 agreement, NDCG@3, and Spearman correlation across seasons and repetitions. Tables 2 and 3 report the metric-level scores and the corresponding average robustness ranks.

Table 2. Metric-level robustness of UFER components.

Metric	UFER	Genie3	URF	URelief
Feature perturbation				
Top-3 agreement	0.835	0.822	0.823	**0.840**
NDCG@3	0.887	0.874	0.874	**0.892**
Spearman	0.949	0.945	0.944	**0.951**
Group deletion				
Top-3 agreement	**0.895**	0.877	0.889	0.880
NDCG@3	**0.933**	0.916	0.924	0.927
Spearman	0.967	0.963	0.965	**0.970**
Irrelevant-feature injection				
Top-3 agreement	0.462	0.486	**0.487**	0.413
NDCG@3	0.478	0.505	**0.506**	0.423
Spearman	0.605	0.634	**0.638**	0.536

Note: Values are mean robustness scores.

Table 3. Average robustness ranks of UFER components.

Perturbation regime	UFER	Genie3	URF	URelief
Feature perturbation	2.00	3.25	3.75	1.00
Group deletion	1.69	2.88	2.50	2.93
Irrelevant-feature injection	3.00	1.67	1.33	4.00

Table 3 shows that the three base rankers exhibit different robustness profiles: URelief performs best under random feature perturbation, URF performs strongly under group-level deletion, and Genie3/URF are more robust under irrelevant-feature injection. Although UFER is not the best method in every setting, it remains consistently competitive, supporting the use of a heterogeneous ensemble rather than a single unsupervised ranking mechanism.

4.2.2. Feature importance and ranking characteristics

After establishing the robustness of the UFER ranking module, we further examine the feature signals that drive the integrated ranking.

Table 4 reports the top 10 features after UFER integration. The component methods emphasize related but nonidentical statistical signals, and the integrated UFER weighting assigns relatively high importance to rebounding, assist-related measures, free-throw pressure, turnover-related variables, team wins, and win-contribution indicators. The resulting importance profile therefore reflects multiple dimensions of player performance rather than being concentrated on a single statistical category. The integrated values also clarify why these variables appear near the top of the UFER feature ranking. Because UFER aggregates normalized component importance scores, a high final weight reflects strong support from multiple estimators or a particularly high score from one estimator combined with non-negligible support from the others. For example, DRB% receives normalized importance scores of 0.9324 from URF and 0.9289 from URelief; TRB receives a Genie3 score of 0.9237 while retaining scores of 0.6789 and 0.6345 under URF and URelief, respectively; and FTr receives scores of 0.9905 and 0.9774 under URF and URelief. Thus, the highest UFER weights arise from the aggregation of heterogeneous feature-relevance signals rather than from a manually imposed preference for a particular basketball statistic. These weights should therefore be interpreted as feature relevance within the observed multivariate representation rather than as isolated causal contributions.

Table 4. Top-ranked features identified by the UFER module.

Feature	Genie3	URF	URelief	UFER	Ranking
DRB%	0.42180	0.9324	0.9289	0.7610	1
TRB	0.9237	0.6789	0.6345	0.7457	2
FTr	0.1479	0.9905	0.9774	0.7053	3
2PA	1.0000	0.6385	0.4711	0.7032	4
TOV%	0.3236	0.7698	1.0000	0.6978	5
AST	0.7479	0.6977	0.6278	0.6911	6
WS	0.7827	0.6443	0.6418	0.6896	7
WIN	0.1197	0.9694	0.9035	0.6642	8
AST%	0.2853	0.9699	0.7249	0.6601	9
WS/48	0.6826	0.7502	0.5433	0.6587	10

Note: Component importance values are standardized to the common scale defined in the UFER integration procedure.

Because several basketball statistics are strongly related, the feature-importance values in Table 4 should be read as the weighting structure under the complete feature representation. Their sensitivity to correlated alternatives and the corresponding ranking-level robustness are examined in Section 4.4.3.

Figure 2 further shows strong heterogeneity and skewness across high-variance variables, especially in assist involvement, free-throw usage, three-point activity, and rebounding. This long-tailed structure indicates substantial variation among elite player-season profiles and further illustrates why player evaluation depends on combinations of multiple statistical signals rather than on a single performance indicator.

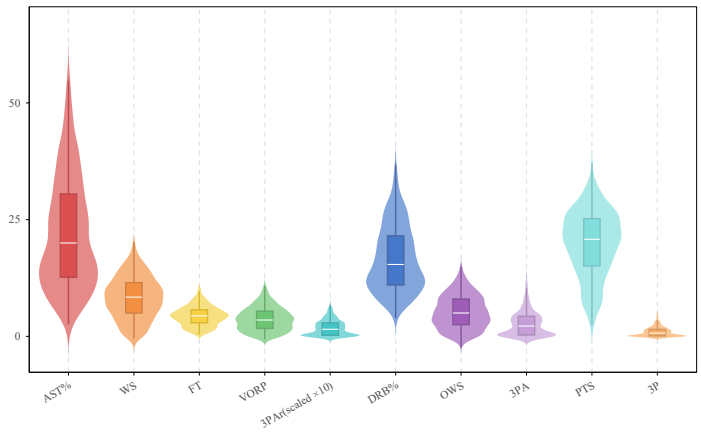

Figure 2. Distribution of the top 10 features by variance.

4.3. Auditing MVP voting decisions through interpretable comparison

4.3.1. Data-driven MVP evaluation results

We first examine whether the official MVP winner is covered by the player profiles learned from the data-driven top-3 labels. RuleFit and Skope-Rules are used as interpretable diagnostic classifiers, and the comparison focuses on coverage consistency between the performance-based labels and official voting outcomes. Across the 46 seasons, both diagnostic models cover the official MVP winner in 17 seasons (37.0%). In 16 seasons, the winner is covered by one model but not the other, while in 13 seasons neither model covers the winner. Among the latter group, four particularly notable cases are Allen Iverson in 2001, Steve Nash in 2005 and 2006, and Kobe Bryant in 2008.

To illustrate these coverage patterns, Table 5 presents selected cases under the data-driven top-3 labeling scheme in which at least one diagnostic model does not cover the official MVP winner.

Several of these seasons have also been widely discussed as debated MVP selections, making them useful examples for interpreting the divergence identified by the diagnostic models. In 2001, Allen Iverson received the official award, whereas the data-driven diagnostic profiles more strongly cover players such as Shaquille O'Neal, Karl Malone, and Tim Duncan. The 2005 and 2006 Nash selections similarly illustrate the tension between statistical production and broader notions of value, while the 2008 season presents another case in which the diagnostic profiles emphasize players such as Chris Paul and LeBron James rather than the official winner.

These discrepancies do not imply that the official selections were incorrect. Rather, they show that the voting outcome does not always coincide with the player profiles most strongly supported by the performance-based benchmark used in this study. MVP voting may also reflect considerations that are not fully represented by the player-level statistical indicators, including team success, leadership, season narrative, reputation, prior recognition, or the perceived value of a player's role within a successful team. Such considerations may help contextualize the observed differences, although the present analysis does not separately identify or quantify their contribution to any particular voting outcome.

Table 5. Diagnostic profile coverage under data-driven top-3 labels.

Year	Official MVP	RuleFit	Skope-Rules
1984	Larry Bird	Larry Bird	Adrian Dantley
		Moses Malone	Moses Malone
2001	Allen Iverson	Shaquille O'Neal	Shaquille O'Neal
		Karl Malone	Karl Malone
		Tim Duncan	Vince Carter
2005	Steve Nash	Tim Duncan	Dirk Nowitzki
		Dirk Nowitzki	Kevin Garnett
		Dwyane Wade	LeBron James
2006	Steve Nash	LeBron James	Kobe Bryant
		Kobe Bryant	Dirk Nowitzki
		Dirk Nowitzki	Dwyane Wade
			Kevin Garnett
2008	Kobe Bryant	LeBron James	Chris Paul
		Chris Paul	Kevin Garnett
		Dwyane Wade	LeBron James

Note: Selected illustrative cases are shown in which at least one diagnostic model does not cover the official MVP winner. Listed names are same-season candidates covered by the diagnostic rules.

Table 6 considers the narrower data-driven MVP labeling scheme, in which only the top-ranked UFER player in each season is assigned to the positive class. Because this definition is more restrictive than the top-3 labeling scheme, the resulting diagnostic profiles produce more frequent disagreement with the official voting-based outcome. The 2011 and 2015 seasons illustrate this pattern: Derrick Rose and Stephen Curry were the official winners, whereas the corresponding data-driven diagnostic results emphasize LeBron James and Anthony Davis, respectively. These cases provide additional examples of divergence between the performance-based benchmark and the voting-based outcome without identifying the specific contextual factors responsible for the historical outcome.

Table 6. Diagnostic profile coverage under data-driven MVP labels.

Year	Historical MVP	RuleFit	Skope-Rules
2025	Gilgeous-Alexander	Gilgeous-Alexander	Antetokounmpo
		Antetokounmpo	Gilgeous-Alexander
		Nikola Jokić	Nikola Jokić
2023	Joel Embiid	Joel Embiid	Nikola Jokić
		Nikola Jokić	
2019	Antetokounmpo	Antetokounmpo	James Harden
2018	James Harden		James Harden
2017	Russell Westbrook		Russell Westbrook
2015	Stephen Curry	Anthony Davis	

Continued on next page

Year	Historical MVP	RuleFit	Skope-Rules
2014	Kevin Durant		Kevin Durant
			LeBron James
2011	Derrick Rose	LeBron James	
2008	Kobe Bryant	LeBron James	LeBron James
2007	Dirk Nowitzki		Dirk
			Nowitzki
2006	Steve Nash	LeBron James	Dirk Nowitzki
			LeBron James
2005	Steve Nash	Kevin Garnett	LeBron James
			Kevin Garnett
2003	Tim Duncan	Tim Duncan	Tracy McGrady
		Kevin Garnett	Kevin Garnett
2001	Allen Iverson		Shaquille
			O'Neal
1999	Karl Malone	Shaquille O'Neal	Shaquille
			O'Neal
1998	Michael Jordan	Michael Jordan	
		Karl Malone	
1996	Michael Jordan		David Robinson
			Michael Jordan
1994	Hakeem Olajuwon	David Robinson	David
			Robinson
1993	Charles Barkley	Hakeem Olajuwon	Hakeem Olajuwon
			Michael Jordan
1990	Earvin Johnson	Michael Jordan	Michael
			Jordan
1989	Earvin Johnson	Michael Jordan	Michael
			Jordan
1987	Earvin Johnson	Earvin Johnson	Larry Bird
			Michael Jordan
1986	Larry Bird		Larry Bird
1985	Larry Bird		Larry Bird
1984	Larry Bird		
1983	Moses Malone	Moses Malone	
1982	Moses Malone	Moses Malone	
1981	Julius Erving		
1980	Kareem Abdul-Jabbar		Kareem
			Abdul-Jabbar

Note: Years in which the outputs of both diagnostic models exactly matched the historical MVP label are omitted. Listed names are same-season candidates covered by the diagnostic rules; blank cells indicate that no candidate is selected by the corresponding model.

Taken together, the top-3 and MVP-label analyses distinguish seasons with relatively strong

agreement from those in which the performance-based diagnostic profiles differ from the official voting-based outcome. The top-3 setting provides a broader test of whether the official winner remains compatible with the performance-oriented candidate profile, whereas the MVP-label setting applies a narrower definition based only on the UFER top-ranked player.

Figure 3 summarizes diagnostic coverage of the official MVP from 1980 to 2025 under the data-driven MVP and top-3 labeling schemes. Filled segments indicate seasons in which the corresponding diagnostic model covers the official winner.

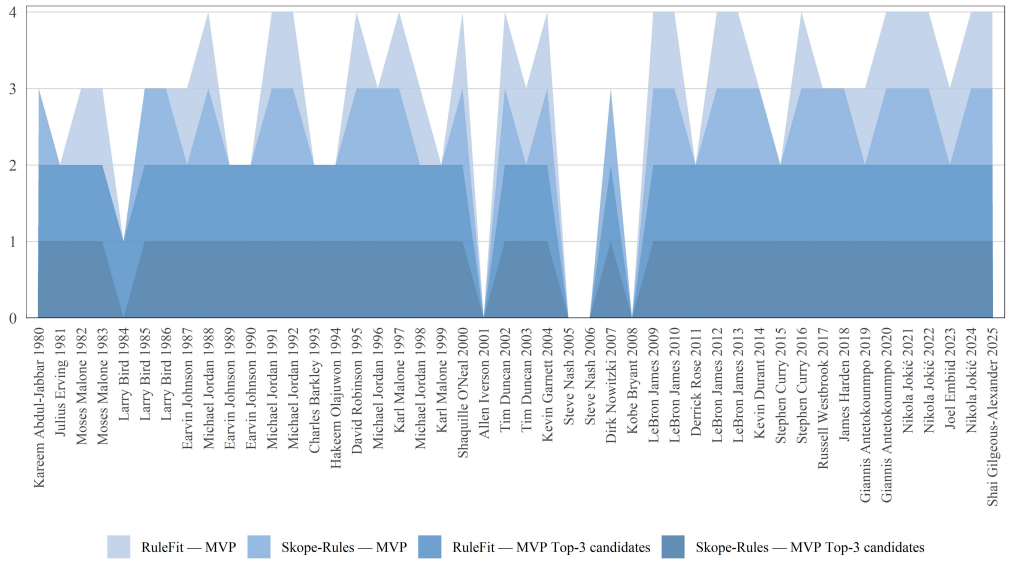

Figure 3. MVP diagnostic coverage by RuleFit and Skope-Rules from 1980 to 2025.

4.3.2. Comparison between data-driven and voting-based diagnostics

To examine how the target-label definition affects the learned diagnostic profiles, we apply the same RuleFit and Skope-Rules model classes under alternative labeling schemes using the same input feature representation. Under the data-driven setting, the target labels are derived from the UFER performance ranking, whereas under the historical setting, they represent the official voting-based outcomes. This matched design allows differences in the resulting diagnostic profiles to be examined primarily in relation to the alternative evaluation perspectives rather than differences in diagnostic model class or input representation.

Table 7 shows the profiles learned when the diagnostic models are trained directly on historical voting-based outcomes. In several seasons, the resulting rules recover the official winner together with other players exhibiting similar statistical profiles. For example, the historical-label diagnostics associate Nikola Jokić with Joel Embiid in 2024 and Michael Jordan with Karl Malone in 1998, illustrating that the learned rules characterize broader statistical profiles associated with historical voting-based outcomes rather than simply reproducing a single player name.

Table 7. Diagnostic coverage under historical MVP labels.

Year	Historical MVP	RuleFit	Skope-Rules
2024	Nikola Jokić	Nikola Jokić	Nikola Jokić
		Joel Embiid	Joel Embiid
2023	Joel Embiid		
2022	Nikola Jokić	Nikola Jokić	Nikola Jokić
		Antetokounmpo	Antetokounmpo
2021	Nikola Jokić	Nikola Jokić	Nikola Jokić
			Joel Embiid
2018	James Harden	James Harden	James Harden
			Stephen Curry
2017	Russell Westbrook		
2016	Stephen Curry	Stephen Curry	Stephen Curry
			Kevin Durant
2013	LeBron James	LeBron James	LeBron James
		Kevin Durant	Kevin Durant
2011	Derrick Rose		Derrick Rose
			LeBron James
2008	Kobe Bryant		
2006	Steve Nash	Dirk Nowitzki	Dirk
			Nowitzki
2005	Steve Nash		Dirk
			Nowitzki
2003	Tim Duncan		Tim Duncan
2002	Tim Duncan		Tim Duncan
2001	Allen Iverson		
1999	Karl Malone		
1998	Michael Jordan	Michael Jordan	Michael Jordan
		Karl Malone	Karl Malone
1997	Karl Malone	Karl Malone	Karl Malone
		Michael Jordan	Michael Jordan
1996	Michael Jordan	Michael Jordan	Michael Jordan
			David Robinson
1994	Hakeem Olajuwon	Hakeem Olajuwon	David
		David Robinson	Robinson
1993	Charles Barkley	Charles Barkley	Charles Barkley
			Michael Jordan
1990	Earvin Johnson	Earvin Johnson	Earvin Johnson
		Michael Jordan	Michael Jordan
1989	Earvin Johnson	Earvin Johnson	Earvin Johnson
		Michael Jordan	Michael Jordan
1984	Larry Bird		
1982	Moses Malone		

Continued on next page

Year	Historical MVP	RuleFit	Skope-Rules
1981	Julius Erving		
1980	Kareem Abdul-Jabbar		

Note: Years in which the outputs of both diagnostic models exactly matched the historical MVP label are omitted. Blank cells indicate that no candidate was selected by the corresponding diagnostic model.

Comparing these results with the data-driven diagnostics in Table 6 reveals that the two labeling schemes can emphasize different player profiles even when the same diagnostic model classes are used. In 2015, for example, the data-driven diagnostic highlights Anthony Davis, whereas Stephen Curry received the official MVP award. In 1990, the data-driven diagnostics emphasize Michael Jordan, while the historical-label diagnostics include Earvin Johnson and therefore more closely reflect the official voting-based outcome. These differences indicate that performance-based and voting-based labels encode partly different evaluation patterns.

The comparison should not be interpreted as establishing which evaluation perspective is correct, nor does it identify the causes of particular historical voting outcomes. Rather, the voting-trained diagnostics describe statistical profiles associated with historical voting-based outcomes, whereas the data-driven diagnostics identify profiles associated with the UFER performance benchmark. The matched use of RuleFit and Skope-Rules therefore provides an interpretable way to examine how the two evaluation perspectives align and where they diverge.

4.3.3. Cumulative MVP-profile analysis

Using RuleFit, we estimate the probability that a player falls within an MVP-like profile in each season. Summing these probabilities over historically awarded MVP seasons yields the cumulative MVP score shown in Figure 4; the figure also reports cumulative top-3 profile scores and actual MVP counts.

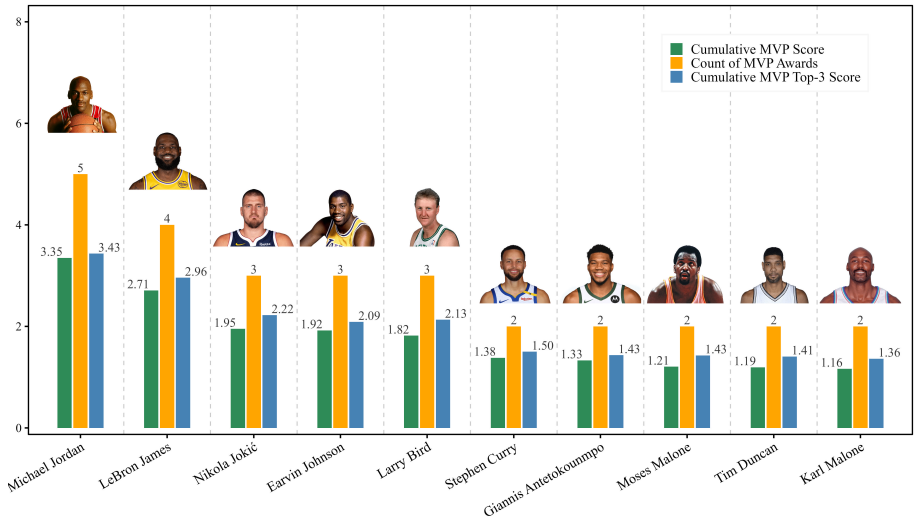

Figure 4. Top-10 players by cumulative RuleFit MVP-profile scores.

Across the 46 seasons under study, 27 players were awarded the regular-season MVP. Figure 4 should be read as a within-sample diagnostic summary, not as an all-time ranking. A high cumulative score means that a player repeatedly receives strong MVP-profile probabilities within the modeled seasons, helping distinguish players with the same award count. This is why players with equal numbers of MVP awards can still be separated: the score reflects how strongly their awarded seasons resemble the learned MVP profile, not only how many awards they won. Because the analysis is restricted to seasons ending from 1980 to 2025, awards outside this window are not included.

4.4. Additional validation analyses

Recent research on sensitivity analysis and validation in multi-criteria decision-making emphasizes the role of robustness, reduced subjectivity, and transparency in assessing decision procedures [29]. Following this validation perspective, the additional analyses in this section examine UFER from three complementary directions: sensitivity to alternative ensemble-weight configurations, robustness relative to independent weighting and aggregation methods, and agreement with an external voting-based reference. These analyses are reported separately so that internal ranking stability, cross-method robustness, and external agreement are not conflated.

4.4.1. Equal-weight rationale and sensitivity analysis

Following the definition in Section 3.3.2, Equal is evaluated as a neutral reference rather than assumed to be uniquely optimal. The sensitivity analysis examines whether comparable rankings can be obtained under alternative admissible weighting configurations.

Sensitivity was assessed over 66 nonnegative weight triples on a 0.1-step simplex grid, covering the interior, edges, and three single-method vertices of the feasible weight space. Equal was evaluated separately because one-third is not included on the 0.1 grid. For concise reporting, three grid points with weights $(0.6, 0.2, 0.2)$ and their permutations are highlighted as dominant configurations; these points belong to the 66-point grid and are not additional off-grid tests. The Spearman correlation and top-1 agreement measure overall and winner-level consistency with Equal, and top-3 Jaccard similarity and player-level rank shifts reveal whether apparently high global correlation conceals meaningful local movement near the top of the ranking. Because the historical data identify the MVP winner rather than a complete external ranking, NDCG@k, MVP Hit@k, and mean winner rank are used for outcome-based evaluation. This separation ensures that internal stability is not confused with agreement with historical outcomes.

The internal sensitivity results are summarized by Figure 5 and Table 8. Across the grid, the Spearman correlation with Equal ranges from 0.9671 to 0.9997 with a median of 0.9945, while top-3 Jaccard similarity ranges from 0.793 to 1.000 with a median of 0.946. The three highlighted 0.6-dominant settings retain Spearman correlations of 0.9940–0.9977 and top-1 agreement of 95.65–97.83%; their mean absolute rank shifts are only 0.124–0.319 positions, their 90th-percentile shifts are one position, and their maximum shifts are two to four positions. By contrast, the Genie3-only setting produces the largest change (mean: 0.983, 90th percentile: 3, maximum: 8) and the lowest top-3 Jaccard similarity (0.793). Thus, the ranking is broadly stable under moderate departures from Equal but becomes less stable when all weight is assigned to a single component.

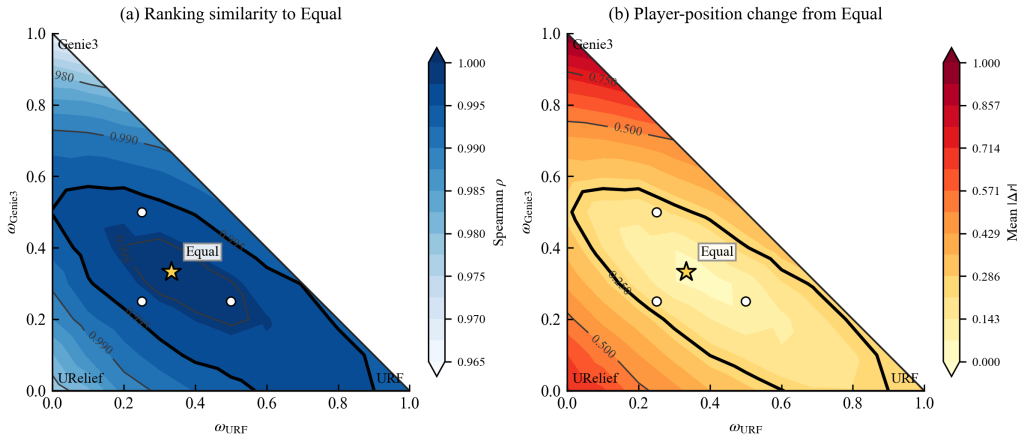

Figure 5. Weight sensitivity across the simplex. The star denotes the separately evaluated Equal setting; circles denote the three 0.6-dominant grid points. Panel (a) shows the Spearman correlation with Equal; panel (b) shows mean absolute rank change. Bold contours mark $\rho = 0.995$ and mean $|\Delta r| = 0.25$.

Table 8. Internal ranking sensitivity relative to Equal under representative weights.

Config.	Weights	ρ	Top-1	Mean $\lvert\Delta r\rvert$	P90	Max	Top-3 J	Top-3 in
Equal	(0.333, 0.333, 0.333)	1.0000	100.00%	0.000	0	0	1.000	0.000
URF only	(1, 0, 0)	0.9942	97.83%	0.289	1	3	0.935	0.130
Genie3 only	(0, 1, 0)	0.9671	89.13%	0.983	3	8	0.793	0.413
URelief only	(0, 0, 1)	0.9783	91.30%	0.724	2	6	0.870	0.261
URF dominant	(0.60, 0.20, 0.20)	0.9977	97.83%	0.124	1	2	0.967	0.065
Genie3 dominant	(0.20, 0.60, 0.20)	0.9940	95.65%	0.296	1	3	0.902	0.196
URelief dominant	(0.20, 0.20, 0.60)	0.9943	97.83%	0.319	1	4	0.957	0.087

Note: All metrics use Equal as the internal reference. ρ and Top-1 measure overall and winner-level agreement; Mean, P90, and Max $|\Delta r|$ summarize absolute player-rank changes; Top-3 J is Top-3 Jaccard similarity, and Top-3 in is the mean number of new top-3 entrants per season. Official MVP outcomes are not used. Weights follow (URF, Genie3, URelief).

Table 9 provides the separate outcome-based agreement check. Under these external measures, Equal attains MVP Hit@1, Hit@3, and Hit@5 values of 69.57%, 91.30%, and 95.65%, respectively, and is tied for the best Hit@1 and Hit@5 across all 67 evaluated settings. Genie3 only yields the weakest NDCG@3 and Hit@1. Equal also remains within 0.0057 of the best normalized discounted cumulative gain (NDCG) value and differs by only 0.04 positions from the best mean MVP rank. Therefore, Equal does not dominate every individual metric, but it combines strong internal stability with near-best external agreement, supporting its use as a stable, interpretable, and method-neutral default rather than a uniquely optimal specification.

Table 9. External agreement with the official MVP under representative weights.

Config.	Weights	NDCG@3	Hit@1	Hit@3	Hit@5	MVP rank
Equal	(0.333, 0.333, 0.333)	0.8129	69.57%	91.30%	95.65%	1.87
URF only	(1, 0, 0)	0.8186	67.39%	93.48%	95.65%	1.83
Genie3 only	(0, 1, 0)	0.7316	58.70%	84.78%	91.30%	2.35
URelief only	(0, 0, 1)	0.8077	69.57%	89.13%	93.48%	1.83
URF dominant	(0.60, 0.20, 0.20)	0.8077	67.39%	91.30%	95.65%	1.87
Genie3 dominant	(0.20, 0.60, 0.20)	0.7723	65.22%	86.96%	95.65%	2.02
URelief dominant	(0.20, 0.20, 0.60)	0.8157	69.57%	91.30%	95.65%	1.85

Note: All metrics use the official MVP as an external voting-based reference. Hit@k is the percentage of seasons in which the official MVP appears within the UFER top-k; NDCG@3 additionally discounts the winner's position within the top 3, and MVP rank is the official winner's mean UFER position. Official outcomes are not used in UFER estimation or coefficient selection. Weights follow (URF, Genie3, URelief).

4.4.2. External weighting and consensus-ranking baselines

To provide broader external validation of the proposed ranking module, UFER is compared with seven weighting and rank-aggregation baselines under the same perturbation protocols. The comparison includes (1) objective criterion-weighting methods, including entropy, CRITIC, MEREC, and LOPCOW [24–26]; (2) projection-based and random-weighting references, including PCA and Random-Z; and (3) a consensus-ranking baseline, Borda Count. All methods use the same analytical sample and criterion-direction specification: TOV, TOV%, and PF are treated as cost criteria, while the remaining 40 variables are treated as benefit criteria. Method-specific normalization is used only for estimating criterion weights. Entropy, CRITIC, and LOPCOW use benefit/cost min–max normalization, MEREC uses its own benefit/cost normalization, and UFER and PCA operate on the direction-aligned standardized feature matrix.

For UFER, entropy, CRITIC, MEREC, LOPCOW, and Random-Z, the resulting feature-weight vector is applied to the common direction-aligned standardized scoring matrix $\mathbf{Z}$. For weighting method q, the player-season score is

$$S_i^{(q)} = \sum_{j=1}^{p} z_{ij} w_j^{(q)},$$

where $w_j^{(q)}$ denotes the weight assigned to feature j by method q. This controlled design holds the downstream scoring rule constant while allowing the feature-weight estimation mechanism to vary.

Only the LOPCOW criterion-weighting component of the LOPCOW–DOBI framework is used, so that the final scoring rule remains comparable with the other weighting methods. PCA ranks players by the first principal-component score after sign alignment, and Random-Z samples nonnegative random weights that sum to one under each paired perturbation. Borda Count is included separately as a consensus-ranking baseline and therefore does not use the feature-weight dot-product scoring rule. For each season s, URF, Genie3, and URelief first generate independent player rankings. If n_s players are evaluated in season s and $r_{m,s}(i)$ denotes the rank assigned to player i by component method m, the Borda score is defined as

$$B_s(i) = \sum_{m \in \{\text{URF,Genie3,URelief}\}} (n_s - r_{m,s}(i) + 1).$$

Players are ranked in descending order of $B_s(i)$. Borda Count therefore performs aggregation at the final ranking level, whereas UFER integrates the three component methods at the feature-importance level before player scoring. All baselines are evaluated under the same feature-perturbation, statistical-group-deletion, and correlated irrelevant-feature-injection protocols used in Section 4.2.1. Weight-based methods are re-estimated after each perturbation, and Borda Count recomputes the component rankings before aggregation.

Table 10 reports the mean ranking-preservation scores under the three perturbation regimes.

Table 10. Metric-level robustness comparison with external weighting and consensus-ranking baselines.

Method	Top-3 Agreement	NDCG@3	Spearman
Feature perturbation			
UFER	0.8354	0.8869	0.9489
PCA	0.8076	0.8499	0.9374
Entropy	0.7196	0.7627	0.8923
CRITIC	0.8088	0.8705	0.9424
MEREC	0.7846	0.8441	0.9179
LOPCOW	**0.8460**	**0.8985**	**0.9512**
Borda Count	0.8303	0.8820	0.9486
Random-Z	0.6515	0.6981	0.8273
Statistical-group deletion			
UFER	**0.8954**	**0.9332**	0.9672
PCA	0.8385	0.8739	0.9476
Entropy	0.8137	0.8500	0.9355
CRITIC	0.8644	0.9192	0.9645
MEREC	0.8644	0.9090	0.9573
LOPCOW	0.8892	0.9328	**0.9683**
Borda Count	0.8861	0.9246	0.9669
Random-Z	0.6863	0.7452	0.8466
Correlated irrelevant-feature injection			
UFER	0.4666	0.4826	0.6063
PCA	0.4353	0.4345	0.3609
Entropy	**0.6469**	**0.6936**	**0.8224**
CRITIC	0.4834	0.5079	0.6390
MEREC	0.3306	0.3319	0.4146
LOPCOW	0.4134	0.4254	0.5321
Borda Count	0.4725	0.4911	0.6135
Random-Z	0.4230	0.4346	0.5548

Note: Values are mean robustness scores averaged over seasons and perturbation settings within each regime. Higher values indicate stronger ranking preservation. Bold values indicate the best result for each metric within each perturbation regime.

The expanded comparison reveals distinct robustness profiles across the competing methods. Under random feature perturbation, LOPCOW achieves the strongest ranking preservation, while

UFER remains highly competitive. Under statistical-group deletion, UFER provides the strongest preservation of the leading candidate set and its ordering, although LOPCOW achieves a slightly higher overall rank-order consistency. A different pattern emerges under correlated irrelevant-feature injection, where entropy shows the strongest preservation, indicating that different weighting mechanisms exhibit different sensitivities to perturbation types. Overall, no method consistently dominates all perturbation regimes, while UFER maintains competitive performance across all settings, particularly when redundancy among statistical feature groups is explicitly considered.

Table 11 summarizes the corresponding average robustness ranks.

Table 11. Average robustness ranks of UFER and external comparison methods.

Method	Feature perturbation (Rank)	Group deletion (Rank)	Irrelevant-feature injection (Rank)
UFER	2.000 (2)	**2.881 (1)**	4.000 (3)
PCA	4.917 (5)	4.381 (5)	5.667 (4)
Entropy	7.000 (7)	5.952 (7)	**1.333 (1)**
CRITIC	4.167 (4)	4.476 (6)	3.000 (2)
MEREC	5.917 (6)	3.286 (2)	7.333 (7)
LOPCOW	**1.000 (1)**	3.810 (4)	5.778 (5)
Borda Count	3.000 (3)	3.548 (3)	3.000 (2)
Random-Z	8.000 (8)	7.667 (8)	5.889 (6)

Note: All reported values are average robustness ranks calculated from top-3 agreement, NDCG@3, and Spearman correlation. Values in parentheses indicate the column-wise dense rank of each method, with lower ranks indicating stronger robustness. Top-1 agreement is not included in the rank calculation.

The rank-based comparison in Table 11 confirms the regime-specific patterns observed above. UFER achieves the strongest overall robustness rank under statistical-group deletion and remains among the top-performing methods under feature perturbation. Although its relative position decreases under correlated irrelevant-feature injection, taken together, the three regime-specific comparisons indicate that UFER maintains a comparatively balanced robustness profile across the evaluated perturbation settings. These results further indicate that robustness depends on the perturbation mechanism, and no single weighting or aggregation strategy is universally optimal across all scenarios.

4.4.3. Robustness to feature dependence and redundancy

The original player-statistics matrix contains conceptually and statistically related performance indicators. Because correlated variables may provide partially substitutable representations of similar information, they can affect the allocation and interpretation of feature importance [30, 31]. We therefore use a correlation-guided ablation design to examine whether feature dependence changes importance estimates and whether such changes propagate to the final player rankings.

The analysis proceeds in three steps. First, correlated feature groups are identified using pairwise absolute Spearman correlations and complete-linkage hierarchical clustering. The primary threshold is $|\rho| \geq 0.85$, with $|\rho| \geq 0.80$ and $|\rho| \geq 0.90$ used as sensitivity checks. Second, in the leave-one-correlated-feature-out (LOCO) experiment, one feature from a correlated group is removed at a time while the remaining features are retained, after which URF, Genie3, URelief, and UFER weights are

re-estimated from the perturbed feature matrix. Third, in the retain-one experiment, one member of a correlated group is retained while the other members are removed, and the retained member is rotated across all members of the group. This provides a stronger redundancy-reduction test than LOCO for groups with three or more features.

For both ablation designs, URF, Genie3, URelief, UFER weighting, player scoring, and season-specific ranking are recomputed under each perturbed configuration. Tree-based components are evaluated with 20 paired random seeds, using the same seed for the full and corresponding perturbed model.

Feature-importance stability is measured by comparing the full-feature and perturbed importance rankings over the common remaining features. For each method

$$m \in \{\text{URF}, \text{Genie3}, \text{URelief}, \text{UFER}\},$$

the stability score is defined as

$$S_{m,j}^{\text{imp}} = \rho_S \left(\mathbf{w}_{0,-j}^{m}, \mathbf{w}_{-j}^{m} \right), \tag{4.1}$$

where $\mathbf{w}_{0,-j}^{m}$ denotes the importance vector from the full-feature model after excluding the removed feature x_j from the comparison, and $\mathbf{w}_{-j}^{m}$ denotes the importance vector re-estimated after x_j is removed from the input data. The correlation is therefore calculated over all features that are present in both configurations. A value closer to one indicates that the relative importance ordering of the remaining features is less affected by the removal of the correlated predictor.

The rankings produced under LOCO and retain-one perturbations are compared with the corresponding full-feature ranking using the top-3 agreement, NDCG@3, and Spearman rank correlation defined in Section 4.2.1. These metrics respectively characterize preservation of the principal candidate set, ordering near the top of the ranking, and overall rank-order consistency.

To provide a direct measure of positional change, we additionally calculate the mean absolute rank shift (MARS),

$$\text{MARS}_{q,s}^{b} = \frac{1}{N_s} \sum_{i=1}^{N_s} \left| r_{q,s}^{b}(i) - r_{q,s}^{0}(i) \right|, \tag{4.2}$$

where N_s denotes the number of evaluated players in season s, $r_{q,s}^{0}(i)$ is the reference rank of player i, and $r_{q,s}^{b}(i)$ is the corresponding rank under the correlated-feature perturbation. Lower MARS values indicate smaller absolute positional changes.

Accordingly, the LOCO experiment is used primarily to assess feature-importance redistribution after the removal of a correlated predictor, whereas the retain-one experiment evaluates whether reducing repeated representations within strongly correlated groups materially changes the resulting player rankings.

We first examine the extent of feature interdependence in the original player-statistics matrix. At the primary threshold of $|\rho| \geq 0.85$, complete-linkage clustering identifies 13 non-singleton correlated groups containing 33 of the 43 performance features. The threshold-sensitivity results are summarized in Table 12.

Table 12. Sensitivity of the identified correlated-feature structure to the Spearman threshold.

$\lvert\rho\rvert$ threshold	Number of correlated groups	Features in correlated groups
0.80	13	34
0.85	13	33
0.90	13	28

The number of identified correlated groups remains 13 across all three thresholds, while the number of features assigned to these groups decreases from 34 to 28 as the criterion becomes more restrictive. This indicates that the broad dependence structure persists across reasonable changes in the operational threshold, although the membership of individual groups becomes more restrictive at higher correlation levels.

At the primary threshold, 13 non-singleton correlated groups are identified. Representative groups include AST–AST%, 3P–3PA–3PAr, DRB–TRB–DRB%–TRB%, FG–FGA–PTS–USG%, PER–OBPM–BPM, and OWS–WS–VORP. Figure 6 summarizes the full Spearman dependence structure of the 43 performance features, with the 13 correlated groups arranged as contiguous blocks and singleton features shown separately. The threshold-sensitivity results in Table 12, together with the LOCO and retain-one results reported below, are used to evaluate whether this feature dependence materially affects feature-importance allocation and the resulting player rankings.

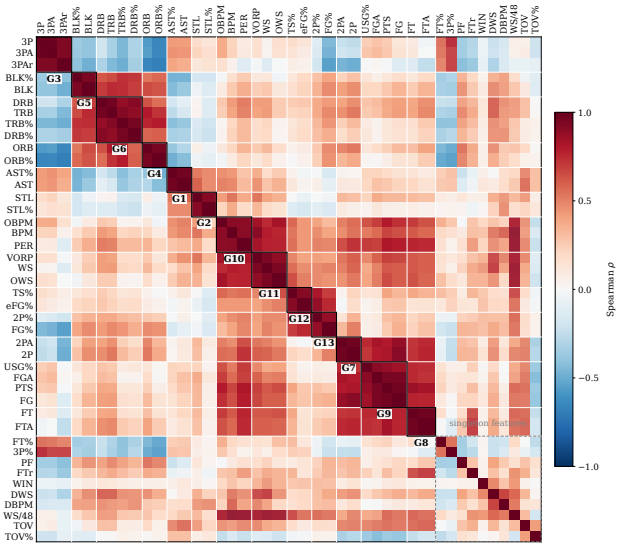

Figure 6. Spearman correlation structure of the 43 performance features. The 13 non-singleton correlated groups identified by complete-linkage clustering at $\lvert\rho\rvert \geq 0.85$ are arranged as contiguous blocks (G1–G13), while singleton features are shown separately. The color range represents the pairwise Spearman correlation coefficient.

We next examine whether this feature dependence affects importance estimation and the resulting player rankings. Table 13 reports feature-importance stability under the 33 LOCO configurations.

The response to correlated-feature removal differs substantially across the component estimators. URelief exhibits very high stability, with a mean importance Spearman correlation of 0.996. In contrast, Genie3 and URF obtain mean correlations of 0.744 and 0.652, respectively, indicating

considerably greater changes in the relative importance ordering after one member of a correlated group is removed. The effect is also present in the integrated UFER weights. The mean UFER importance Spearman correlation under LOCO is 0.706. Thus, equal-weight integration does not preserve an invariant feature-importance ordering when correlated predictors are removed.

The more important question for the proposed ranking framework is whether these feature-weight changes materially propagate to the final player ordering. Table 14 reports the corresponding UFER ranking-stability results.

Table 13. Feature-importance stability under leave-one-correlated-feature-out perturbations.

Method	Mean Spearman	SD	Range
URF	0.652	0.071	0.448–0.831
Genie3	0.744	0.058	0.522–0.895
URelief	**0.996**	0.002	0.992–0.999
UFER	0.706	0.064	0.529–0.870

Note: Statistics summarize 33 correlated-feature removals evaluated with 20 paired random seeds, yielding 660 LOCO re-estimations per method. The Spearman correlation compares the full-model and perturbed feature-importance orderings over features available in both configurations. Higher values indicate greater importance stability.

Table 14. UFER ranking stability under correlated-feature perturbations.

Experiment	Top-3 Agreement	NDCG@3	Spearman	MARS
LOCO	0.965	0.982	0.994	0.272
Retain-one	0.944	0.970	0.989	0.430

Note: Values are averaged over correlated-feature perturbation configurations, 46 seasons, and 20 paired random seeds. Higher top-3 agreement, NDCG@3, and Spearman values indicate greater ranking preservation; lower MARS values indicate smaller absolute positional changes.

Although correlated-feature removal changes the feature-weight allocation, the induced player rankings remain highly stable. Under LOCO, UFER maintains a top-3 agreement of 0.965, an NDCG@3 of 0.982, and a Spearman rank correlation of 0.994, with a mean absolute rank shift of 0.272. The retain-one redundancy-reduction experiment produces larger changes overall because multi-member correlated groups are reduced to a single retained representation, while two-member groups coincide with the corresponding one-feature-removal setting. Even under this setting, the corresponding values remain 0.944, 0.970, and 0.989, with a MARS of 0.430. These results indicate that the principal ranking structure is substantially less sensitive to correlated-feature representation than the underlying feature-weight vector.

5. Discussion

5.1. Implications of data-driven benchmarking for subjective evaluation

The results demonstrate that data-driven ranking and subjective recognition systems address related but different evaluation objectives. UFER is not designed to predict historical MVP selections; instead, it constructs a performance-based benchmark from measurable indicators and provides an independent reference for examining how voting-based outcomes relate to statistical evidence. Therefore, the comparison between UFER and historical MVP results should be interpreted as an

analysis of different evaluation perspectives rather than a competition between supervised prediction and unsupervised ranking.

The empirical findings reveal two complementary aspects of this benchmark. At the feature level, different unsupervised relevance estimators provide partially overlapping information, and their ensemble integration reduces dependence on a single importance estimation strategy. At the ranking level, the robustness experiments show that changes in feature attribution do not necessarily translate into equivalent changes in player ordering. This distinction between feature attribution and ranking stability is important when interpreting multi-criteria evaluation models.

The role of UFER is therefore not to replace expert judgment or define a unique value standard, but to provide a transparent quantitative reference for auditing subjective decisions. By separating benchmark construction from historical voting outcomes, the framework enables researchers to examine when measurable performance aligns with subjective recognition and when the two evaluation perspectives diverge.

5.2. Interpreting divergence between the performance-based benchmark and voting-based outcomes

The comparison between UFER rankings and historical MVP selections indicates that performance-based evaluation and voting-based recognition are strongly related but not identical. The substantial agreement observed across many seasons suggests that measurable player performance provides an important basis for award decisions. At the same time, divergence cases indicate that voting-based recognition may incorporate evaluation dimensions beyond the statistical indicators represented in the current benchmark.

The interpretable diagnostic analysis further clarifies this difference. Rules derived from the UFER-based benchmark describe statistical patterns associated with measurable player contribution, whereas rules learned from historical MVP labels characterize profiles associated with actual award outcomes. Therefore, the diagnostic comparison does not determine which evaluation system is correct; instead, it reveals that the two systems may emphasize different aspects of player value. The top-k external agreement analysis and the stricter diagnostic comparison serve different purposes: the former identifies seasons with larger winner-ranking discrepancies (e.g., 2001, 2005, 2006, and 2008), whereas the latter examines cases where the learned evaluation patterns differ between performance-oriented and voting-based labels.

Examples such as 2015, 1990, and 1994 illustrate this distinction under the top-1 diagnostic setting. These cases should be interpreted as differences between two evaluation structures rather than direct evidence of incorrect voting decisions. Additional contextual considerations may contribute to subjective recognition, but such factors are not explicitly modeled or quantitatively tested in the present framework. Therefore, the diagnostic results identify where measurable performance and voting outcomes differ, while avoiding causal interpretations of voter behavior.

5.3. Robustness, limitations, and broader implications

The robustness analysis highlights an important issue in feature-weighted ranking systems: correlated criteria may affect the allocation of feature importance without necessarily compromising the stability of final decisions. In multi-indicator evaluation problems, individual feature weights should therefore be interpreted as relative contributions within the observed feature representation

rather than isolated causal effects of individual variables.

The LOCO, retain-one, and perturbation analyses consistently show that feature attribution and ranking stability represent different levels of robustness. Correlated variables can redistribute importance among related indicators, but the resulting player rankings remain substantially more stable. This finding suggests that evaluating only the invariance of individual weights may provide an incomplete assessment of feature-weighted ranking frameworks. For decision-oriented applications, stability of the final ranking structure can be a more relevant criterion when assessing the reliability of the resulting evaluation.

The ensemble design of UFER contributes to this robustness by integrating multiple unsupervised relevance perspectives rather than relying on a single estimator. However, UFER should not be considered completely invariant to feature dependence. The resulting benchmark remains conditional on the selected indicators, normalization choices, and modeling assumptions. The contribution of the framework is therefore not to eliminate all uncertainty in feature weighting, but to provide a transparent and empirically validated ranking procedure under the specified evaluation criteria.

Several limitations should also be acknowledged. First, although the selected indicators capture major measurable aspects of player contribution, some contextual information involved in real-world award decisions is not explicitly represented. Second, UFER provides a performance-based benchmark rather than an absolute definition of player value; therefore, divergence between UFER rankings and MVP outcomes should not be interpreted as causal evidence explaining voting behavior. Third, the current validation is based on NBA MVP evaluation, and extensions to other subjective ranking scenarios require domain-specific indicators and independent validation.

The conclusions are also conditional on the available statistics and feature representation. Changes in data coverage, measurement conventions, feature selection, direction, or normalization may influence the resulting benchmark. The framework should therefore be viewed as a transparent and reproducible reference under specified criteria rather than a value-free ground truth.

These limitations also define the broader applicability of the proposed framework. Rather than providing a universal value standard, UFER offers an interpretable auditing perspective for examining subjective ranking decisions under specified criteria. Future extensions to other domains may benefit from richer decision information, broader indicator systems, and independent validation to further evaluate the generalizability of this framework.

6. Conclusions

This study proposed an interpretable unsupervised feature-weighted ensemble ranking framework for multi-criteria player evaluation. By integrating unsupervised random forest, Genie3, and URelief, UFER constructs a common performance benchmark from player statistics without relying on historical MVP voting outcomes during model construction. RuleFit and Skope-Rules are then used as post-ranking diagnostic tools to compare data-driven MVP profiles with historical voting-based outcomes.

The empirical analysis of NBA regular-season data from 1980 to 2025 shows that UFER captures a stable performance structure while also identifying seasons in which the performance-based benchmark diverges from official MVP outcomes. Historically debated seasons such as 2001, 2005, 2006, and 2008 exhibit relatively large discrepancies between statistical performance signals and

voting-based outcomes, whereas several high-consensus seasons show closer agreement. These deviations should be interpreted as diagnostic differences between a performance-based statistical benchmark and voting-based outcomes rather than as direct evidence that either outcome is intrinsically correct or incorrect.

The additional validation analyses further clarify the scope of the framework. External comparisons show that different weighting and consensus-ranking methods have distinct robustness strengths across perturbation regimes, with no single method dominating every setting. Feature-dependence analysis also shows that correlated variables can affect the allocation of feature importance, but the resulting UFER rankings remain substantially more stable than the underlying feature-weight representation. These findings support the use of UFER as a competitive and interpretable benchmark, while also indicating that its feature weights should be read within the selected multivariate feature representation.

The proposed framework should therefore not be interpreted as a replacement for human judgment. Instead, it provides a transparent and reproducible benchmark for examining when subjective ranking outcomes align with or depart from measurable performance evidence. Although the empirical analysis focuses on NBA MVP evaluation, the underlying framework can be extended to other sports awards and broader multi-criteria decision settings in which quantitative indicators and subjective judgments coexist. These findings should be interpreted in light of the historical data, selected performance variables, and modeling choices considered in this study, with broader applicability requiring further domain-specific validation.

Author contributions

Zheng Fang and Wei Liang: Conceptualization, Methodology, Supervision, and Writing–review and editing. Xinyue Weng and Dongding Zhang: Software, Data curation, Formal analysis, Validation, Visualization, and Writing–original draft. Junjie Zhou: Investigation and Resources. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

During the preparation of this work, the authors used generative AI tools only to improve sentence-level language clarity and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments

This work was supported by the Aesthetic and General Education Curriculum Program of Xiamen University (U10303800138). The authors thank Haoran Qiu, Wangming Qin, Yihao Zhu, Hongxiang Fang, and Si Zhang for their assistance in collecting relevant literature.

Data availability statement

The data used in this study were obtained from the public databases described in the Data and Feature Description section. The replication materials are available from the corresponding author

upon reasonable request.

Conflict of interest

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

References

1. B. J. Coleman, J. M. DuMond, A. K. Lynch, An examination of NBA MVP voting behavior: Does race matter? *J. Sports Econ.*, **9** (2008), 606–627. https://doi.org/10.1177/1527002508320653
2. Y. Zhai, T. Xu, Novel metric to predict NBA regular season MVP, In: *2024 10th IEEE International Conference on High Performance and Smart Computing (HPSC)*, New York: IEEE, 2024, 36–42. https://doi.org/10.1109/HPSC62738.2024.00014
3. C. Cheng, Moral hazard in teams with subjective evaluations, *RAND J. Econ.*, **52** (2021), 22–48. https://doi.org/10.1111/1756-2171.12360
4. S. Ishiguro, Y. Yasuda, Moral hazard and subjective evaluation, *J. Econ. Theory*, **209** (2023), 105619. https://doi.org/10.1016/j.jet.2023.105619
5. V. S. Maas, R. Torres-Gonzalez, Subjective performance evaluation and gender discrimination, *J. Bus. Ethics*, **101** (2011), 667–681. https://doi.org/10.1007/s10551-011-0763-7
6. S. Jacob, F. Willits, Objective and subjective indicators of community evaluation: A Pennsylvania assessment, *Soc. Indic. Res.*, **32** (1994), 161–177. https://doi.org/10.1007/BF01078733
7. B. Anderson, R. Ryan, W. Goudy, Consistency in subjective evaluations of community attributes, *Soc. Indic. Res.*, **14** (1984), 165–175. https://doi.org/10.1007/BF00293408
8. A. I. Chatzimouratidis, P. A. Pilavachi, Objective and subjective evaluation of power plants and their non-radioactive emissions using the analytic hierarchy process, *Energy Policy*, **35** (2007), 4027–4038. https://doi.org/10.1016/j.enpol.2007.02.003
9. A. Kuhn, In the eye of the beholder: Subjective inequality measures and individuals' assessment of market justice, *Eur. J. Polit. Econ.*, **27** (2011), 625–641. https://doi.org/10.1016/j.ejpoleco.2011.06.002
10. J. E. Rockoff, C. Speroni, Subjective and objective evaluations of teacher effectiveness: Evidence from New York City, *Labour Econ.*, **18** (2011), 687–696. https://doi.org/10.1016/j.labeco.2011.02.004
11. C. A. Parsons, J. Sulaeman, M. C. Yates, D. S. Hamermesh, Strike three: Discrimination, incentives, and evaluation, *Am. Econ. Rev.*, **101** (2011), 1410–1435. https://doi.org/10.1257/aer.101.4.1410
12. H. Choi, S. Kim, School ties and evaluation outcomes: Evidence from the Korean basketball league, *Can. J. Econ.*, **57** (2024), 1337–1359. https://doi.org/10.1111/caje.12746
13. V. Scoppa, Are subjective evaluations biased by social factors or connections? An econometric analysis of soccer referee decisions, *Empir. Econ.*, **35** (2008), 123–140. https://doi.org/10.1007/s00181-007-0146-1

14. J. Lee, Outlier aversion in subjective evaluation: Evidence from World Figure Skating Championships, *J. Sports Econ.*, **9** (2008), 141–159. https://doi.org/10.1177/1527002507299203

15. J. Hollinger, *Pro Basketball Forecast*, Washington: Potomac Books, 2005.

16. J. Sill, Improved NBA adjusted ± using regularization and out-of-sample testing, In: *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*, Cambridge, 2010.

17. N. Nguyen, B. Ma, J. Hu, Predicting National Basketball Association players performance and popularity: A data mining approach, In: *Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Lecture Notes in Computer Science*, Cham: Springer, **12496** (2020), 293–304. https://doi.org/10.1007/978-3-030-63007-2_23

18. Y. Ke, R. Bian, R. Chandra, A unified machine learning framework for basketball team roster construction: NBA and WNBA, *Appl. Soft Comput.*, **153** (2024), 111298. https://doi.org/10.1016/j.asoc.2024.111298

19. X. Sun, J. Davis, O. Schulte, G. Liu, Cracking the black box: distilling deep sports analytics, In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, 3154–3162. https://doi.org/10.1145/3394486.3403367

20. Y. Wang, W. Liu, X. Liu, Explainable AI techniques with application to NBA gameplay prediction, *Neurocomputing*, **483** (2022), 59–71. https://doi.org/10.1016/j.neucom.2022.01.098

21. M. Petkovic, D. Kocev, B. Skrlj, S. Dzeroski, Ensemble- and distance-based feature ranking for unsupervised learning, *Int. J. Intell. Syst.*, **36** (2021), 3068–3086. https://doi.org/10.1002/int.22390

22. A. Sarkar, S. S. Goswami, D. K. Behera, D. Bozanic, Criteria weighting methods in multi-criteria decision making: A comprehensive review of subjective, objective, and hybrid approaches, *J. Contemp. Decis. Sci.*, **3** (2026), 1–34. https://doi.org/10.67334/cds202619

23. A. Sarkar, S. S. Goswami, A comprehensive review: The novel weighting methods for multi-criteria decision-making (MCDM), *Spec. Oper. Res.*, 2026, 1–26. https://doi.org/10.31181/sor202781

24. D. Diakoulaki, G. Mavrotas, L. Papayannakis, Determining objective weights in multiple criteria problems: The CRITIC method, *Comput. Oper. Res.*, **22** (1995), 763–770. https://doi.org/10.1016/0305-0548(94)00059-H

25. M. Keshavarz-Ghorabaee, M. Amiri, E. K. Zavadskas, Z. Turskis, J. Antucheviciene, Determination of objective weights using a new method based on the removal effects of criteria (MEREC), *Symmetry*, **13** (2021), 525. https://doi.org/10.3390/sym13040525

26. F. Ecer, D. Pamucar, A novel LOPCOW-DOBI multi-criteria sustainability performance assessment methodology: An application in developing country banking sector, *Omega*, **112** (2022), 102690. https://doi.org/10.1016/j.omega.2022.102690

27. P. Emerson, The original Borda count and partial voting, *Soc. Choice Welfare*, **40** (2013), 353–358. https://doi.org/10.1007/s00355-011-0603-9

28. J. H. Friedman, B. E. Popescu, Predictive learning via rule ensembles, *Ann. Appl. Stat.*, **2** (2008), 916–954. https://doi.org/10.1214/07-AOAS148

29. A. Sarkar, S. S. Goswami, K. K. Gupta, Sensitivity analysis and validation in MCDM methods: A comprehensive review with advancements, applications, and future directions, *Spec. Decis. Mak. Appl.*, **4** (2026), 1–14. https://doi.org/10.31181/sdmap41202769

30. C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinform.*, **8** (2007), 25. https://doi.org/10.1186/1471-2105-8-25

31. B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests, *Stat. Comput.*, **27** (2017), 659–678. https://doi.org/10.1007/s11222-016-9646-1

AIMS Press