A Supervised Machine Learning Framework for Sports Championship Prediction: Case Studies in the NBA and FIFA World Cup

João Vitor T. Arroyo, Lucas V. Fujii, Gustavo de O. Gimenes, Murilo C. Naldi

Department of Computing – Universidade Federal de São Carlos (UFSCar)
São Carlos – SP – Brazil

{joaoarroyo, lucasfujii, ggimenes}@estudante.ufscar.br

naldi@ufscar.br

Abstract. *This paper proposes a generalizable machine learning framework for sports championship prediction, leveraging a structural analogy where qualifying phases provide features and championship phases yield targets. Evaluated on the NBA (2000–2025) and FIFA World Cup (1958–2022) across five classifiers, it addresses class imbalance via AUC-ROC optimization and threshold calibration. Logistic Regression achieved a 0.933 AUC-ROC in the NBA, while Random Forest and SVM led the World Cup with 0.921 AUC. Validation via DeLong and McNemar tests, bootstrap intervals, and temporal splits supports result stability. The framework readily adapts to other competitions, offering actionable insights for sports analytics.*

Keywords: *Sports Championship Prediction, Supervised Machine Learning, Class Imbalance, Temporal Validation, Feature Engineering, Sports Data Mining.*

1. Introduction

Predictive analytics in professional sports has expanded substantially over the past two decades, driven by detailed digital performance data and advances in machine learning (ML) [Terner and Franks 2021]. Beyond forecasting outcomes, sports data mining yields insights that inform coaching, scouting, and long-term roster planning [Horvat and Job 2020]. Predicting which team will reach or win a championship involves compounding challenges: severe class imbalance (only one team wins per season), temporal autocorrelation as teams evolve, high randomness in single-game knockouts, and few positive examples relative to all competitors [Loeffelholz et al. 2009]. This work is not aimed at betting or odds estimation; instead, it shows how supervised ML can convert qualifying-phase performance statistics into interpretable indicators for team analytics, scouting, and strategic planning.

Sport-specific prediction studies face two recurring issues: they usually focus on a single competition, limiting generalizability, and they often ignore temporal ordering when splitting data, causing data leakage and inflated performance [Horvat and Job 2020]. This paper addresses both by proposing a generalizable two-phase championship prediction

framework: the *qualifying phase* supplies per-team statistics as features, and the *championship phase* provides the prediction target. We evaluate the framework on two structurally different competitions—the NBA (American basketball, 82-game regular season) and the FIFA World Cup (global football, qualifying campaign + ELO)—whose contrasting structures, geographies, and data constraints test the framework's robustness.

A central methodological question in championship prediction is whether to use classification or ranking (see Section 2). Because of extreme class imbalance (about 3.4% positives), accuracy is misleading, so we use AUC-ROC as the main objective, with Precision–Recall curves, Average Precision, and systematic threshold optimization to ensure well-calibrated binary deployment decisions, thus linking ranking-based model selection to threshold-based deployment; Learning-to-Rank is left for future work. The paper also stresses methodological transparency, detailing temporal splitting, z-score normalization, confederation-level bias correction, ELO computation, and statistical significance testing (Section 3), to support replication and extension to other competitions such as the Brasileirão and the NFL/Super Bowl.

2. Related Work

2.1. Sport-Specific Prediction: Basketball and Football

Basketball prediction has progressed from neural networks for game outcomes to logistic regression, Naïve Bayes, and decision trees, where feature selection highlighted defensive over scoring statistics [Thabtah et al. 2019]. Reviews report that team-level metrics outperform player-level statistics for championship forecasting [Horvat and Job 2020], motivating our feature design: per-season z-scores to remove era-level biases from pace changes, a normalization used in both case studies [Terner and Franks 2021]. Recent ensemble methods (gradient boosting, random forests) improve NBA playoff prediction but often overfit small test sets due to the few championships per season [Bunker et al. 2024]. We extend this work by systematically comparing five classifiers with different inductive biases under temporal validation.

FIFA World Cup forecasting has been refined through regularized team-specific covariates and hybrid ELO–random forest approaches that yielded substantial performance gains for the 2018 tournament [Groll et al. 2019]. Among pre-tournament indicators, ELO ratings computed exclusively from neutral-ground matches are particularly informative [Gilch and Müller 2018], a finding that directly underpins our Case Study 2. More recent contributions integrate gradient boosting with ELO ratings [Bunker et al. 2024], employ explainable AI techniques [Song et al. 2024], and demonstrate that classical classifiers remain competitive for professional football leagues [Moya et al. 2025]. In contrast to Groll et al. [Groll et al. 2019], who primarily addressed match-level forecasting within a single World Cup, our framework is explicitly designed for pre-tournament championship prediction across multiple editions, leveraging aggregated qualifying-phase statistics rather than within-tournament temporal dynamics.

Reviews of over 100 sports ML studies identify three recurring flaws: randomly splitting temporal data, relying on accuracy alone under severe class imbalance, and omitting out-of-window validation [Horvat and Job 2020]—all of which our framework explicitly avoids. A later review (2019–2024) confirmed the dom-

inance of supervised and ensemble methods and the growing role of explainable AI [Felice et al. 2024]. Data leakage largely explains inflated accuracy in basketball, football, and rugby [Bunker and Susnjak 2022], while NFL and volleyball studies further support the cross-sport validity of the two-phase methodology [Lin et al. 2024].

When the positive class is under 4% of observations, accuracy is uninformative and the F1-score highly threshold-dependent. AUC-ROC measures the probability that a random positive is ranked above a random negative, making it threshold-agnostic and suitable for model comparison [Hand 2009]. Yet with extremely rare positives, AUC-ROC can overstate practical performance, as a model that ranks most positives relatively high may still miss the true champion at any realistic threshold [Davis and Goadrich 2006]. We therefore complement AUC-ROC with Precision–Recall curves, Average Precision, and structured threshold analysis that links ranking quality to binary deployment decisions. Learning-to-Rank methods, which directly optimize candidate ordering, are left for future work.

3. The Proposed Framework

The framework is based on the common two-phase structure of major sports competitions: a *qualifying phase* that yields per-team statistics over many matches, followed by a *championship phase* whose outcome (finalist or champion) is the prediction target. This pattern appears in regular season $\rightarrow$ playoffs (NBA); qualifiers/ELO/titles $\rightarrow$ knockout (FIFA World Cup); and, as future applications, league rounds $\rightarrow$ champion (Brasileirão), regular season $\rightarrow$ Super Bowl (NFL), and group stage $\rightarrow$ knockout (UEFA Champions League). We instantiate the framework in two structurally distinct domains: the NBA (data-rich, 82-game season) and the FIFA World Cup (sparse data, cross-confederation heterogeneity).

3.1. Framework Methodology

The framework defines six reusable steps implemented identically across both case studies, differing only in domain-specific feature definitions. **(1) Data collection** gathers team-level statistics from publicly available, versioned datasets. **(2) Target construction** defines two binary targets (`reached_final` and `champion`), both exhibiting severe class imbalance (3.4%–4.1%), which renders accuracy misleading. **(3) Feature engineering** computes per-team statistics and normalizes them *within each edition* using z-scores:

$$z_{i,t}^{(f)} = \frac{x_{i,t}^{(f)} - \mu_t^{(f)}}{\sigma_t^{(f)}}, \tag{1}$$

where i indexes teams, t editions, and f features. Per-edition normalization removes era-level confounders and makes features comparable across years [Terner and Franks 2021]. **(4) Temporal train/test split** partitions data chronologically, using `TimeSeriesSplit` for internal cross-validation [Horvat and Job 2020], supplemented by domain-specific robustness checks: Leave-One-Cup-Out (LOCO-CV) for the World Cup and walk-forward analysis for the NBA. **(5) Model training** evaluates five classifiers (Logistic Regression, Random Forest, XGBoost, SVM, and KNN) with class-weight balancing, training-only scaling, and hyperparameter tuning via AUC-ROC. Finally, **(6) Evaluation and forecasting** assesses models via AUC-ROC, Average Precision, and F1-score, alongside rigorous

statistical validation (DeLong [DeLong et al. 1988], McNemar [McNemar 1947], and bootstrap intervals, $B = 2000$). A threshold sweep identifies F1-maximizing cutoffs for calibration diagnostics, while prospective forecasting retrains models on the full historical dataset to rank candidates by probability.

4. Case Study 1: NBA (2000–2025)

4.1. Dataset and Preprocessing

The *NBA Team Summaries* dataset [Sumitro Datta 2024] contains regular-season statistics for all NBA franchises since 1946; we use seasons 2000–2025 (26 seasons, 775 team-seasons, 29–30 teams per season). Franchise relocations are standardized (e.g., *Seattle SuperSonics* $\rightarrow$ *OKC Thunder*). The two binary targets are `reached_finals` (52 positives, 6.7%) and `champion` (26 positives, 3.4%); the 96.7% naive test accuracy shows that accuracy is an unsuitable evaluation metric.

The final NBA feature set was derived from exploratory analysis on the full dataset (correlation screening, class separation). From 12 candidates, we compute per-season z-scores (Eq. 1) for NRtg and WinPct. Features with $r > 0.98$ (W/L%, SRS, NRtg, FT/FGA) are removed to avoid multicollinearity, leaving: {`ORtg`, `DRtg`, `eFG%`, `NRtg_zscore`, `WinPct_zscore`}.

TOV%, ORB%, and Pace show almost complete distributional overlap between finalists and non-finalists and are excluded. NRtg_zscore and WinPct_zscore show the strongest class separation. The five retained features are mostly non-redundant ($\max |r| < 0.92$ for all pairs except NRtg_zscore and WinPct_zscore, which are nearly perfectly correlated because both depend on win-loss records; both are kept because tree-based regularization exploits their complementary signals). Distribution plots for all 12 candidate features and the full correlation matrix (including the discarded collinear cluster) are available in the supplementary repository.[1]

4.2. Temporal Split and Training Protocol

The primary temporal split uses seasons 2000–2018 for training (19 seasons, 565 instances, 19 champions) and 2019–2025 for testing (7 seasons, 210 instances, 7 champions). Five classifiers are tuned with `GridSearchCV` and `TimeSeriesSplit(n_splits=5)`, optimizing AUC-ROC and using `class_weight=balanced`. Best models: Logistic Regression $C = 0.1$; Random Forest n_est$= 200$, max_depth$= 3$; XGBoost n_est$= 200$, depth$= 2$, lr$= 0.05$; SVM $C = 1$, kernel=linear; KNN $k = 21$, weights=uniform. The full grid (parameter ranges) is in the supplementary repository.[1]

4.3. Results

Table 1 reports the evaluation metrics for both prediction tasks. For finalist prediction, Logistic Regression attains the highest AUC (0.858), while Random Forest yields the best F1-score (0.286). For champion prediction, Logistic Regression reaches an AUC of 0.933 and Recall of 0.857, correctly identifying 6 of 7 champions, whereas XGBoost achieves

[1]Code, notebooks, and supplementary figures: `https://github.com/jvarroyo33/Artigo_ENIAC`

the highest F1-score (0.381) by trading recall for precision. SVM and KNN make no positive predictions at the default 0.5 threshold, despite AUCs above 0.82, motivating the threshold calibration.

Table 1. NBA results: finalist (test 2019–2025; 14 pos.) and champion (7 pos.) prediction. Bold = best per metric.

Task	Model	AUC	TP	FP	FN	TN	Recall	F1
Finalist	**Log. Regression**	**0.858**	11	55	3	141	0.786	0.275
	Random Forest	0.775	10	46	4	150	0.714	**0.286**
	XGBoost	0.775	10	52	4	144	0.714	0.263
	SVM	0.569	0	0	14	196	0.000	0.000
	KNN	0.839	0	0	14	196	0.000	0.000
Champion	**Log. Regression**	**0.933**	6	31	1	172	**0.857**	0.273
	Random Forest	0.916	3	18	4	185	0.429	0.214
	XGBoost	0.916	4	10	3	193	0.571	**0.381**
	SVM	0.828	0	0	7	203	0.000	0.000
	KNN	0.891	0	0	7	203	0.000	0.000

Table 2 presents Average Precision and DeLong-test AUC comparisons against the best model (Logistic Regression): LR's champion-prediction AUC is significantly higher than SVM's ($p = 0.046$) and KNN's ($p = 0.038$), but not significantly different from Random Forest's ($p = 0.376$) or XGBoost's ($p = 0.626$) at $\alpha = 0.05$. McNemar's test on overall correctness counter-intuitively favors SVM and KNN over LR and RF ($p < 0.01$; e.g., SVM vs LR: $b = 31$, $c = 6$, $p = 0.0001$), because under 3.4% prevalence, never predicting the positive class maximizes accuracy — illustrating why accuracy-based paired tests are ill-suited to imbalanced tasks and motivating our reliance on AUC-ROC, Average Precision, and ranking metrics for model selection.

Table 2. NBA champion prediction: Average Precision (AP), AUC, and DeLong p-values comparing each model to Logistic Regression (reference). Bootstrap 95% CI for AUC in parentheses.

Model	AUC	95% CI	AP	p (DeLong)
Log. Regression	0.933	[0.878, 0.975]	0.391	—
Random Forest	0.916	[0.865, 0.958]	0.197	0.376
XGBoost	0.916	[0.838, 0.973]	0.264	0.626
SVM	0.828	[0.689, 0.928]	0.118	0.046
KNN	0.891	[0.825, 0.950]	0.162	0.038

Ranking metrics (Top-1/Top-3 = share of seasons where the champion appears in the top-1/top-3 ranked teams; MRR = mean reciprocal rank) show XGBoost leading in both Top-3 (5/7) and MRR (0.540); LR, XGBoost, and KNN tie on Top-1 (2/7), but KNN (MRR 0.505) and LR (MRR 0.469) trail XGBoost, placing other champions lower when they rank the champion first. SVM trails in all three (1/7, 2/7, 0.319), consistent with its abstention pattern; Random Forest trails similarly (1/7, 3/7, 0.362). The full ranking table is in the supplementary repository.

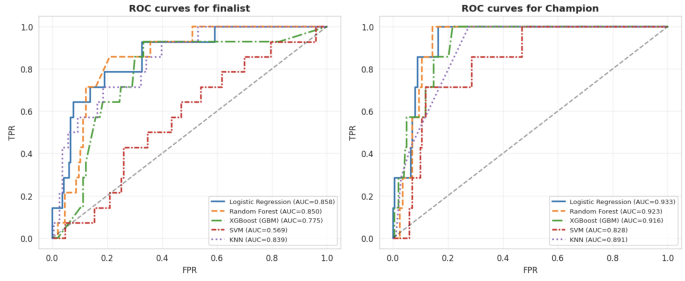


Figure 1. ROC curves for finalist (left) and champion (right) prediction, NBA test 2019–2025. LR achieves the highest AUC in both tasks. Confusion matrices (threshold=0.5) are summarized numerically in Table 1 and shown in full in the supplementary repository.

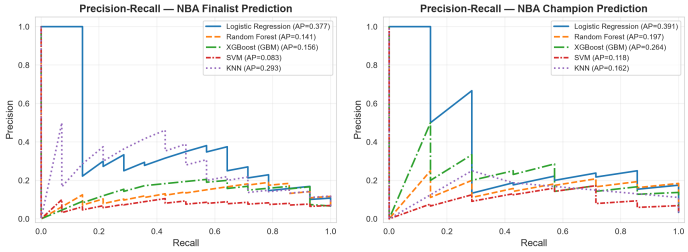


Figure 2. Precision–Recall curves for NBA finalist and champion prediction. LR achieves the highest AP in both tasks (0.377 and 0.391, respectively); the sharp decline at high recall for champion prediction reflects the severe class imbalance (7 positives in 210)

Feature importance (chart in the supplementary repository) shows NRtg_zscore is the most influential feature for Random Forest (≈40%) and XGBoost (≈48%) in finalist prediction, indicating that relative seasonal performance is more discriminative than absolute performance metrics; WinPct_zscore ranks second for Random Forest (34%), while eFG% ranks second for XGBoost (34%), suggesting that shot-making efficiency adds signal beyond overall team performance. For champion prediction, the pattern shifts: WinPct_zscore dominates for both Random Forest (41%) and XGBoost (58%), while NRtg_zscore contributes 37% and 22% respectively, indicating that sustained winning consistency becomes the primary discriminator at the championship level.

Reliability diagrams (before/after isotonic calibration, supplementary repository[1]) show that Random Forest is overconfident in the 0.3–0.6 range, while KNN improves markedly after calibration, aligning with the view that under class imbalance, raw probabilities from tree- and distance-based models benefit from post-hoc calibration (shown here only for RF and KNN in the NBA case study).

Sweeping decision thresholds on the test set shows that the F1-optimal cutoff for Logistic Regression in the champion prediction task is 0.959, increasing F1 from 0.273 (at 0.5) to 0.400 while Recall drops from 0.857 to 0.286. Because this threshold is tuned on the test set, these gains are diagnostic of calibration issues rather than unbiased deployment estimates. For Random Forest, the best threshold is 0.415 (F1=0.323, Recall=0.714 vs. F1=0.214 at 0.5). XGBoost is comparatively well-calibrated (F1=0.381 at both 0.5 and its optimal 0.523). SVM and KNN, which predict no positives at 0.5, reach F1 scores of 0.278 and 0.267 at thresholds 0.040 and 0.143. Overall, the default 0.5 threshold is unsuitable under class imbalance; practical use requires explicit threshold tuning and calibration. Full threshold–metric tables are in the supplementary repository.

The best-AUC model (Logistic Regression, $C = 0.1$) was retrained on 2000–2024 and validated on 2025: the OKC Thunder (actual champion) was correctly ranked **#1** with P(champion)=13.84%. Retrained on 2000–2025 and applied to 2025–26 statistics, OKC again tops both rankings for 2026 (8.4% Finals, 14.9% champion); the full forecast chart is in the supplementary repository.

5. Case Study 2: FIFA World Cup Pre-Tournament (1958–2022)

The *International Football Results* repository [Mart 2024] includes every international match since 1872 (49 000+ matches). The World Cup features parallel the NBA's: **elo_z** corresponds to NRtg_zscore (overall strength), **q_win_pct_conf_z** to WinPct_zscore (qualifier win rate), and **q_gd_pg_conf_z** to absolute NRtg (scoring margin per game); **titles_z** (histor-

ical pedigree) is WC-specific with no NBA analogue. The targets are `reached_final` (34 positives, 8.2%) and `champion` (17 positives, 4.1%), with positive rates close to the NBA's (6.7%, 3.4%), supporting the structural analogy (though not independently validating it).

ELO ratings are computed cumulatively over all matches since 1872, with a snapshot taken approximately one month before each World Cup, following the update rule $\text{ELO}_i' = \text{ELO}_i + k \cdot \delta(m) \cdot (S_i - E_i)$ [Gilch and Müller 2018, Groll et al. 2019], where k is a tournament-importance factor (60 for the World Cup, 40 for qualifiers, 20 for friendlies), $\delta(m)$ is a goal-difference multiplier (margin 1→1.0; 2→1.5; 3→1.75; ≥4 → increasing), and a home bonus of 75 ELO points is applied on non-neutral ground.

The final WC feature set was chosen after exploratory analysis of confederation win-rate distributions and the full correlation structure; all subsequent modeling uses only these features. From qualifying statistics, we compute per-season confederation-level z-scores (Eq. 1). Because CONMEBOL's median win% (≈0.50) is lower than UEFA's (≈0.70) due to its full round-robin among 10 nations, global z-scoring would underrate CONMEBOL teams; confederation-level z-scoring fixes this. The final features are **elo_z** (global ranking), **q_win_pct_conf_z** (qualifying win rate), **q_gd_pg_conf_z** (goal difference per game), and **titles_z** (historical pedigree). elo_z has the strongest individual correlation with both targets (0.30, 0.23) and best separates finalists from non-finalists. titles_z, correlated with elo_z (0.51) and bimodally distributed due to title concentration, still adds independent signal. The four features are non-redundant ($\max|r| = 0.67$). Confederation win-rate distributions, the full correlation matrix, per-feature class distributions, and the Random Forest feature-importance chart are in the supplementary repository.

5.1. Temporal Split

The primary split trains on editions 1958–2010 (14 World Cups, 317 instances, excluding teams with incomplete qualifying data) and tests on 2014–2022 (3 World Cups, 96 instances). With only 3 champion positives in the test set, each misclassification changes AUC by ≈ 0.33, so we also use Leave-One-Cup-Out cross-validation (LOCO-CV), holding out each edition in turn (up to 17 folds). Because each LOCO-CV training set includes all *other* editions, including later ones, we treat LOCO-CV as a robustness check across tournament configurations rather than strict temporal validation. Temporal validation is instead provided by the main train–test split and the walk-forward analysis in Section 6. Within each training window, we use `TimeSeriesSplit` with 5 folds for internal CV.

5.2. Results

Table 3 reports the full evaluation. For finalist prediction, SVM yields the highest AUC (0.870) and, with Random Forest, the highest Recall (0.833); Random Forest achieves the best F1 (0.400). For champion prediction, **Random Forest and SVM tie with the highest AUC (0.921)**, but behave very differently at threshold 0.5: SVM attains perfect Recall (1.000, F1=0.273), while Random Forest finds 2 of 3 test champions (Recall=0.667, best F1=0.286). XGBoost and KNN, despite AUCs of 0.885 and 0.900, output no positive predictions at threshold 0.5 (F1=0.000), mirroring the high-AUC, zero-Recall pattern seen for some NBA models but affecting XGBoost and KNN instead of SVM. This confirms that the default 0.5 threshold is unsuitable for several models under severe class imbalance,

disqualifying them from ensemble forecasting despite strong ranking performance.

Table 3. World Cup results: finalist and champion prediction. Bold = best per metric.

Task	Model	AUC	TP	FP	FN	TN	Recall	F1
Finalist	**SVM**	**0.870**	5	16	1	74	**0.833**	0.370
	Random Forest	0.862	5	14	1	76	**0.833**	**0.400**
	KNN	0.817	1	1	5	89	0.167	0.250
	XGBoost	0.816	0	1	6	89	0.000	0.000
	Log. Regression	0.809	4	17	2	73	0.667	0.296
Champion	**Random Forest**	**0.921**	2	9	1	84	0.667	**0.286**
	SVM	**0.921**	3	16	0	77	**1.000**	0.273
	Log. Regression	0.903	2	16	1	77	0.667	0.190
	KNN	0.900	0	0	3	93	0.000	0.000
	XGBoost	0.885	0	0	3	93	0.000	0.000

Table 4 extends the analysis with Average Precision, bootstrap confidence intervals, and DeLong tests, using Random Forest as the reference. None of the champion-prediction AUC differences are significant at $\alpha = 0.05$ ($p = 0.384$ for LR, 0.205 for XGB, 1.000 for SVM, 0.460 for KNN), so all five models are statistically indistinguishable in ranking quality despite their very different binary behavior at threshold 0.5. McNemar's test on overall correctness favors the abstaining models (XGB, KNN) over LR ($p = 0.0013$ for LR vs. XGB); SVM does not differ significantly from LR or Random Forest ($p = 1.000$ and 0.109) but does differ from XGB and KNN ($p = 0.0044$ each), consistent with SVM no longer abstaining at threshold 0.5. As in the NBA case, this shows that accuracy-based paired tests are ill-suited for imbalanced tasks and motivates our reliance on AUC-ROC, Average Precision, and other ranking metrics for model selection.

Table 4. WC champion prediction: Average Precision (AP), AUC, and DeLong p-values comparing each model to Random Forest (reference). Bootstrap 95% CI for AUC. Test set contains only 3 champions; statistical power is limited.

Model	AUC	95% CI	AP	p (DeLong)
Log. Regression	0.903	[0.789, 0.979]	0.210	0.384
Random Forest	0.921	[0.821, 0.989]	0.250	—
XGBoost	0.885	[0.747, 0.979]	0.190	0.205
SVM	0.921	[0.800, 1.000]	0.328	1.000
KNN	0.900	[0.809, 0.979]	0.194	0.460

We complement binary metrics with a ranking evaluation: for each World Cup in the test set (2014, 2018, 2022), we rank all 32 teams by predicted champion probability and check if the true champion is in the top-k. In 2014 (Germany), LR, RF, and SVM rank the champion #2 (XGBoost #3, KNN #4). In 2022 (Argentina), only SVM ranks it first; RF and XGBoost place it #2, LR and KNN #3. The main difficulty is 2018 (France), where no model ranks the champion above #6 (SVM best at #6, XGBoost worst at #9). SVM attains the highest MRR (0.556, Top-3=2/3), ranking the champion #1 in 2022 and #2 in 2014, and now combines this with perfect binary Recall (1.000), unlike earlier NBA-style high-AUC/zero-Recall behavior. LR, RF, and XGBoost never rank a champion first (Top-1=0/3; MRR 0.319, 0.381, 0.315), and KNN is the only model to miss the top-3 in 2014, yielding the worst Top-3 (1/3) and MRR (0.242). Full per-model rankings and confusion matrices at threshold 0.5 are in the supplementary repository.

On the WC champion test set, Random Forest's optimal threshold is 0.658, yielding F1=0.444 (vs. 0.286 at 0.5) with Recall=0.667. SVM has Recall=1.000 at the default threshold (F1=0.273); its optimal threshold (0.125) reduces recall to 0.667 to reach F1=0.444. Because these thresholds are tuned on test, the metrics are diagnostic, not deployment-ready. XGBoost and KNN, which emit no positives at 0.5, improve at low thresholds

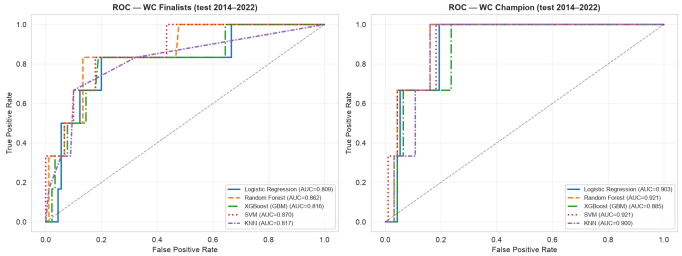


Figure 3. ROC curves for finalist (left) and champion (right), WC test 2014–2022. For champion prediction, Random Forest and SVM tie for the highest AUC (0.921).

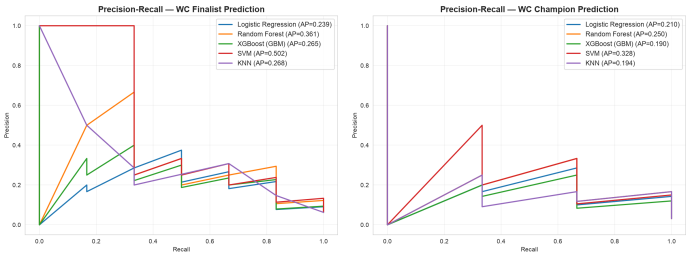


Figure 4. Precision–Recall curves for WC finalist and champion prediction. The curves illustrate the precision–recall trade-off for the two prediction tasks, with the champion prediction showing lower precision at higher recall due to the severe class imbalance.

($t^* \approx 0.07$) to F1=0.364 (Recall=0.667) and F1=0.286 (Recall=1.000), respectively. As in the NBA case study, under class imbalance, threshold calibration is critical: 0.5 can be too conservative (XGBoost, KNN, previously SVM) or adequate (Random Forest), so there is no universal default. The full threshold table is in the supplementary repository.

Models retrained on 1958–2018 correctly ranked Argentina as the top pre-tournament favourite for 2022 (P=9.1%), with the ELO feature capturing the strength of traditional powers despite uneven qualifying. For 2026, the ensemble forecast identifies France as favourite (6.3%), followed by Argentina (6.1%), Brazil (5.9%), and Germany (5.8%); the full chart and per-model breakdown are in the supplementary repository.

6. Cross-Domain Results and Discussion

Table 5 summarizes champion-prediction performance across both case studies, revealing consistent observations.

Table 5. Champion prediction summary: best AUC, Recall, and F1 at threshold 0.5 per case study. Test sets are temporally disjoint from all training data.

Case Study	Best model (AUC)	AUC	Best Recall	Best F1
NBA (2019–2025)	Log. Regression	0.933	0.857 (LR)	0.381 (XGB)
WC (2014–2022)	RF / SVM (tie)	0.921	1.000 (SVM)	0.286 (RF)
Naive accuracy baseline (majority class)		—	—	96.7% / 96.9%

Qualifying-phase statistics carry substantialpredictive signals, and the best model varies with data richness: AUCs above 0.90 in both sports empirically support the framework's structural-analogy claim. While direct comparison to prior work is impossible because of different tasks, our 0.921–0.933 AUCs exceed Groll et al. [Groll et al. 2019] (0.75–0.85, match-level World Cup prediction) and Gilch and Müller [Gilch and Müller 2018] (≈0.70, ELO-only 2018 World Cup model), largely because aggregating many qualifying matches reduces noise relative to single-match prediction (though it cannot predict specific matches). For the NBA, Logistic Regression performs best (82-game season, 5 features, 19 training seasons), consistent with Terner and Franks [Terner and Franks 2021], who show simple

linear models can rival complex ones when features are strong and data abundant. In the sparser World Cup setting, Random Forest and SVM share the top champion AUC (0.921), but only Random Forest yields a well-calibrated default decision (Recall=0.667 at threshold 0.5); SVM reaches Recall=1.000 at this threshold but with lower precision.

High AUC does not guaranty useful binary classification: In the NBA, KNN and SVM achieve high AUCs (>0.82) but no positive predictions at threshold 0.5; in the World Cup, XGBoost and KNN show the same pattern (AUC 0.885 and 0.900, Recall=0), while SVM—previously excluded due to model-selection instability (Section 5)—reaches perfect Recall once fixed. Thus, the high-AUC/zero-Recall pattern is real but model- and domain-specific, and threshold calibration or F1-based thresholding from the Precision–Recall curve is still needed before deployment. DeLong tests (Tables 2, 4) show that in the NBA, LR's AUC is significantly higher than SVM's and KNN's ($p = 0.046$, $p = 0.038$), whereas in the World Cup, AUC differences for champion prediction are not significant at $\alpha = 0.05$. Still, binary performance gaps (Recall, Precision, F1) remain large, underscoring the need to assess models at multiple operating points rather than by AUC alone.

Qualifier stats and feature importance: In the WC Random Forest, elo_z explains ≈45% of variance, matching [Gilch and Müller 2018]; qualifying stats contribute 12–15% each after confederation normalization, and titles_z 28–29%, indicating largely non-redundant signals. In the NBA, NRtg_zscore dominates finalist prediction (≈40–48%), consistent with work highlighting relative within-season performance, while WinPct_zscore dominates champion prediction (41–58%), reflecting sustained winning consistency. This points to a cross-sport pattern where relative performance metrics (z-scored net rating or ELO) outperform absolute metrics, meriting further validation.

Cross-validation supports stability in the NBA, less so in the World Cup: In the NBA, a walk-forward analysis with three temporal cut points (2015, 2016, 2017, each tested on the next seven seasons) yields champion-prediction AUCs of 0.938, 0.927, and 0.925 (mean 0.930 ± 0.006), closely matching the primary-split AUC of 0.933 and indicating a stable predictive signal over time. In the World Cup, Leave-One-Cup-Out cross-validation (LOCO-CV; Section 5) used as a cross-configuration robustness check yields mean champion-prediction AUCs across 17 folds of 0.810 ± 0.162 for Random Forest and 0.843 ± 0.169 for Logistic Regression, both lower and more variable than the single-split result of 0.921. This variability shows that WC performance is less stable across tournament eras than in the NBA and reinforces the limitation that single-split WC estimates, based on only three test champions, should be interpreted cautiously.

Practical implications: Logistic Regression has the best NBA Recall (0.857) and SVM the best World Cup Recall (1.000), but threshold calibration is needed because some high-AUC models still predict no positives at the default threshold. Feature importance identifies which performance dimensions (offensive efficiency, qualifying win rate, historical pedigree) best predict championships, informing coaching strategy, scouting, and roster planning.

7. Conclusion

This paper proposes and validates a generalizable supervised ML framework for predicting sports champions, formalizing the common two-phase qualifying-to-championship

structure. Applied to the NBA (2000–2025) and FIFA World Cup (1958–2022) under a temporal protocol that prevents data leakage (all preprocessing, scaling, tuning, and model selection confined to training/validation), it attains champion-prediction AUCs of 0.933 and 0.921, respectively (World Cup test set spans only three tournaments; interpret cautiously). Qualifying-phase statistics are strongly predictive across sports, though the best model depends on data richness. High AUC does not guarantee useful binary classification at a fixed threshold: at 0.5, most models in both domains output no positive predictions, with the few exceptions being domain-specific. DeLong tests, McNemar tests, and bootstrap confidence intervals quantify uncertainty in performance differences. NBA results are stable across evaluation windows (walk-forward mean AUC 0.930 ± 0.006), whereas World Cup results are less so (LOCO-CV mean AUC 0.810–0.843, $\sigma \approx 0.17$ over 17 folds), weakening stability claims there. Nonetheless, the held-out test sets rank the OKC Thunder as the top NBA candidate for 2025 and Argentina as the top candidate for the 2022 World Cup.

Several limitations qualify these results. First, the World Cup test set has only three champion instances, so each misclassification substantially alters performance estimates (ΔAUC ≈ 0.33 per error); temporal cross-validation reduces but does not eliminate this, leaving wide confidence intervals (Table 4) and requiring cautious interpretation. Second, features are limited to collective performance statistics and omit player injuries, squad market value, coaching changes, and home advantage beyond the ELO home bonus. Third, probability calibration is shown only for the NBA; extending it to the World Cup would strengthen claims about practical deployment. Fourth, the framework assumes structurally similar qualifying and championship phases, which may fail in competitions with different qualifying formats (e.g., promotion/relegation leagues). Fifth, although the McNemar test is correctly applied, under extreme class imbalance its statistic is dominated by true negatives, yielding counterintuitive results; we therefore rely mainly on AUC-ROC, Average Precision, and ranking metrics for model comparison. Finally, small test sets limit the DeLong tests' power to detect moderate effect sizes.

Promising directions include extending the framework to more competitions (e.g., Brasileirão and NFL/Super Bowl [Duque 2024, Gifford and Bayrak 2023]); incorporating pre-qualifying information such as FIFA rankings and squad market values; using Learning-to-Rank formulations that directly optimize candidate ordering, better matching the structure of championship prediction; applying systematic probability calibration (Platt scaling, isotonic regression) across domains and models; modeling finalists and champions jointly via multi-target prediction to exploit tournament hierarchy; and extending to women's and youth competitions to further test generalizability.

References

Bunker, R. and Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73:1285–1322.

Bunker, R., Yeung, C., and Fujii, K. (2024). Machine learning for soccer match result prediction. In *Artificial Intelligence, Optimization, and Data Sciences in Sports*, pages 7–49. Springer.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845.

Duque, A. (2024). Brasileirão dataset. GitHub repository.

Felice, C. et al. (2024). Artificial intelligence and machine learning in sports prediction: A systematic review. *Electronics*, 13(17):3444.

Gifford, M. and Bayrak, T. (2023). A predictive analytics model for forecasting outcomes in the national football league games using decision tree and logistic regression. *Decision Analytics Journal*, 8:100296.

Gilch, L. A. and Müller, S. (2018). On Elo based prediction models for the FIFA World Cup 2018. *Journal of Quantitative Analysis in Sports*, 14(4):187–200.

Groll, A., Ley, C., Schauberger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4):271–287.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):23–39.

Horvat, T. and Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1380.

Lin, H.-Y., Lien, J.-J. J., and Chang, Y.-N. (2024). Machine learning in sports analytics: Volleyball match outcome prediction. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6. IEEE.

Loeffelholz, B., Bednar, E., and Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1).

Mart, J. (2024). International football results from 1872 to present. GitHub repository.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Moya, D., Tipantuña, C., Villa, G., Calderón-Hinojosa, X., Rivadeneira, B., and Álvarez, R. (2025). Machine learning applied to professional football: Performance improvement and results prediction. *Machine Learning and Knowledge Extraction*, 7(3):85.

Song, G. et al. (2024). Predicting the FIFA World Cup Qatar 2022 using machine learning and SHAP analysis. *Applied Sciences*, 14(8):3169.

Sumitro Datta (2024). NBA/ABA/BAA Stats. Kaggle dataset.

Terner, Z. and Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, 8:1–23.

Thabtah, F., Zhang, L., and Abdelhamid, N. (2019). Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116.