ORIGINAL ARTICLE

Temporal machine learning-based prediction of handball player performance using a composite player performance score: An expanding-window Bundesliga study

Marouane Baadi . *LM2I. Faculty of Science and Technology. Sultan Moulay Slimane University. Beni Mellal, Morocco.*
Zakaria Khoudi. *LM2I. Faculty of Science and Technology. Sultan Moulay Slimane University. Beni Mellal, Morocco.*
Lekbir Afraites. *LM2I. Faculty of Science and Technology. Sultan Moulay Slimane University. Beni Mellal, Morocco.*
Soufiane Lyaqini. *LAMDA Laboratory. National School of Applied Sciences. Sultan Moulay Slimane University. Khouribga, Morocco.*

ABSTRACT

Predicting next-season handball performance requires temporally valid evaluation that accounts for roster turnover and information leakage. This study developed a Composite Player Performance Score (CPPS) using per-game offensive and disciplinary indicators and evaluated its predictability. We analysed 5,825 records from seven German Bundesliga seasons (2017–2018 to 2023–2024). Goalkeepers were excluded because available indicators inadequately represented their performance. Positive indicators were standardized using preceding-season reference distributions. Five expanding-window out-of-time evaluations compared Ridge Regression, Gradient Boosting, and Random Forest against a historical player-mean baseline. Logistic Regression predicted four prospective performance tiers. Ridge achieved the lowest mean absolute error (MAE = 0.3651) and highest coefficient of determination (R^2 = .4280), outperforming Gradient Boosting (MAE = 0.3732; R^2 = .4072), Random Forest (MAE = 0.3776; R^2 = .3948), and the historical baseline (MAE = 0.4239). This represented a 13.9% MAE reduction relative to baseline. For tier classification, Logistic Regression achieved accuracy of .4663 and Macro-F1 of .4425, compared with .3987 and .3958 for the baseline. These findings demonstrate moderate predictability among returning outfield players, although CPPS requires independent validation and results cannot be generalized to departing players or new entrants.
Keywords: performance analysis, handball, sports analytics, composite player performance score, temporal validation, ridge regression, player retention.

Cite this article as:
Baadi, M., Khoudi, Z., Afraites, L., & Lyaqini, S. (2027). Temporal machine learning-based prediction of handball player performance using a composite player performance score: An expanding-window Bundesliga study. *Journal of Human Sport and Exercise, 22*(1), 232-248. https://doi.org/10.55860/r89b3454

Corresponding author. *LM2I, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco.*
E-mail: marouane.baadi@usms.ma
Submitted for publication September 11, 2026.
Accepted for publication November 02, 2026.
Published November 24, 2026.
Journal of Human Sport and Exercise. ISSN 1988-5202.
©Asociación Española de Análisis del Rendimiento Deportivo. Alicante. Spain.
doi: https://doi.org/10.55860/r89b3454

© 2027 ARD Asociación Española

INTRODUCTION

Handball performance is multidimensional. Goals, assists, shooting efficiency, and ball recoveries describe some productive actions, whereas technical errors and disciplinary sanctions can reduce effective contribution. Nevertheless, no single season-level statistic captures all tactical, physical, and contextual determinants of player value. A composite index can offer a transparent operational summary, provided that its scope and limitations are explicit (Wagner et al., 2023).

Machine-learning applications in sport increasingly combine multiple performance indicators, but retrospective predictive accuracy does not necessarily translate into realistic prospective forecasting. When the same players appear across seasons, randomly partitioned observations can mix information from the future with the training period. Furthermore, strong historical persistence means that complex algorithms must be evaluated against simple player-level baselines rather than against weak population averages (Hastie et al., 2009).

The previous version of this study used a Composite Player Development Score (CPDS). Its season-total disciplinary penalties, participation weighting, and target-season standardization made the index difficult to interpret as a comparable next-season outcome. In response, we introduce the Composite Player Performance Score (CPPS), using per-game exposure adjustment and reference distributions estimated only from prior seasons. The term "performance" is deliberately preferred to "development": year-to-year performance prediction is not equivalent to estimating an athlete's developmental trajectory.

The objective is not to establish CPPS as a definitive measure of player quality. Instead, this study asks whether next-season CPPS can be predicted using the preceding season's observable performance, whether a more complex nonlinear model improves upon a regularized linear model, and whether classification into prospective CPPS tiers adds value beyond a historical player-mean rule.

The principal contributions are: (i) a revised, dimensionless CPPS based on per-game variables and historically fixed reference values; (ii) a five-season expanding-window evaluation with historical-only model fitting; (iii) comparisons with a player-specific historical baseline; (iv) an explicit retention and turnover audit; and (v) a permutation-importance analysis to support cautious interpretation.

Related work

Handball match-analysis research has established the importance of technical and tactical indicators but also emphasized their context dependence (Ferrari et al., 2019). Wagner et al. (2023) introduced the PlayerScore as a systematic observation instrument for individual handball performance, illustrating the value of incorporating both productive and negative actions. That work is distinct from the present study because CPPS is derived from season-level administrative statistics and has not undergone independent observation-based validation.

Bassek et al. (2023) used graph-based deep learning to study contextualized intensity in elite handball, demonstrating the value of richer spatial information unavailable in the present dataset. Felice and Ley (2025) examined machine learning for handball match prediction with estimated team strengths. These approaches address different prediction units and outcomes from next-season individual-player CPPS.

The statistical literature provides a rationale for the model comparisons. Ridge Regression stabilizes linear prediction under correlated covariates (Hoerl & Kennard, 1970); Random Forest aggregates randomized

trees (Breiman, 2001); and Gradient Boosting fits sequential additive learners (Friedman, 2001). Their relative effectiveness is empirical and must be established on untouched future seasons.

Explainability is useful for assessing the dependence of a fitted model on its inputs, but importance should not be confused with causal influence or construct validity. This distinction is especially relevant when the outcome is a composite function of statistics related to the predictors.

MATERIALS AND METHODS

Figure 1 summarizes the study workflow, from cohort preparation and historical CPPS construction to expanding-window training, five held-out test seasons, and robustness analyses.

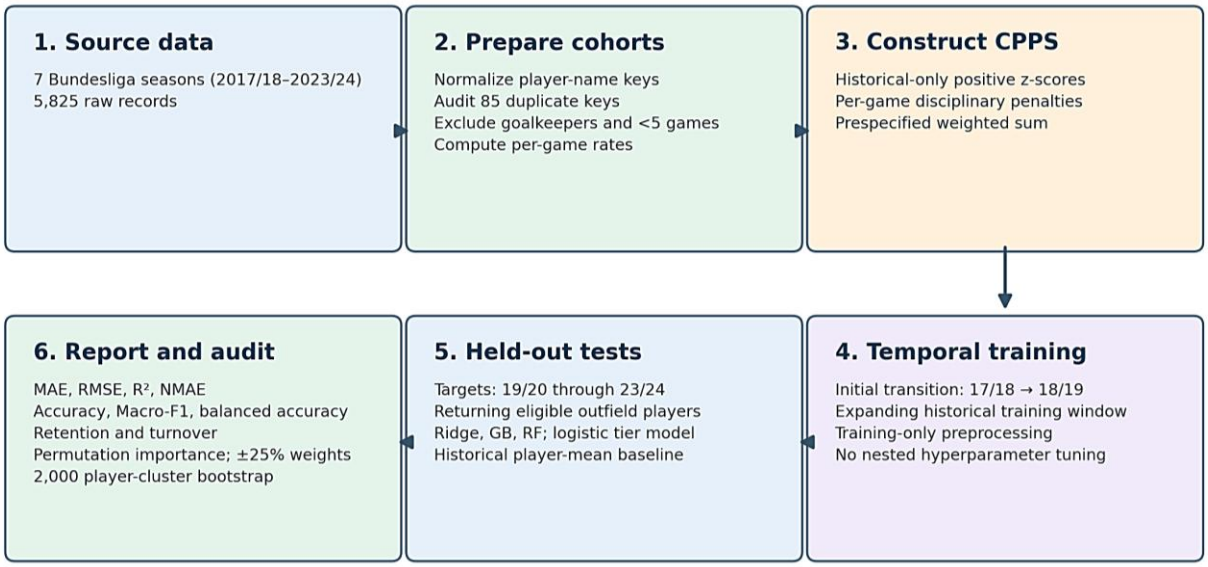

Figure 1. Methodological workflow of the historical-only CPPS forecasting and evaluation framework.

Study design and dataset

We used the publicly available German Bundesliga handball dataset attributed to Mansilla (2024), comprising 5,825 raw records from seven seasons (2017–2018 through 2023–2024). Variables include player name, team, position, games played, goals, missed shots, assists, steals, blocks, shooting accuracy, technical faults, yellow cards, two-minute exclusions, and red cards. No minutes played, age, injury history, opponent strength, or independent expert ratings were available.

The unit of modelling was a matched outfield player observed in both season $t-1$ and season t. Goalkeepers were not scored, since their primary contribution cannot be represented adequately by the available outfield offensive indicators. Players with fewer than five recorded games in a player-season were excluded from CPPS construction and modelling. The resulting inference is conditional on eligibility and consecutive-season observation.

Identity resolution and aggregation

Player keys were constructed from normalized concatenations of recorded surnames and given names. Where more than one row shared the same season–name key, counting statistics and games played were

 © 2027 ARD Asociación Española

summed. This operation may incorrectly combine namesakes or double-count participation where source records are ambiguous. The modal recorded position was retained. Shooting accuracy was recomputed as total goals divided by total recorded goals plus missed shots, rather than averaged without regard to attempt counts. When no attempts were recorded, accuracy was missing.

Table 1. Dataset variables, measurement definitions, and inclusion criteria.

Variable / group	Operational definition	Role
Player key	Normalized given name + surname; not independently verified	Matching
Season / team / position	Recorded season, team, and playing position	Cohort / categorical predictor
Games Played	Recorded games in season; minimum 5	Eligibility / predictor
Goals, Assists, Steals, Blocks	Season totals divided by games played	Per-game predictors
Shooting Accuracy	Goals / (goals + missed shots)	CPPS component / predictor
Technical Faults, Yellow Cards, 2 Min Exclusion, Red Cards	Season totals divided by games played	Penalty components / predictors
CPPS	Historically standardized positive rates minus weighted per-game penalties	Next-season target
Goalkeepers	Excluded from CPPS prediction; included in retention counts	Exclusion
Historical player mean	Training-history mean CPPS for each player; global training mean fallback	Comparator

The name-based approach can merge namesakes and cannot establish whether every apparent cross-season match represents the same individual. The code identified 85 duplicated season–name keys in the raw records. Some may reflect multiple team records or transfers; others require manual source verification. The study therefore reports this as a data-quality limitation rather than asserting that all identities were resolved.

Exposure-adjusted performance variables

For each count statistic X, the per-game rate was defined as $X_pg = X / G$, where G denotes recorded games played. Positive CPPS components were goals per game, assists per game, shooting accuracy, and steals per game. Penalty components were technical faults per game, yellow cards per game, two-minute exclusions per game, and red cards per game. Blocks per game and games played were retained as predictive variables but not directly weighted in CPPS.

The per-game adjustment reduces direct dependence on season length and participation counts; it does not substitute for minutes played. Players with equal games played may have very different court exposure. The five-game eligibility rule is an analytic choice, not a validated reliability threshold.

CPPS formulation and historical normalization

For each positive indicator k and target season s, the historical reference mean $\mu(k,<s)$ and standard deviation $\sigma(k,<s)$ were calculated from eligible outfield player-season observations in seasons strictly earlier than s. The standardized positive indicator was $z(i,s,k) = [x(i,s,k) - \mu(k,<s)] / \sigma(k,<s)$, with a fallback standard deviation of 1 if the reference dispersion was zero. Missing positive values were replaced by the corresponding historical reference mean, producing $z = 0$.

The score was defined as CPPS(i,s) = 0.35 z(Goals_pg) + 0.25 z(Assists_pg) + 0.25 z(Shooting Accuracy) + 0.15 z(Steals_pg) − 0.10 TechnicalFaults_pg − 0.15 YellowCards_pg − 0.10 TwoMinuteExclusions_pg − 0.30 RedCards_pg. This is the implemented additive specification. The positive weights sum to 1; negative coefficients were prespecified rather than learned from an external quality criterion. The score is dimensionless and may be negative. A one-unit change is not a clinically or practically validated unit of player value.

For season 2017–2018, no earlier reference season was available, so no CPPS target was calculated. CPPS for 2018–2019 was referenced to 2017–2018; later seasons used all preceding seasons. Crucially, reference parameters for each season did not use observations from that same or a future season. The resulting CPPS values may nevertheless shift with the evolving historical reference distribution; this property should be considered when comparing absolute scores over time.

Longitudinal cohorts and retention

Consecutive-season player sets were compared to describe retained players (present in both years), lost players (present only in the first year), and new players (present only in the second year). These descriptive cohorts include all eligible positions and are not identical to the outfield-only prediction cohorts. The recorded retained counts across the six transitions were 500, 503, 514, 545, 530, and 506. The associated retention rates ranged from 68.47% to 78.59%.

Performance modelling used matched eligible outfield players only. This conditional design introduces potential survivorship or selection bias: players who exit the observed competition may differ systematically in age, fitness, playing time, or ability. The current experiments do not estimate their unobserved next-season performance.

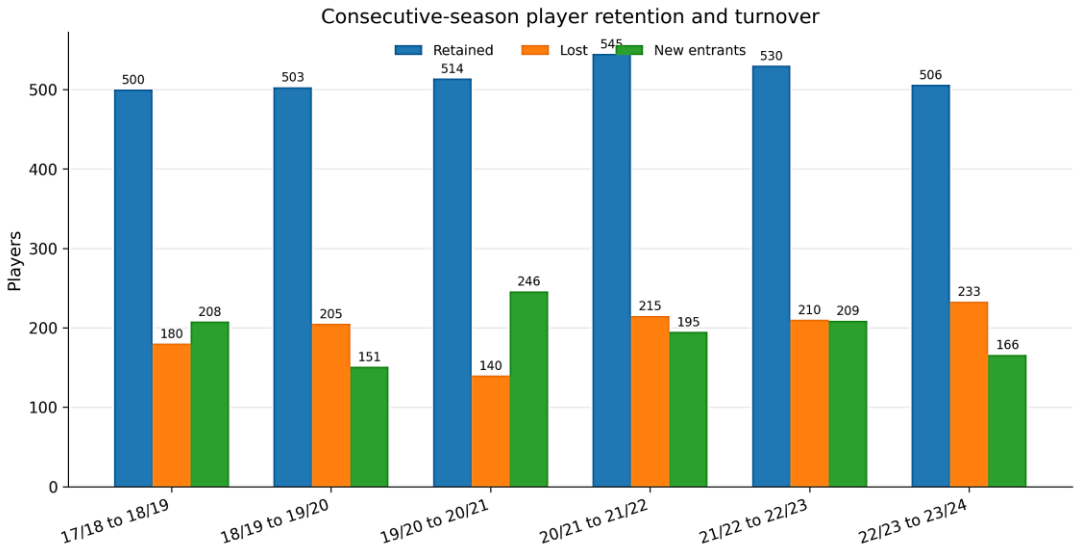

Figure 2. Consecutive-season player retention, loss, and entry.

Predictors and preprocessing

Predictors were obtained from the preceding season and included goals per game, assists per game, shooting accuracy, steals per game, technical faults per game, yellow cards per game, two-minute exclusions per game, red cards per game, blocks per game, games played, and playing position.

 © 2027 ARD Asociación Española

Numerical predictors were median-imputed using parameters estimated exclusively from the chronological training data. Ridge Regression and Logistic Regression additionally used training-only feature standardization. Playing position was represented using one-hot encoding fitted on the training observations, with previously unseen categories ignored during transformation. All preprocessing transformations were re-estimated within each expanding-window training period to prevent information leakage.

Expanding-window forecasting

Six adjacent-season transitions were available. The 2017–2018 to 2018–2019 transition provided the initial historical training observations. Formal out-of-time tests were conducted on target seasons 2019–2020, 2020–2021, 2021–2022, 2022–2023, and 2023–2024. For each test season, the model was fitted using only transitions with earlier target seasons and evaluated on the next unseen transition.

The analysis used Ridge Regression ($\alpha = 10$), Gradient Boosting Regression (120 estimators, learning rate 0.05, maximum depth 2), and Random Forest Regression (120 estimators, minimum leaf size 5). Randomized estimators used random_state = 42. These are prespecified configurations in the supplied executable script; this rerun did not perform a nested chronological hyperparameter search. Accordingly, the reported comparison should not be described as a tuned-versus-default experiment.

Historical baselines and classification

The continuous historical baseline predicted each player's target-season CPPS using the mean of that player's target CPPS values available in earlier training transitions. For players without such a history, it used the historical training-target population mean. The baseline was evaluated on the same future-season cohort as each regression model.

For four-class classification, quartile thresholds were calculated from CPPS targets in historical training transitions only and frozen before evaluating the target season. These cut points defined four ordered prospective classes. A Logistic Regression classifier (Bishop, 2006) was fitted to the preceding-season predictors and historical class labels. The historical player-mean classification baseline applied the same frozen thresholds to its continuous baseline predictions. Class thresholds were re-estimated only as the expanding training window advanced, not using future test targets.

Performance evaluation and interpretability

Regression was evaluated using mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and normalized MAE, defined here as MAE divided by the standard deviation of observed CPPS within the target test season. Classification was assessed using accuracy, Macro-F1, and balanced accuracy. Model summaries report arithmetic means of the five season-specific metrics; they are not necessarily equal to pooled-observation metrics.

Permutation importance was computed for Gradient Boosting on the held-out target-season data, with three feature permutations and the increase in MAE as the importance criterion. Because CPPS is explicitly built from related statistical components, high importance for goals or assists is expected and does not establish that these features independently validate the index. No causal interpretation is warranted.

The extended analysis includes ±25% weight perturbations, within-season quartile stability and a paired player-cluster bootstrap for regression MAE contrasts. These analyses assess internal robustness and uncertainty, not independent construct validity. Player identities remain unverified.

CPPS weight sensitivity and player-cluster bootstrap

We tested ±25% scaling of all positive weights jointly, all disciplinary penalty weights jointly, and each positive weight individually. For individual positive weights, the positive-weight vector was renormalized to its original total of one. Alternative CPPS values were compared with the original score using Spearman rank correlation and within-season quartile agreement, where quartiles were recomputed separately for each score specification. This is an internal index-stability analysis; it does not re-estimate forecasting models or validate CPPS against an external criterion. For uncertainty, 2,000 paired bootstrap samples (random seed 42) resampled normalized player-name clusters with replacement, retaining all their out-of-time test records. The estimate was the pooled mean difference in absolute error between each comparator and Ridge, with percentile 95% intervals. Clustering by names is imperfect when distinct players share a name. These intervals condition on the five observed seasons and do not capture uncertainty from sampling new seasons.

RESULTS

Player turnover and eligible cohorts

Across the six consecutive-season transitions, retained eligible players ranged from 500 to 545, with corresponding loss and entry in every year. Retention rates ranged from 68.47% to 78.59%. These values describe the season-level retention audit, including goalkeeper records; the prediction sample was restricted to matched outfield players.

The pronounced turnover confirms that a forecasting system trained only on consecutive-season matched players cannot be assumed to generalize to new entrants or to athletes who disappear from the competition. This distinction is important for interpreting the practical population to which the model applies.

Table 2. Player retention, attrition, and new entry by consecutive-season transition.

Transition	Retained	Lost	New	Retention
17/18 to 18/19	500	180	208	.7353
18/19 to 19/20	503	205	151	.7105
19/20 to 20/21	514	140	246	.7859
20/21 to 21/22	545	215	195	.7171
21/22 to 22/23	530	210	209	.7162
22/23 to 23/24	506	233	166	.6847

Note. Retention cohorts include eligible goalkeepers; prediction samples are outfield players only. Name-key identity matching is not independently verified.

Out-of-time regression performance

Ridge Regression achieved the lowest mean season-specific MAE (0.3651) and highest mean R^2 (.4280). Gradient Boosting achieved MAE = 0.3732 and R^2 = .4072, while Random Forest achieved MAE = 0.3776 and R^2 = .3948. The player-specific historical mean baseline achieved MAE = 0.4239.

Table 3. Five-season mean MAE, RMSE, R^2, and normalized MAE for regression models and the historical baseline.

Model	MAE	RMSE	R^2	NMAE
Gradient boosting	0.3732	0.4848	.4072	0.5924
Historical player mean	0.4239	0.5538	.2263	NR
Random forest	0.3776	0.4898	.3948	0.5995
Ridge	0.3651	0.4761	.4280	0.5797

Note. Unweighted mean of season-specific metrics. MAE and RMSE are dimensionless CPPS units; NMAE is MAE divided by target-season standard deviation. NR = not reported for the historical baseline in the supplied model summary.

 © 2027 ARD Asociación Española

The absolute difference in mean MAE between Ridge and the historical baseline was 0.0588 CPPS units, equivalent to a 13.9% reduction relative to the baseline mean. This is a descriptive comparison of five future seasons and should not be interpreted as proof of a causal effect or external validity. The mean season-level normalized MAE for Ridge was 0.5797.

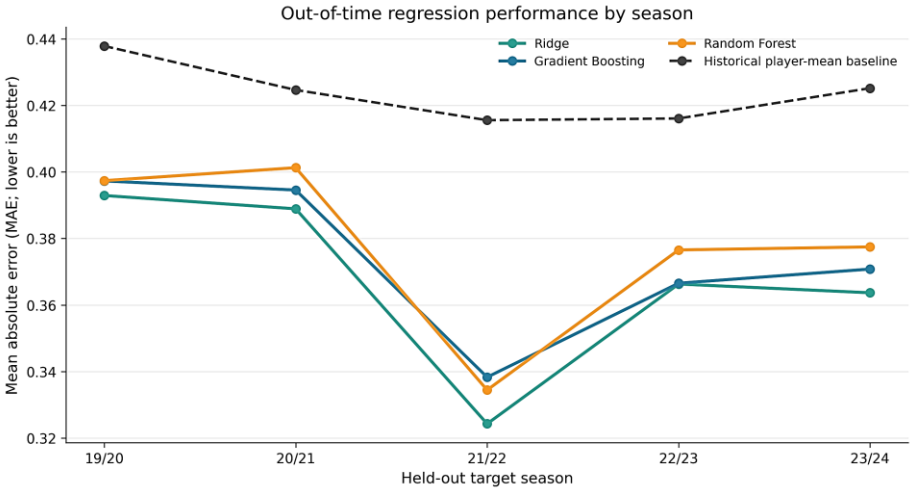

Figure 3. Season-specific MAE of regression models and the historical player-mean baseline.

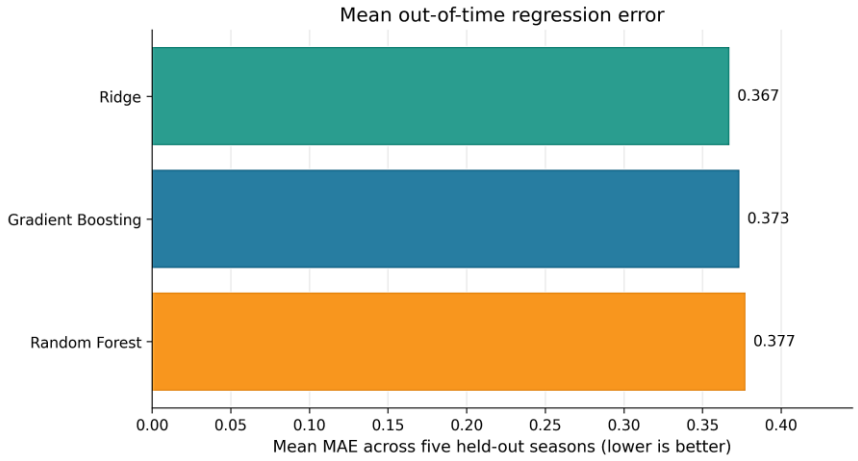

Figure 4. Average regression MAE comparison.

Variation across target seasons

Regression accuracy varied by target season. All three evaluated algorithms achieved their lowest MAE in 2021–2022, while their largest errors occurred in the earlier evaluation seasons. Ridge had the highest mean R^2 across the five target seasons, but the relative model ordering was not identical in every season. The pronounced variation underscores the value of reporting the full out-of-time sequence instead of a single randomly sampled holdout.

No nested tuning comparison was included in this CPPS rerun. The present results therefore supersede the previous CPDS numerical values and cannot be directly interpreted as an improvement in the same target scale.

Table 4. Season-specific out-of-time regression metrics and baseline errors.

Target season	Model	*n*	MAE	RMSE	R^2
19/20	Historical player mean	440	0.4379	0.5605	.1896
19/20	Ridge	440	0.3897	0.4932	.3725
19/20	Gradient boosting	440	0.3969	0.5032	.3470
19/20	Random forest	440	0.3961	0.5056	.3407
20/21	Historical player mean	443	0.4246	0.5586	.2783
20/21	Ridge	443	0.3837	0.5053	.4093
20/21	Gradient boosting	443	0.3942	0.5179	.3795
20/21	Random forest	443	0.4011	0.5215	.3709
21/22	Historical player mean	470	0.4156	0.5557	.2055
21/22	Ridge	470	0.3263	0.4378	.5070
21/22	Gradient boosting	470	0.3382	0.4497	.4797
21/22	Random forest	470	0.3365	0.4491	.4812
22/23	Historical player mean	458	0.4161	0.5306	.2455
22/23	Ridge	458	0.3635	0.4650	.4205
22/23	Gradient boosting	458	0.3659	0.4644	.4220
22/23	Random forest	458	0.3763	0.4752	.3948
23/24	Historical player mean	439	0.4252	0.5635	.2127
23/24	Ridge	439	0.3623	0.4791	.4307
23/24	Gradient boosting	439	0.3707	0.4887	.4077
23/24	Random forest	439	0.3783	0.4974	.3864

Note. All models are evaluated on the same held-out outfield-player cohort within each season.

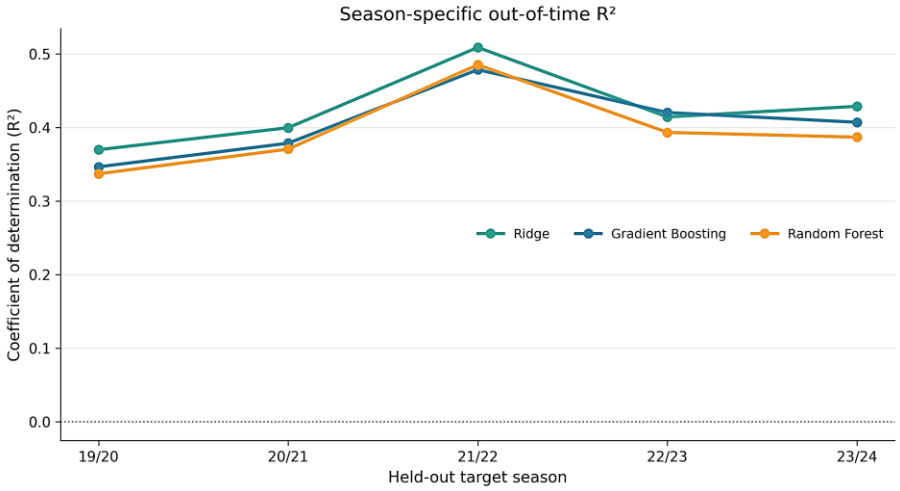

Figure 5. Season-specific regression R^2.

Prospective tier classification

The four-class Logistic Regression model achieved mean accuracy = .4663, Macro-F1 = .4425, and balanced accuracy = .4593. The historical player-mean tier baseline achieved accuracy = .3987, Macro-F1 = .3958, and balanced accuracy = .3943.

Logistic Regression exceeded the baseline by 0.0467 Macro-F1 points on the mean season-level metric. Nevertheless, Macro-F1 remained below .45, demonstrating substantial difficulty in assigning future-season performance tiers. Prospective quartiles defined using historical training targets do not guarantee balanced classes in the subsequent season.

 © 2027 ARD Asociación Española

Table 5. Mean prospective four-tier classification performance.

Model	Accuracy	Macro-F1	Balanced accuracy
Historical player mean	.3987	.3958	.3943
Logistic Regression	.4663	.4425	.4593

Note. Tier cutoffs are derived from historical training targets only.

Table 6. Season-specific prospective classification.

Target season	Model	*n*	Accuracy	Macro-F1	Balanced accuracy
19/20	Logistic Regression	440	.4159	.3841	.3915
19/20	Historical player mean	440	.3727	.3690	.3698
20/21	Logistic Regression	443	.4673	.4260	.4541
20/21	Historical player mean	443	.4063	.4011	.3991
21/22	Logistic Regression	470	.4851	.4624	.4850
21/22	Historical player mean	470	.4170	.4166	.4128
22/23	Logistic Regression	458	.4803	.4651	.4800
22/23	Historical player mean	458	.4192	.4184	.4164
23/24	Logistic Regression	439	.4829	.4749	.4858
23/24	Historical player mean	439	.3781	.3739	.3736

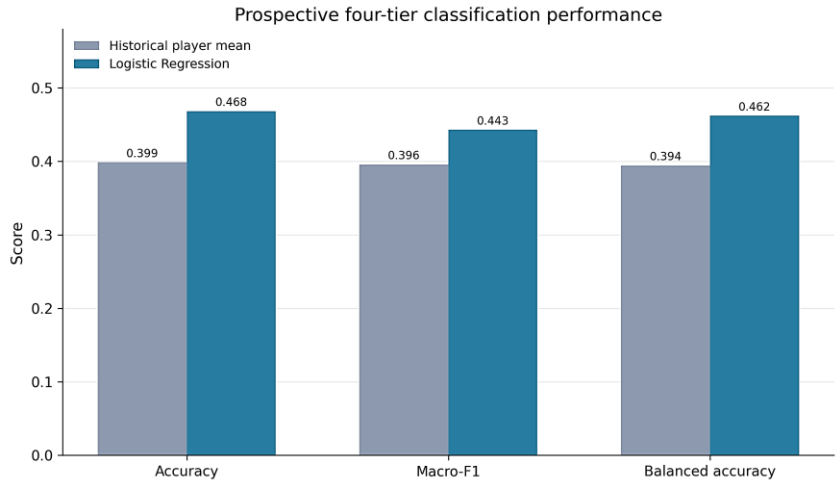

Figure 6. Accuracy, Macro-F1, and balanced accuracy for the two classification approaches.

Observed versus predicted CPPS

The pooled observed-versus-predicted diagram for Ridge displays a positive association but substantial residual dispersion. Predictions were compressed toward the centre of the score distribution: unusually high observed CPPS values were frequently underpredicted, while unusually low observed values were often overpredicted. This is consistent with regularization and with the limited ability of previous-season aggregate statistics to anticipate extreme subsequent-season outcomes.

The scatterplot should be read as a diagnostic visualization, not as an independent validation of the CPPS construct.

Permutation importance and score distribution

Permutation importance for Gradient Boosting identified preceding-season goals per game as the dominant predictor, followed by assists per game. Steals per game and shooting accuracy made smaller contributions.

Most disciplinary variables had comparatively low importance in the fitted forecasting model. These findings describe predictive reliance, not the real-world value of any handball action.

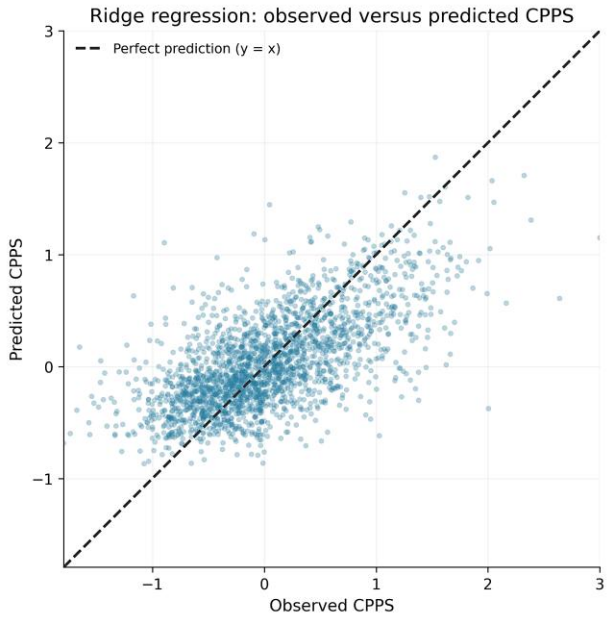

Figure 7. Observed versus predicted CPPS for Ridge across out-of-time test observations.

The out-of-time evaluation sample's CPPS distribution was approximately unimodal with a right tail and included both negative and positive values. Its numeric scale is dimensionless because the positive components were historically standardized. The score is not a probability, and zero does not mean an absence of performance.

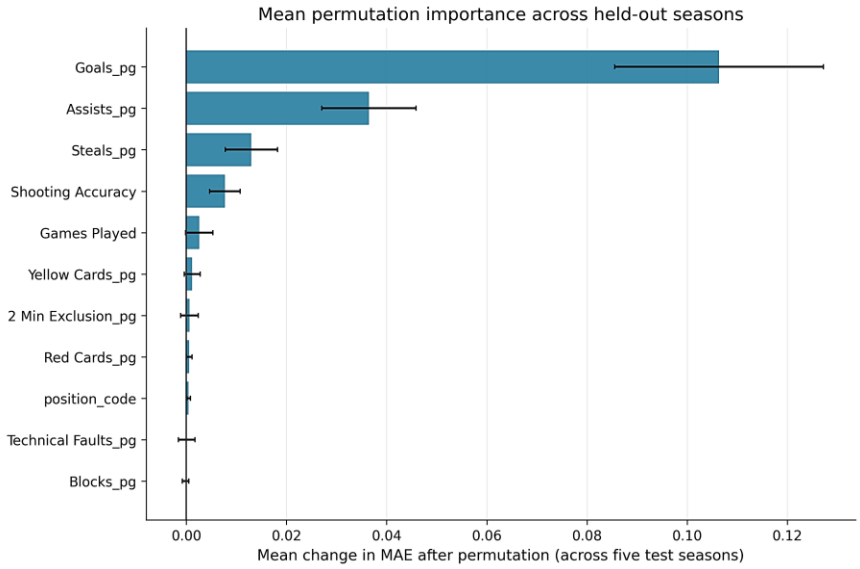

Figure 8. Held-out permutation importance for Gradient Boosting, measured as increase in MAE.

 © 2027 ARD Asociación Española

Supplementary Table S1. Mean held-out permutation importance for Gradient Boosting.

Predictor	Mean MAE increase
Goals_pg	0.1074
Assists_pg	0.0363
Steals_pg	0.0128
Shooting Accuracy	0.0072
Games Played	0.0026
Yellow Cards_pg	0.0014
Position	0.0011
2 Min Exclusion_pg	0.0008
Red Cards_pg	0.0005
Technical Faults_pg	0.0001
Blocks_pg	-0.0000

Note. Permutation importance is descriptive and does not establish causality or external score validity.

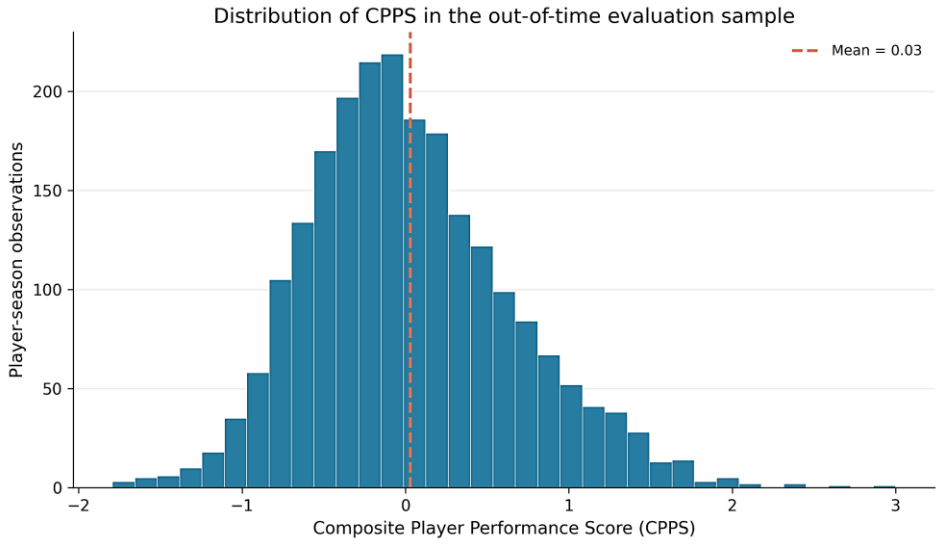

Figure 9. Distribution of CPPS in the out-of-time evaluation sample.

CPPS weight-sensitivity analysis

Across 12 prespecified ±25% coefficient perturbations, Spearman rank correlations between the original and modified CPPS ranged from .9945 to .9997. Within-season quartile agreement ranged from 91.04% to 98.10%.

The lowest quartile agreement (91.04%) occurred when the shooting-accuracy coefficient was reduced by 25%. Although player rankings were generally stable under the examined perturbations, up to approximately 9.0% of player-season observations changed quartile assignment.

These findings indicate internal robustness under the tested coefficient modifications but do not establish independent construct validity. Moreover, the within-season quartiles used in this sensitivity analysis differ from the historically determined thresholds used in prospective classification.

Detailed results are presented in Table 7.

Table 7. CPPS coefficient sensitivity.

Perturbation	Spearman ρ	Quartile agreement	Mean absolute CPPS change
original	1.0000	1.0000	0.0000
positive_weight_x0.75	.9992	.9723	0.1302
penalty_weight_x0.75	.9996	.9756	0.0254
positive_weight_x1.25	.9997	.9810	0.1302
penalty_weight_x1.25	.9995	.9788	0.0254
Goals_pg_weight_x0.75	.9974	.9603	0.0381
Goals_pg_weight_x1.25	.9984	.9544	0.0320
Assists_pg_weight_x0.75	.9970	.9430	0.0434
Assists_pg_weight_x1.25	.9977	.9451	0.0383
Shooting Accuracy_weight_x0.75	.9945	.9104	0.0524
Shooting Accuracy_weight_x1.25	.9954	.9381	0.0463
Steals_pg_weight_x0.75	.9988	.9631	0.0233
Steals_pg_weight_x1.25	.9990	.9739	0.0216

Paired player-cluster bootstrap uncertainty

Across pooled out-of-time predictions, the mean absolute-error difference between the historical player-mean baseline and Ridge Regression was 0.0591 CPPS units (95% player-cluster bootstrap confidence interval [CI: 0.0471, 0.0711]).

The corresponding differences were 0.0081 (95% CI [0.0037, 0.0126]) for Gradient Boosting minus Ridge and 0.0125 (95% CI [0.0072, 0.0177]) for Random Forest minus Ridge.

All intervals were calculated from 2,000 paired bootstrap resamples clustered by normalized player-name keys. These pooled differences should not be confused with differences between unweighted season-specific mean MAEs. The intervals do not account for uncertainty across future seasons, and name-based player identification remains a limitation.

Detailed comparisons are presented in Table 8.

Table 8. Paired player-cluster bootstrap differences in MAE.

Comparison	Pooled MAE difference	95% CI lower	95% CI upper	Resamples	Cluster
Historical player mean MAE minus Ridge MAE	0.0591	0.0471	0.0711	2,000	Normalized player-name key
Gradient Boosting MAE minus Ridge MAE	0.0081	0.0037	0.0126	2,000	Normalized player-name key
Random Forest MAE minus Ridge MAE	0.0125	0.0072	0.0177	2,000	Normalized player-name key

Note. 2,000 resamples clustered by normalized player-name key. Positive difference favours Ridge. Pooled paired difference is not the difference of unweighted season-mean MAEs.

DISCUSSION

The main finding of the CPPS rerun is that Ridge Regression provided the strongest average next-season regression performance among the three evaluated algorithms. The advantage of Ridge over Gradient Boosting and Random Forest is compatible with a forecasting signal dominated by stable, approximately

 © 2027 ARD Asociación Española

additive associations in season-level performance rates. It does not establish that nonlinear relationships are absent, because only a limited set of configurations was evaluated.

Compared with the historical player-mean baseline, Ridge reduced mean absolute error by approximately 13.9%. This indicates incremental predictive information in the preceding-season feature set, conditional on player retention and the chosen CPPS definition. Because historical target values and current predictors originate from overlapping types of performance statistics, the experiment evaluates predictability of a constructed outcome rather than independent player quality.

Classification performance was weaker in absolute terms. Although Logistic Regression exceeded the historical tier baseline, its mean Macro-F1 of .443 indicates considerable class confusion. Prospective thresholds derived from historical distributions also permit shifts in class prevalence. The classification experiment should therefore be presented as exploratory and not as a ready-to-deploy selection tool.

The importance results are substantively plausible but partially built into the target. Goals, assists, shooting accuracy, and steals enter the CPPS equation directly, so their influence on forecasting is not evidence that CPPS is valid as an external performance construct. Comparison against blinded coach assessments, independently coded match actions, or team-level impact measures is necessary before such a claim could be made.

The revised index addresses several methodological weaknesses of the original CPDS. Per-game penalties avoid direct inflation with season length, while historical-only standardization avoids using target-season statistics to set the score scale. However, historical reference parameters evolve over time, the coefficients remain normative, and games played is an imperfect exposure proxy. These choices should be examined through prespecified sensitivity analyses.

The turnover audit emphasizes selection bias. In real-world deployment, analysts would need a separate retention or availability model and a strategy for new entrants. Predictions conditional on a player appearing in the following season should not be described as unconditional career forecasts. A two-stage approach—first predicting retention, then conditional CPPS—would be a useful extension.

From a practical perspective, the most defensible use is a supplemental screening indicator for already-observed outfield players. Analysts should review predicted scores alongside position, playing time, tactical role, injury information, and qualitative match assessment. A model-generated score should not replace professional judgment.

Interpretive caution: CPPS is an operational index derived from selected offensive and disciplinary statistics, not an independently validated measure of overall handball performance. Its predictability does not establish criterion or construct validity. Independent assessment against blinded expert ratings, role-specific evaluations, playing-time-adjusted measures, or other external criteria is necessary before individual selection decisions can be supported.

Strengths, limitations, and future work

The study benefits from five genuinely prospective season-level test periods, historical-only model fitting, direct comparison with a player-specific baseline, transparent score construction, and explicit reporting of roster turnover. The Python workflow exports per-season predictions and model metrics for reproducibility.

Important limitations remain. First, the raw data lack stable unique player identifiers; 85 repeated season–name keys were flagged and require manual checking. Second, the data do not include minutes played, injuries, opponent quality, age, team tactics, or expert evaluations. Third, the CPPS coefficients were chosen a priori and have not been externally validated. Fourth, goalkeepers were excluded because the source lacks adequate goalkeeper-specific performance measures. Fifth, the model is evaluated only for players observed in consecutive seasons, introducing survivorship bias.

Sixth, position was represented using training-fitted one-hot encoding in the final rerun; however, role-specific modelling was not evaluated. Seventh, the current model set is narrower than the original CPDS experiment, and no chronological hyperparameter search was performed. Eighth, only five formal test seasons are available, limiting precision of season-level comparisons. Ninth, sensitivity to positive and disciplinary weights has been assessed, but eligibility-threshold sensitivity and predictive-model retraining under alternative score weights have not been evaluated.

Future work should include manually verified player identities; a position-aware score or separate role-specific subindices; independent external validation; season-block uncertainty assessments given the small number of test seasons; sensitivity to minimum games and model retraining under alternative CPPS definitions; nested temporal tuning; and two-stage modelling of player retention and conditional performance.

CONCLUSION

Using seven seasons of German Bundesliga data, we developed a historically referenced, exposure-adjusted Composite Player Performance Score and evaluated its next-season predictability under five expanding-window out-of-time tests. Ridge Regression achieved the strongest average continuous prediction performance (MAE = 0.3651; R^2 = .4280), improving on the historical player-mean baseline (MAE = 0.4239).

Logistic Regression exceeded the historical classification baseline but achieved only moderate tier-prediction performance (Macro-F1 = .4425).

These results demonstrate measurable predictability of the constructed CPPS outcome, not independent validation of player quality. The framework is most appropriately interpreted as a reproducible temporal benchmark for eligible returning outfield players, pending identity verification, further sensitivity analyses, and external validation.

AUTHOR CONTRIBUTIONS

All authors meet the criteria for authorship in accordance with established ethical guidelines. Contributions are specified according to the CRediT (Contributor Roles Taxonomy) as follows:

Conceptualisation: Marouane Baadi, Lekbir Afraites. Methodology: Marouane Baadi, Zakaria Khoudi, Lekbir Afraites, Soufiane Lyaqini. Formal analysis: Marouane Baadi. Investigation: Marouane Baadi. Data curation: Marouane Baadi, Zakaria Khoudi, Soufiane Lyaqini. Writing – original draft: Marouane Baadi. Writing – review & editing: Marouane Baadi, Zakaria Khoudi, Lekbir Afraites, Soufiane Lyaqini. Supervision: Lekbir Afraites, Soufiane Lyaqini.

All authors have critically reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

 © 2027 ARD Asociación Española

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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 manuscript.

ETHICS APPROVAL

This study was based exclusively on secondary analysis of publicly available observational sports-performance data. No experimental intervention, direct interaction with human participants, or collection of new personal data was undertaken. The research was conducted in accordance with applicable ethical principles and research regulations. No institutional ethics approval or exemption reference number is reported.

DATA AVAILABILITY

The data analysed in this study were obtained from the publicly accessible Bundes Handball Dataset—7 Seasons Until Half 2024, attributed to Mansilla (2024) and hosted on Kaggle. The dataset includes German Bundesliga handball player statistics covering seven seasons, from 2017–2018 to 2023–2024. The original dataset is accessible through the Kaggle platform, subject to the access and reuse conditions specified by the dataset provider.

The analyses were conducted using Python and the scikit-learn library. The reproducibility scripts are available from the corresponding author upon reasonable request.

AI USE DISCLOSURE

In accordance with current publishing ethics and transparency recommendations, generative artificial intelligence tools were used solely to assist with language editing, improving textual clarity, and manuscript formatting. No AI tools were used in the generation of scientific content, including the study design, data collection, analysis, interpretation of results, or the formulation of conclusions. The authors retain full responsibility for the content of the manuscript and confirm its originality, integrity, and accuracy.

REFERENCES

Bassek, M., Raabe, D., Banning, A., Memmert, D., & Rein, R. (2023). Analysis of contextualized intensity in men's elite handball using graph-based deep learning. *Journal of Sports Sciences, 41*(13), 1299–1308. https://doi.org/10.1080/02640414.2023.2268366

Bishop, C. M. (2006). Pattern *recognition* and *machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Felice, F., & Ley, C. (2025). Predicting handball matches with machine learning and statistically estimated team strengths. *Journal of Sports Analytics, 11*, 1–12. https://doi.org/10.1177/22150218251313937

Ferrari, W. R., Sarmento, H., & Vaz, V. (2019). Match analysis in handball: A systematic review. *Montenegrin Journal of Sports Science and Medicine, 8*(2), 63–76. https://doi.org/10.26773/mjssm.190909

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Mansilla, J. A. (2024). Bundes *handball dataset—7 seasons until half 2024* [Data set]. Kaggle.

Wagner, H., Hinz, M., Melcher, K., Radic, V., & Uhrmeister, J. (2023). The PlayerScore: A systematic game observation tool to determine individual player performance in team handball competition. *Applied Sciences, 13*(4), Article 2327. https://doi.org/10.3390/app13042327

This work is licensed under a Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0 DEED).

 © 2027 ARD Asociación Española