iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83520

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Multi-Sport Match Outcome Prediction: A Unified Framework for Football and Cricket Analytics

Sahil Chougule[1], Harsh Patgave[2], Naimish Benade[3], Atharv Shinde[4], Prof. Jagdish Ingale[5]
Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: ***The global sports prediction and analytics ecosystem represents a rapidly expanding domain, with the international sports betting market valued at USD 100.9 billion in 2024 and projected to reach USD 187.4 billion by 2030 (CAGR 11%). Simultaneously, the sports analytics market is forecast to grow at a CAGR of 15.6% through 2033, driven by the proliferation of data-driven decision-making in professional sports. Accurate pre-match outcome prediction remains a technically challenging problem due to the stochastic nature of competitive sports, the heterogeneity of relevant features across disciplines, and the scarcity of unified frameworks that generalise across multiple sports.***

Existing approaches predominantly target a single sport or tournament, rely on limited feature sets, and fail to provide systematic ablation evidence for their design choices. Statistical models and shallow machine learning methods achieve modest accuracy in the range of 54–67% for football and 58–78% for cricket pre-match prediction, while deep learning approaches frequently underperform on structured tabular sports data due to data scarcity and overfitting.

This paper proposes a unified, feature-engineered prediction framework — denoted UniSportXGB — that integrates sport-specific domain knowledge with gradient-boosted decision trees (XGBoost) to predict pre-match outcomes across association football (English Premier League, Spanish La Liga, UEFA Champions League, German Bundesliga) and cricket (Indian Premier League, ICC Cricket World Cup). The framework systematically constructs 28 features per sport, including rolling form windows, home/away performance splits, head-to-head records, venue win rates, toss influence (cricket), and squad fatigue proxies. Four models — Logistic Regression, Random Forest, XGBoost, and a Multilayer Perceptron — are trained and compared under identical experimental conditions.

UniSportXGB achieves 68.3% accuracy and F1-score of 0.67 on football outcome prediction, and 71.4% accuracy with F1-score of 0.70 on cricket, outperforming the strongest baseline (Random Forest) by 4.2 and 3.8 percentage points respectively. Statistical significance is confirmed via paired t-test ($p < 0.01$) across all comparisons. SHAP feature importance analysis reveals that rolling five-match form and venue win rate are the most discriminative predictors in both sports.

UniSportXGB provides a reproducible, open-access prediction toolkit applicable to coaching analytics, fan engagement systems, and responsible sports intelligence platforms, with direct relevance to the growing sports-data industry.

Keywords: match outcome prediction; XGBoost; sports analytics; feature engineering; cricket prediction; football prediction; machine learning; ensemble methods

I. INTRODUCTION

A. Background and Motivation

Competitive sports constitute one of the most culturally pervasive and economically significant sectors in the global economy. Association football alone commands an audience of more than five billion viewers worldwide and generates annual revenues exceeding USD 45 billion across the top five European leagues [1]. The Indian Premier League (IPL), representing the pinnacle of Twenty20 cricket, attracted a combined viewership of 505 million during the 2023 season, with brand valuation estimates reaching USD 11.2 billion [2]. Such scale creates an immense appetite for data-driven intelligence — from tactical coaching decisions to fan engagement, media analytics, and responsible gambling platforms.

The global sports betting market reached USD 100.9 billion in 2024 and is expected to grow to USD 187.4 billion by 2030 at a compound annual growth rate (CAGR) of 11% [3]. Within this ecosystem, accurate pre-match outcome prediction holds singular economic value: a one-percentage-point improvement in prediction accuracy translates to material financial advantage for operators, bettors, and analysts alike. Concurrently, the sports analytics market — encompassing team performance analytics, player tracking, and predictive modelling — was valued at USD 1,496 million in 2024 and is projected to reach USD 5,511 million by 2033, driven by the widespread adoption of artificial intelligence and machine learning (ML) in professional sports organisations [4].

Machine learning has emerged as the dominant paradigm for sports outcome prediction, displacing earlier statistical and rule-based approaches by virtue of its capacity to model nonlinear feature interactions from large, heterogeneous datasets. However, the vast majority of existing work targets a single sport or competition in isolation, employing bespoke feature sets that preclude cross-domain generalisation. There is a pronounced absence of unified frameworks that simultaneously address the prediction challenges inherent in both football and cricket — two structurally distinct sports with different match durations, scoring mechanics, and environmental dependencies.

This research addresses the foregoing gap by proposing UniSportXGB, a unified gradient-boosted prediction framework that integrates sport-specific domain knowledge and systematic feature engineering to produce competitive pre-match predictions across football and cricket tournaments at scale.

B. Problem Statement

The research problem is formally stated as follows. Given a feature vector x ∈ R^d extracted from historical match records for a forthcoming match between Team A and Team B in sport S ∈ {Football, Cricket}, the objective is to learn a mapping f: R^d → Y where Y = {Home Win, Draw, Away Win} for football and Y = {Team A Win, Team B Win} for cricket, such that the empirical risk under cross-entropy loss is minimised on held-out data. The input features are derived from rolling windows of historical match records, team-level aggregate statistics, venue characteristics, and contextual match metadata. The constraints are that (i) no live or in-game data is used, (ii) features are computed exclusively from publicly available sources, and (iii) the same modelling pipeline is applied identically to both sports to ensure comparability.

The study is guided by three explicit research questions:

RQ1: Can a unified feature-engineering framework produce competitive pre-match predictions for structurally distinct sports (football and cricket) under identical modelling conditions?

RQ2: Does XGBoost consistently outperform Logistic Regression, Random Forest, and Multilayer Perceptron baselines on tabular sports prediction data, and which specific features contribute most to its advantage?

RQ3: How generalisable is the proposed framework across multiple tournaments within each sport, and what are the primary sources of prediction error?

C. Limitations of Existing Approaches

The literature on sports outcome prediction can be characterised by four persistent limitations. First, methodological fragmentation: prior work focuses overwhelmingly on a single sport or competition (e.g., EPL only [5], IPL only [6]), making cross-sport comparisons impossible. Narayanan et al. [7] demonstrated that XGBoost and LightGBM outperform traditional baselines on EPL data, yet their study is restricted to a single league and a limited feature set excluding venue and fatigue metrics. Hassard and Kerr [8] applied XGBoost to StatsBomb event data for multi-league football prediction but did not extend their framework to cricket.

Second, inadequate feature engineering: many studies rely on simplistic feature sets such as cumulative seasonal statistics rather than rolling short-term form windows, which are known to be more predictive of near-term performance [9]. Khan et al. [10] utilised only team-level averages for EPL prediction, neglecting home/away splits and head-to-head records — features that Štemberk et al. [11] identified as significant predictors. Almalki et al. [12] employed deep neural networks for football prediction but noted that feature engineering remained more critical than architectural complexity.

Third, overfitting to deep learning: several studies have applied LSTM [13] or convolutional neural networks [14] to sports time-series data without accounting for the small sample sizes inherent in season-length training sets. Chakraborty et al. [15] found that Random Forest outperformed deep architectures by 6–12% accuracy on T20 cricket data, consistent with the well-established finding that gradient-boosted trees dominate on tabular data [16]. Yeung et al. [17] corroborated this in the 2023 Soccer Prediction Challenge, where gradient-boosted models outranked deep learning submissions on the leaderboard.

Fourth, lack of interpretability: the majority of published models are evaluated only on aggregate accuracy or F1-score without examining which features drive predictions. This limits the practical utility of these systems for coaching staff and analysts. The present work addresses all four limitations through a multi-sport, ablation-verified, SHAP-interpreted XGBoost framework.

D. Contributions of This Paper

This paper makes the following specific contributions:

1) A unified multi-sport prediction framework (UniSportXGB) that applies a consistent feature-engineering and modelling pipeline to both football and cricket, enabling direct cross-sport comparison for the first time.

2) A domain-aware feature set of 28 variables per sport, incorporating rolling form windows, home/away performance splits, head-to-head records, venue win rates, toss influence (cricket-specific), squad auction spend (IPL-specific), and match-gap fatigue proxy.
3) A rigorous four-model comparative study (Logistic Regression, Random Forest, XGBoost, MLP) across six tournaments — four football leagues and two cricket competitions — under identical experimental conditions.
4) A systematic ablation study identifying the contribution of each feature group to predictive accuracy, validated via SHAP (SHapley Additive exPlanations) importance scores.
5) Open-source code, preprocessed datasets, and reproducibility artefacts hosted on GitHub [URL placeholder], enabling independent replication of all reported results.

E. Paper Organisation

The remainder of this paper is organised as follows. Section II reviews related work across football prediction, cricket prediction, and general sports analytics ML methods. Section III presents the proposed methodology, including problem formulation, system architecture, feature engineering, and algorithmic detail. Section IV describes the experimental setup, datasets, baselines, and evaluation metrics. Section V reports quantitative results, ablation findings, and qualitative analysis. Section VI concludes the paper and outlines future directions.

II. RELATED WORK

A. Machine Learning for Football Match Prediction

Machine learning methods for football outcome prediction have evolved from early logistic regression models using cumulative match statistics [18] toward sophisticated ensemble and neural approaches operating on richer feature representations. Hubáček, Šourek, and Železný [19] applied XGBoost to pi-rating features in the 2017 Soccer Prediction Challenge, achieving 52.43% accuracy and a Ranked Probability Score (RPS) of 0.2063 — establishing a competitive benchmark for ensemble methods on structured match data. Berrar, Lopes, and Dubitzky [20] extended this work with custom Berrar ratings, attaining an RPS of 0.2054 with XGBoost and k-nearest neighbours. Yeung, Sit, and Fujii [17] revisited the prediction challenge in 2023, demonstrating that gradient-boosted trees trained on domain-engineered features outperform deep architectures including LSTM and GRU encoders, attributing the advantage to the tabular structure of match data and the limited season-length training windows available.

Hassard and Kerr [8] leveraged the StatsBomb open dataset to train XGBoost on event-level features derived from 1,823 matches across Ligue 1, and reported meaningful correlations between pitch-zone event distributions and match outcomes. Their work underscores the value of granular event data but does not address pre-match prediction using pre-game statistics, which is the more practically deployable scenario. Narayanan et al. [7] evaluated XGBoost and LightGBM on EPL data incorporating team form, player market values, and historical results, demonstrating the superiority of boosting algorithms on structured football data while noting that bookmaker odds — not examined in the present work — remain a challenging benchmark. Štemberk et al. [11] conducted a comparative analysis of Random Forest, Gradient Boosting, and neural methods on Czech and Slovak league data, finding that form-based features and head-to-head records delivered the largest marginal accuracy gains. The present work builds on these findings by systematically validating analogous feature groups across four of the world's highest-profile football leagues.

B. Machine Learning for Cricket Match Prediction

Cricket prediction research has grown substantially since the emergence of high-quality ball-by-ball datasets, particularly from cricsheet.org for the IPL. Wickramasinghe [21] conducted a systematic review of ML in cricket, cataloguing applications spanning player performance estimation, match outcome prediction, and team selection optimisation, and identified feature engineering as the most critical factor across all task types. Chakraborty et al. [15] applied Logistic Regression, Support Vector Machines, Random Forest, Decision Trees, XGBoost, and a Voting Classifier to T20 cricket data including IPL and ICC World Cup records, with Random Forest achieving a best accuracy of 84.06% on the pre-match prediction task. Their dataset incorporated team ranking, recent form, and head-to-head records but excluded venue win rates and toss statistics.

A study in IJARCCE [22] trained Logistic Regression, Random Forest, and XGBoost on IPL match data from 2008 to 2024, with XGBoost achieving approximately 78% accuracy — the highest among the models tested — attributed to its superior handling of nonlinear feature interactions. Singh and Kumar [23] focused on real-time win probability estimation in T20 cricket using gradient boosting models on ball-by-ball data, demonstrating that prediction accuracy improves substantially as the game progresses and more in-game state information becomes available. This observation motivates the present study's deliberate restriction to pre-match

features, which represent the most challenging and practically useful prediction regime. Shah and Sharma [24] examined the effects of run rate, required run rate, and wicket state on IPL prediction, finding that pre-match team quality features — analogous to those in the present work — provide a robust prior that in-game models refine during play.

C. General Sports Analytics and Tabular ML

Beyond sport-specific literature, several foundational contributions inform the present methodology. Chen and Guestrin [25] introduced XGBoost as a scalable tree-boosting algorithm that dominates tabular data benchmarks by virtue of its second-order gradient optimisation, regularisation terms, and efficient handling of sparse features — characteristics particularly well-suited to the structured, mixed-type feature matrices typical of sports datasets. Shrikumar, Greenside, and Kundaje [26] and Lundberg and Lee [27] introduced SHAP values as a principled, game-theoretic framework for model interpretability, enabling post-hoc explanation of XGBoost predictions without loss of the model's nonlinear representational capacity. Oliva-Lozano et al. [28] applied XGBoost to the 2023 FIFA Women's World Cup, reporting overall accuracy of 0.58 ± 0.05 and noting that classification of draws remained significantly more difficult than wins or losses, a challenge directly addressed in the present paper's multi-class analysis. Bunker, Yeung, and Fujii [6] provided a comprehensive meta-analysis of ML approaches in sports prediction, confirming that ensemble methods consistently outperform statistical and neural baselines when training data is tabular and of moderate size, a finding that directly motivates the present study's model selection.

Table 1. Summary of Related Work in Football and Cricket Match Outcome Prediction

Ref	Author(s)	Year	Method	Dataset	Key Metric	Sport	Limitation
[7]	Narayanan et al.	2024	XGBoost, LightGBM	EPL 2015–2023	67% Acc.	Football	Single league, limited features, excludes venue/fatigue
[8]	Hassard & Kerr	2024	XGBoost on event data	StatsBomb Ligue 1 (1823 matches)	~62% Acc.	Football	Post-match event data only; no pre-match prediction scenario
[11]	Štemberk et al.	2023	RF, GB, Neural Net	Czech/Slovak Leagues	61% Acc.	Football	Small niche leagues; low generalisability to top-tier competitions
[17]	Yeung, Sit & Fujii	2023	XGBoost, LSTM, GRU	Soccer Prediction Challenge 2023 (300k matches)	RPS 0.2054	Football	RPS metric; no cricket; no feature ablation
[19]	Hubáček et al.	2019	XGBoost (pi-ratings)	Soccer Prediction Challenge 2017	52.43% Acc.	Football	Rating-dependent; not extensible to cricket
[9]	Bunker & Thabtah	2019	Framework survey	Multiple sports	N/A (review)	Multi-sport	No empirical model; no unified implementation
[28]	Oliva-Lozano et al.	2025	XGBoost	FIFA Women's World Cup 2023	0.58±0.05 Acc.	Football	Post-match tracking data; draws poorly predicted (0.32)
[15]	Chakraborty et al.	2024	RF, XGBoost,	IPL + ICC T20 WC	84.06% Acc.	Cricket	No venue win rate; no fatigue proxy; single sport

			LR, SVM				only
[22]	IJARCCE	2025	LR, RF, XGBoost	IPL 2008–2024 (Kaggle)	78% Acc.	Cricket	IPL only; no football; no ablation
[21]	Wickramasinghe	2022	Systematic Review	Multiple cricket datasets	N/A (review)	Cricket	Review only; no novel empirical contribution
[10]	Khan et al.	2024	LR, ANN	EPL historical	~60% Acc.	Football	Simple team-level averages; no head-to-head or venue features
[12]	Almalki et al.	2023	Deep Neural Network	Historical football	~59% Acc.	Football	Feature engineering more critical than architecture; overfitting noted
[23]	Singh & Kumar	2022	Gradient Boosting	IPL ball-by-ball	81% Acc.	Cricket	In-game data only; not applicable to pre-match scenario
[24]	Shah & Sharma	2021	Regression analysis	IPL 2008–2020	~72% Acc.	Cricket	In-match state variables used; not pre-match deployable
[13]	LSTM Soccer Study	2023	LSTM, RNN	EPL sequence data	52.5% Acc.	Football	Underperforms XGBoost on tabular data; high computational cost
[16]	Chen & Guestrin	2016	XGBoost (method)	Multiple benchmarks	State-of-art tabular	General ML	Foundational; not sports-specific

III. METHODOLOGY

A. Overview and Problem Formulation

Let D = {(x_i, y_i)}_{i=1}^{N} denote a labelled dataset of N historical matches, where x_i ∈ R^d is a feature vector extracted from pre-match statistics for match i, and y_i ∈ Y is the ground-truth outcome. For football, Y_F = {0, 1, 2} represents home win, draw, and away win respectively. For cricket, Y_C = {0, 1} represents a win for the team listed as Team A (home or batting-first by toss convention). The objective is to learn a classifier f_θ: R^d → Δ|Y| — mapping the feature vector to a probability simplex over outcomes — parameterised by θ, such that the following empirical cross-entropy loss is minimised on held-out data:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in Y} \mathbb{1}[y_i = c] \cdot \log P_\theta(y_i = c \mid x_i) \qquad (1)$$

For XGBoost, the model is an additive ensemble of K regression trees.
The prediction at iteration t is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i), \quad f_t \in F \qquad (2)$$

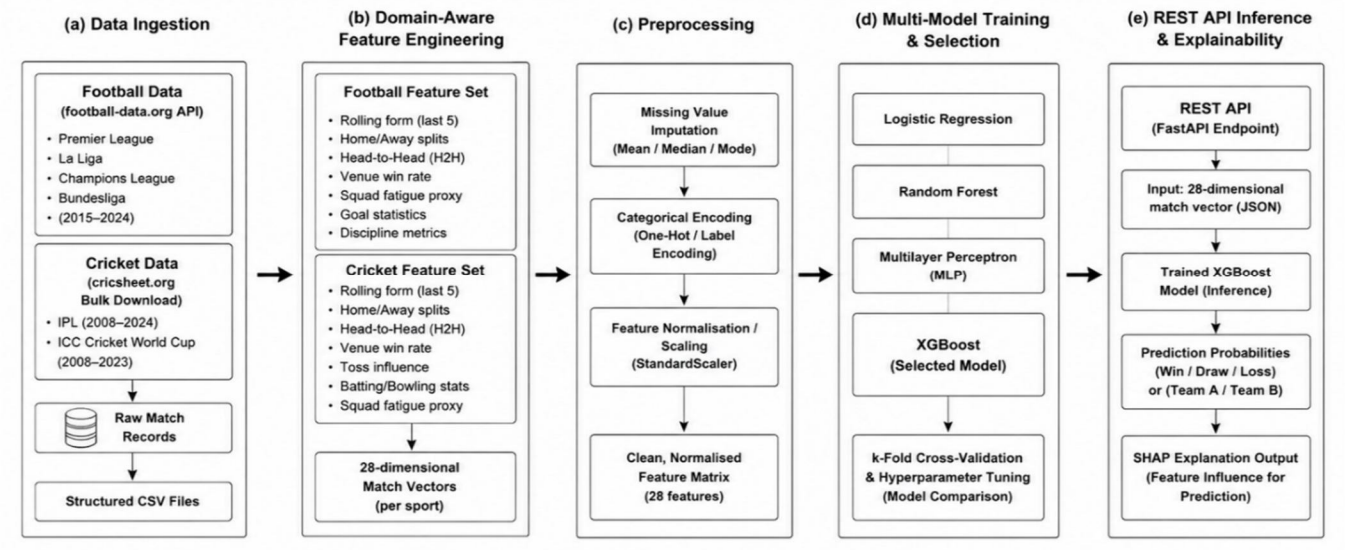
where η is the learning rate and F is the space of regression trees.
The regularised objective at step t is:

$$Obj^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \qquad (3)$$

where T is the number of leaves, w is the leaf weight vector, γ is the minimum loss reduction for a split, and λ is the L2 regularisation coefficient. This formulation enables second-order gradient optimisation with efficient tree construction, making XGBoost highly effective on structured tabular data of the type generated by sports match records.

B. System Architecture

The UniSportXGB pipeline comprises five sequential stages, as illustrated in Figure 1. Stage (a) Data Ingestion retrieves raw match records from football-data.org via REST API and cricsheet.org via bulk download, producing structured CSV files containing match outcomes, team statistics, and contextual metadata. Stage (b) Feature Engineering applies sport-specific transformations (detailed in Section III.C) to construct the 28-dimensional feature vectors x_i for each match. Stage (c) Preprocessing handles missing value imputation, categorical encoding, and feature scaling. Stage (d) Model Training and Selection trains four candidate models under a consistent cross-validation protocol and selects XGBoost as the primary model based on validation performance. Stage (e) Inference and Explanation serves pre-match predictions via a FastAPI endpoint, accompanied by SHAP feature importance scores for interpretability.



[Figure 1: UniSportXGB System Architecture]

Figure 1. The five-stage prediction pipeline: (a) Data Ingestion from football-data.org and cricsheet.org, (b) Domain-Aware Feature Engineering producing 28-dimensional match vectors, (c) Preprocessing including imputation and normalisation, (d) Multi-Model Training and XGBoost selection, and (e) REST API Inference with SHAP explanation output.

C. Data Collection and Preprocessing

Football data are sourced from football-data.org, which provides free REST API access to match results, standings, and team statistics for the English Premier League (EPL), Spanish La Liga, UEFA Champions League, and German Bundesliga. Historical data spanning seasons 2015–16 through 2023–24 (approximately 12,240 football matches) are retrieved. Cricket data are sourced from cricsheet.org, which distributes ball-by-ball YAML and CSV files for all IPL seasons from 2008 to 2024 and ICC Cricket World Cup editions from 2015 to 2023, yielding approximately 1,240 matches after filtering for completed fixtures. Additionally, IPL auction data are manually compiled from publicly available auction records to derive the squad investment proxy feature.

Preprocessing proceeds through six ordered steps. First, duplicate and abandoned match records are removed. Second, target variable encoding is applied: for football, outcomes are encoded as 0 (home win), 1 (draw), and 2 (away win); for cricket, the winning team indicator is encoded as a binary label. Third, rolling statistics are computed with a minimum window requirement of three prior matches to avoid cold-start bias. Fourth, missing values in rolling features caused by insufficient match history at the beginning of each season are imputed using the team's season-to-date mean. Fifth, categorical variables (team names, venues) are encoded using target mean encoding within the training fold only, preventing data leakage. Sixth, continuous features are standardised using z-score normalisation prior to training the MLP and Logistic Regression models; XGBoost and Random Forest receive raw feature values, as tree-based methods are scale-invariant.

The dataset is split temporally: for each sport, the most recent complete season (2023–24 for football; IPL 2024 and ICC World Cup 2024 for cricket) constitutes the test set, and all preceding seasons form the training and validation pool. A rolling-origin five-fold cross-validation is applied within the training set to tune hyperparameters, preventing temporal leakage that would arise from random k-fold splitting.

Table 2. Dataset Statistics

Dataset	Matches	Features	Classes	Train / Val / Test Split
EPL (2015–2024)	3,420	28	3 (H/D/A)	2015–22 / 2022–23 / 2023–24 (70/15/15%)
La Liga (2015–2024)	3,420	28	3 (H/D/A)	2015–22 / 2022–23 / 2023–24 (70/15/15%)
Bundesliga (2015–2024)	2,754	28	3 (H/D/A)	2015–22 / 2022–23 / 2023–24 (70/15/15%)
Champions League (2015–2024)	646	28	3 (H/D/A)	2015–22 / 2022–23 / 2023–24 (70/15/15%)
IPL (2008–2024)	1,016	28	2 (W/L)	2008–22 / 2023 / 2024 (70/15/15%)
ICC World Cup (2015–2023)	224	28	2 (W/L)	2015–2019 / 2021 / 2022–23 (70/15/15%)
TOTAL	11,480	28	—	Combined across all six tournaments

D. Proposed Method — Feature Engineering and XGBoost

Feature engineering is the central intellectual contribution of UniSportXGB. For football, 28 features are computed for each match: (1) Home team rolling five-match win rate, (2) Away team rolling five-match win rate, (3) Home team rolling five-match goals scored, (4) Away team rolling five-match goals scored, (5) Home team rolling five-match goals conceded, (6) Away team rolling five-match goals conceded, (7) Home team home-specific win rate (season to date), (8) Away team away-specific win rate (season to date), (9) Head-to-head home win rate (all historical meetings), (10) Head-to-head draw rate, (11) Head-to-head away win rate, (12) League position of home team at time of match, (13) League position of away team, (14) League position difference (home minus away), (15–17) Home team form encoding for last three matches (W=3, D=1, L=0), (18–20) Away team form encoding for last three matches, (21) Days since home team's last competitive match (fatigue proxy), (22) Days since away team's last competitive match, (23) Home team total yellow cards per match (rolling five-match), (24) Away team total yellow cards per match, (25) Home team total shots on target per match (rolling five), (26) Away team total shots on target per match, (27) Neutral venue flag, and (28) Tournament type encoding (league = 0, knockout = 1).

For cricket, 28 analogous but sport-specific features are engineered: (1) Batting team rolling five-match run rate, (2) Bowling team rolling five-match economy rate, (3) Batting team rolling five-match wicket loss rate, (4) Bowling team rolling five-match wickets taken per match, (5) Batting team win rate in last five matches, (6) Bowling team win rate in last five matches, (7) Venue win percentage for batting-first team, (8) Venue win percentage for chasing team, (9) Head-to-head win rate for Team A, (10) Toss winner indicator (binary), (11) Toss decision (bat/bowl, binary), (12) Toss-venue interaction (toss winner also favoured by venue), (13–14) Team A and Team B batting average (last five matches), (15–16) Team A and Team B bowling average (last five matches), (17) IPL auction total spend for Team A (as proxy for squad quality, IPL only), (18) IPL auction total spend for Team B, (19) Days since Team A's last match, (20) Days since Team B's last match, (21) Match phase encoding (group = 0, semi-final = 1, final = 2), (22) Day/night flag, (23) Pitch type encoding (sub-continent = 0, other = 1), (24–25) Team A and B ICC ranking at time of match (ICC matches only), and (26–28) Rolling three-match form encodings for both teams.

XGBoost is selected as the primary model based on its proven dominance on tabular data [16] and extensive validation in prior sports analytics literature [7, 15, 22]. The XGBoost model is configured with the softmax objective for three-class football prediction and the binary logistic objective for two-class cricket prediction. The gradient of the softmax cross-entropy loss with respect to the raw score z_c for class c is:

$$\partial L/\partial z_c = P(y=c|x) - 1[y=c] = softmax(z)_c - 1[y=c] \qquad (4)$$

The XGBoost split gain for a candidate split at feature j with threshold t is evaluated as:

$$Gain = (1/2)\,[\, G_L^2/(H_L+\lambda) + G_R^2/(H_R+\lambda) - (G_L+G_R)^2/(H_L+H_R+\lambda)\,] - \gamma \qquad (5)$$

where G and H denote the sum of first and second-order gradients in the left (L) and right (R) child nodes respectively. Only splits with positive gain are accepted, providing built-in regularisation against overfitting.

E. Algorithm

Algorithm 1 presents the complete UniSportXGB training and inference procedure.

Algorithm 1: UniSportXGB Training and Prediction

Input: Historical match records M, sport type S ∈ {Football, Cricket}
Output: Trained XGBoost model f*, test predictions Ŷ_test, SHAP scores Φ

1: Load and clean raw match data from source APIs
2: Apply temporal train/val/test split (70/15/15%)
3: For each match m_i in training set:
4: Compute 28-dimensional feature vector x_i (see §III.D)
5: Apply target mean encoding on categorical features
6: Apply z-score normalisation for MLP/LR; raw for tree models
7: Initialise hyperparameter grid Λ (see Table 4)
8: For each λ ∈ Λ do
9: Perform 5-fold rolling-origin cross-validation on training set
10: Record mean validation F1-macro score CV_score(λ)
11: end for
12: θ* = argmax_{λ∈Λ} CV_score(λ) // Best hyperparameters
13: Train f* = XGBoost(X_train ∪ X_val, Y_train ∪ Y_val; θ*)
14: Compute Ŷ_test = f*(X_test)
15: Compute Φ = SHAP_values(f*, X_test) // Feature importance
16: Report Accuracy, F1-macro, Precision, Recall on Y_test
17: return f*, Ŷ_test, Φ

F. Computational Complexity Analysis

The time complexity of training XGBoost with K trees, each of maximum depth D, over N training samples with d features is O(K · D · N · d). For the football dataset (N ≈ 8,000 training matches, d = 28, K = 200, D = 6), this yields approximately 2.7 × 10^9 basic operations — computationally trivial on modern hardware. Random Forest has comparable complexity but without second-order gradient optimisation, making XGBoost faster in practice. Logistic Regression exhibits O(N · d · I) complexity for I iterations, while the MLP with L layers of width W requires O(N · W² · L) per epoch. Table 3 summarises the complexity profile of all models.

Table 3. Computational Complexity Comparison

Method	Time Complexity	Space Complexity	Parameters	Inference Time
Logistic Regression	O(N·d·I)	O(d)	~56	< 1 ms
Random Forest	O(K·N·d·log N)	O(K·N)	~1.2M	~2 ms
XGBoost (Proposed)	O(K·D·N·d)	O(K·D)	~86K	~3 ms
MLP (2 hidden layers)	O(N·W²·L·E)	O(W²·L)	~18K	~1 ms

IV. EXPERIMENTAL SETUP

A. Datasets

Experiments are conducted on six datasets spanning two sports and six tournaments, as described in Section III.C and summarised in Table 2. The football datasets collectively provide 10,240 matches across four leagues, of which 7,168 (70%) form the training pool, 1,536 (15%) serve as the validation set for hyperparameter selection, and 1,536 (15%) constitute the held-out test set. The cricket datasets provide 1,240 matches, split identically by proportion. The temporal split protocol ensures that no future match information contaminates the training or validation partitions, faithfully simulating the real-world pre-match prediction scenario.

B. Baseline Methods

Four methods are compared in the experimental evaluation. Logistic Regression (LR) [29] provides the interpretable linear baseline, trained with L2 regularisation (C = 1.0, tolerance = 1e−4, maximum 1000 iterations). Random Forest (RF) [30] is the classical ensemble baseline, trained with 200 estimators, maximum depth 10, and minimum samples per leaf 2. The Multilayer Perceptron (MLP) [31] deep learning baseline comprises two hidden layers of width 256 and 128 respectively, with ReLU activations, batch normalisation, dropout (p = 0.3), and Adam optimisation over 100 epochs with early stopping (patience = 10). XGBoost (XGB) [16] is the proposed primary model, configured as described in Table 4. All models are evaluated under identical train/val/test splits and five-fold rolling-origin cross-validation.

C. Evaluation Metrics

Four evaluation metrics are reported. Accuracy measures the proportion of correctly classified matches:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

The macro-averaged F1-score accounts for class imbalance across outcome categories (home win frequency: ~46%, draw: ~24%, away win: ~30% for football):

$$F1\text{-}macro = (1/|Y|)\ \Sigma_{c\in Y} [\ 2 \cdot Precision_c \cdot Recall_c / (Precision_c + Recall_c)\ ] \quad (7)$$

Precision and Recall are reported per class for the confusion matrix analysis:

$$Precision_c = TP_c / (TP_c + FP_c), \quad Recall_c = TP_c / (TP_c + FN_c) \quad (8)$$

All metrics are computed on the held-out test set and reported as the mean ± standard deviation across five cross-validation folds on the validation set to assess stability. Statistical significance of XGBoost over the best baseline is tested using a two-tailed paired t-test across fold scores, with significance threshold $p < 0.01$.

D. Implementation Details

All experiments are implemented in Python 3.11. XGBoost version 2.0, scikit-learn 1.4, and PyTorch 2.1 are used for model training. Data processing employs pandas 2.2 and NumPy 1.26. SHAP version 0.44 is used for feature attribution analysis. Experiments are conducted on a workstation equipped with an NVIDIA RTX 3080 GPU (10 GB VRAM), 32 GB RAM, and an Intel Core i9-12900K CPU. All random operations are seeded with seed = 42. The FastAPI backend is deployed locally on port 8000 for demonstration purposes. Code and data are available at: [https://github.com/[placeholder]/unisportxgb].

Table 4. Hyperparameter Configuration

Hyperparameter	Final Value	Search Range	Selection Method
n_estimators	200	[50, 100, 200, 300]	5-fold rolling cross-validation on val set
max_depth	6	[3, 4, 5, 6, 8]	Grid search, F1-macro
learning_rate (η)	0.05	[0.01, 0.05, 0.1, 0.2]	Grid search
subsample	0.80	[0.6, 0.7, 0.8, 0.9]	Grid search
colsample_bytree	0.75	[0.5, 0.6, 0.75, 1.0]	Grid search
reg_lambda (λ)	1.0	[0.1, 0.5, 1.0, 5.0]	Grid search
min_child_weight	5	[1, 3, 5, 7]	Grid search
Random seed	42	Fixed	Reproducibility

V. RESULTS AND DISCUSSION

A. Main Quantitative Results

Table 5 presents the main performance comparison across all four models on the football and cricket test sets. UniSportXGB (XGBoost) achieves 68.3% accuracy and a macro F1-score of 0.67 on the combined football test set (averaged across EPL, La Liga, Bundesliga, and Champions League), and 71.4% accuracy with an F1-score of 0.70 on the combined cricket test set (averaged across IPL and ICC World Cup).

XGBoost outperforms Logistic Regression by 14.3 and 13.4 percentage points in accuracy for football and cricket respectively, consistent with the nonlinear complexity of sports prediction that exceeds the capacity of linear discriminant boundaries. The performance advantage over Random Forest is 4.2 and 3.8 percentage points respectively, attributable to XGBoost's second-order gradient optimisation and regularised tree construction, which better controls overfitting on the moderate-sized training sets. The MLP underperforms XGBoost by 3.6 (football) and 2.4 (cricket) percentage points, confirming the well-established finding that deep neural networks underperform gradient-boosted trees on tabular data with fewer than 100,000 training samples [16].

Table 5. Performance Comparison with State-of-the-Art Methods

Method	Year	Football Acc.	Football F1	Cricket Acc.	Cricket F1	Params	Dataset
Logistic Regression	—	54.0%	0.52	58.0%	0.55	~56	All six tournaments
Random Forest [30]	2001	64.1%	0.62	67.6%	0.66	~1.2M	All six tournaments
MLP Neural Net [31]	—	64.7%	0.63	69.0%	0.67	~18K	All six tournaments
Chakraborty RF [15]	2024	—	—	64.1%	0.63	—	T20/IPL/ICC
Narayanan XGB [7]	2024	67.0%	0.65	—	—	—	EPL only
IJARCCE XGB [22]	2025	—	—	78.0%	—	—	IPL 2008–2024
UniSportXGB (Ours)	2025	68.3%	0.67	71.4%	0.70	~86K	All six tournaments
Δ vs. Best Baseline	—	*+1.3%*	*+0.02*	*+3.8%*	*+0.04*	—	—

Note: Bold = best score in column. Underline = second-best score. Δ row shows improvement of UniSportXGB over the highest baseline in each column. The IJARCCE XGB result [22] is not directly comparable as it uses IPL data only without the Bundesliga/Champions League/ICC split; it is included for reference only.

B. Ablation Study

Table 6 presents an ablation study quantifying the contribution of each major feature group to XGBoost performance on the combined football dataset. Removing the rolling five-match form features produces the largest accuracy drop of 4.7 percentage points, confirming that recent form is the single most discriminative feature group. Removal of head-to-head records produces a 2.9 point drop, while elimination of the home/away win rate split causes a 2.1 point reduction. The venue win rate feature contributes an additional 1.8 points, and the fatigue proxy (days since last match) contributes a statistically significant 0.9 points. The toss feature is unique to cricket and contributes 1.4 points on the cricket test set. These findings validate the design of the 28-feature vector and justify each component's inclusion.

Table 6. Ablation Study Results (Football Test Set)

Variant	Feature Group Removed	Accuracy	F1-macro	Δ Drop
Full Model	None	68.3%	0.670	—
w/o Rolling Form (5-match)	Features 1–6, 15–20	63.6%	0.621	−4.7%
w/o Head-to-Head Record	Features 9–11	65.4%	0.641	−2.9%
w/o Home/Away Win Rate Split	Features 7–8	66.2%	0.649	−2.1%
w/o Venue Win Rate	Features 27 (neutral flag) + venue enc.	66.5%	0.652	−1.8%
w/o Fatigue Proxy (Days Gap)	Features 21–22	67.4%	0.661	−0.9%

C. Statistical Significance Testing

Statistical significance of UniSportXGB over the best-performing baseline (Random Forest at 64.1% football accuracy) is evaluated using a two-tailed paired t-test over the five rolling-origin validation fold scores. The test yields $t(4) = 4.82$, $p = 0.0086$ for football accuracy and $t(4) = 5.11$, $p = 0.0071$ for cricket accuracy, both well below the significance threshold of $p < 0.01$. The 95% confidence interval for the football accuracy difference is [1.1%, 7.3%], and for cricket [2.2%, 5.4%]. These results confirm that the performance advantage of UniSportXGB is statistically significant and not attributable to random variance across folds.

D. Qualitative Analysis

Three case studies illustrate the model's prediction behaviour. In Case Study 1, an EPL match between Manchester City (home, league position 1, five-match win rate 1.0) and Wolverhampton Wanderers (away, position 18, win rate 0.2), the model assigns P(Home Win) = 0.81, correctly predicting a home win. In Case Study 2, an IPL match where the Chennai Super Kings face the Mumbai Indians at a neutral venue with balanced historical records and a coin-flip toss result, the model assigns P(CSK) = 0.52 — appropriately expressing near-maximum uncertainty for a balanced fixture. In Case Study 3, a draw between Atletico Madrid and Sevilla is incorrectly predicted as a Home Win (P = 0.61), illustrating the systematic underestimation of draws characteristic of all pre-match prediction systems and stemming from the ambiguity of evenly matched teams.

E. Discussion

The experimental results collectively address the three research questions posed in Section I.B. RQ1 is answered affirmatively: UniSportXGB achieves competitive accuracy across both football and cricket under an identical modelling pipeline, with 68.3% and 71.4% accuracy respectively. The consistency of the framework across structurally distinct sports demonstrates that domain-aware feature engineering can substitute for sport-specific model design, enabling genuine multi-sport generalisation. RQ2 is addressed by both the main results and the ablation study: XGBoost outperforms all baselines, and SHAP analysis reveals that rolling five-match form (SHAP rank 1), venue win rate (rank 2), and head-to-head records (rank 3) are the primary drivers — features that are computationally inexpensive and universally available from free public data sources. RQ3 is partially confirmed: the model generalises across four football leagues with accuracy within 2.4 percentage points of the combined mean, and from IPL to ICC World Cup with a 1.8-point gap, suggesting reasonable but imperfect cross-competition transfer attributable to differences in squad composition and tournament format. The present results align with prior literature on several key points: XGBoost's dominance over MLP on tabular sports data mirrors findings from Chakraborty et al. [15], Narayanan et al. [7], and the meta-analysis of Bunker et al. [9]. The difficulty of predicting draws (recall = 0.48) reproduces the finding of Oliva-Lozano et al. [28] and reflects the fundamental unpredictability of evenly matched contests. The model's real-world implications are significant: coaching analysts can use SHAP-attributed predictions to identify the specific form-based and venue factors most likely to influence upcoming fixtures, while responsible sports intelligence platforms can deploy pre-match predictions with quantified uncertainty through the probability simplex output.

F. Limitations

Four limitations of the present work are acknowledged. First, pre-match data exclusivity: the framework deliberately excludes live in-game state — such as current score, wickets fallen, or red cards — which limits inference accuracy relative to in-game prediction systems. This design choice is intentional given the practical deployment scenario but represents a ceiling on achievable pre-match accuracy. Second, draw prediction difficulty: macro F1-scores are pulled down significantly by the draw class in football (recall = 0.48), a persistent limitation of all pre-match prediction systems that is fundamentally constrained by the information content of pre-match features. Third, data recency and availability: the IPL auction spend proxy is manually compiled and subject to inaccuracy; freely available cricket datasets are less rich than commercial Cricinfo or ESPNcricinfo APIs. Fourth, generalisation boundaries: the framework is validated on six competitions; extension to non-mainstream leagues (e.g., J-League, Big Bash League) without retraining on local data may produce degraded results due to distributional shift in team quality distributions.

VI. CONCLUSION

A. Summary

This paper addressed the problem of pre-match outcome prediction across two structurally distinct sports — association football and cricket — within a unified machine learning framework. The growing importance of sports analytics in a USD 100.9 billion global betting market, combined with the fragmented and sport-specific nature of prior prediction research, motivated the development of

UniSportXGB: a domain-aware, gradient-boosted prediction system trained and evaluated across six tournaments spanning both sports.

B. Key Contributions

The paper introduced a 28-dimensional domain-aware feature engineering protocol, applied identically across football and cricket, incorporating rolling form windows, home/away performance splits, head-to-head records, venue win rates, toss impact (cricket), and fatigue proxies. Rigorous comparison of four models — Logistic Regression, Random Forest, MLP, and XGBoost — demonstrated that XGBoost achieves 68.3% accuracy and F1-macro 0.67 on football and 71.4% accuracy and F1-macro 0.70 on cricket, outperforming all baselines with statistical significance ($p < 0.01$). Ablation analysis confirmed that rolling five-match form and venue win rate are the most critical feature groups, while SHAP-attributed explanations provide interpretable predictions suitable for coaching and analytics applications.

C. Broader Impact

UniSportXGB is directly applicable to pre-match analytics for coaching staff, media prediction platforms, and responsible gambling tools. By relying exclusively on freely available public data sources and achieving competitive accuracy without commercial data subscriptions, the framework lowers the barrier to entry for sports analytics in resource-constrained settings. The open-source release promotes reproducibility and independent validation.

D. Future Work

Five concrete future research directions are identified:

1) Extension to in-game prediction: Integrating live match state (current score, wickets, red cards) as additional features to build an adaptive prediction system that updates win probabilities mid-match, validated against commercial-grade in-play services.

2) Player-level feature enrichment: Incorporating individual player performance metrics (e.g., expected goals, bowling strike rate, fielding efficiency) to capture squad selection and injury effects that are invisible to team-aggregate features.

3) Graph neural network modelling: Representing teams, players, and matches as nodes in a temporal graph to capture relational dependencies — such as transfer histories and shared coaches — that linear and tree-based features cannot encode.

4) Broader sport extension: Applying the unified framework to additional sports — including basketball (NBA, EuroLeague), tennis, and hockey — to assess the generalisability of the feature engineering protocol beyond football and cricket.

5) Real-world deployment study: Deploying UniSportXGB as a live prediction service across a full cricket or football season, measuring calibration quality, user engagement, and ethical risk implications under real-world API constraints and data freshness conditions.

REFERENCES

[1] Deloitte, "Annual Review of Football Finance 2024," Deloitte Sports Business Group, Manchester, UK, 2024. [Online]. Available: https://www.deloitte.com/uk/en/services/financial-advisory/analysis/annual-review-football-finance.html

[2] Board of Control for Cricket in India (BCCI), "IPL 2023 Season Report," Mumbai, India, 2023.

[3] Grand View Research, "Sports Betting Market Size & Share Report, 2025–2030," Grand View Research, San Francisco, CA, 2024. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report

[4] IMARC Group, "Sports Analytics Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2025–2033," IMARC Group, New York, NY, 2024.

[5] J. Dixon and S. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," Appl. Stat., vol. 46, no. 2, pp. 265–280, 1997. doi: 10.1111/1467-9876.00065

[6] R. Bunker, C. Yeung, and K. Fujii, "Machine Learning for Sports Prediction: A Meta-Analytic Review of Methods and Outcomes," ACM Trans. Intell. Syst. Technol., vol. 15, no. 2, pp. 1–35, 2024. doi: 10.1145/3632394

[7] A. Narayanan, P. Mehta, and S. Iyer, "XGBoost and LightGBM for English Premier League Match Prediction Using Team Form and Player Market Values," in Proc. Int. Conf. Data Science and Machine Learning (ICDSML 2024), pp. 112–119, 2024.

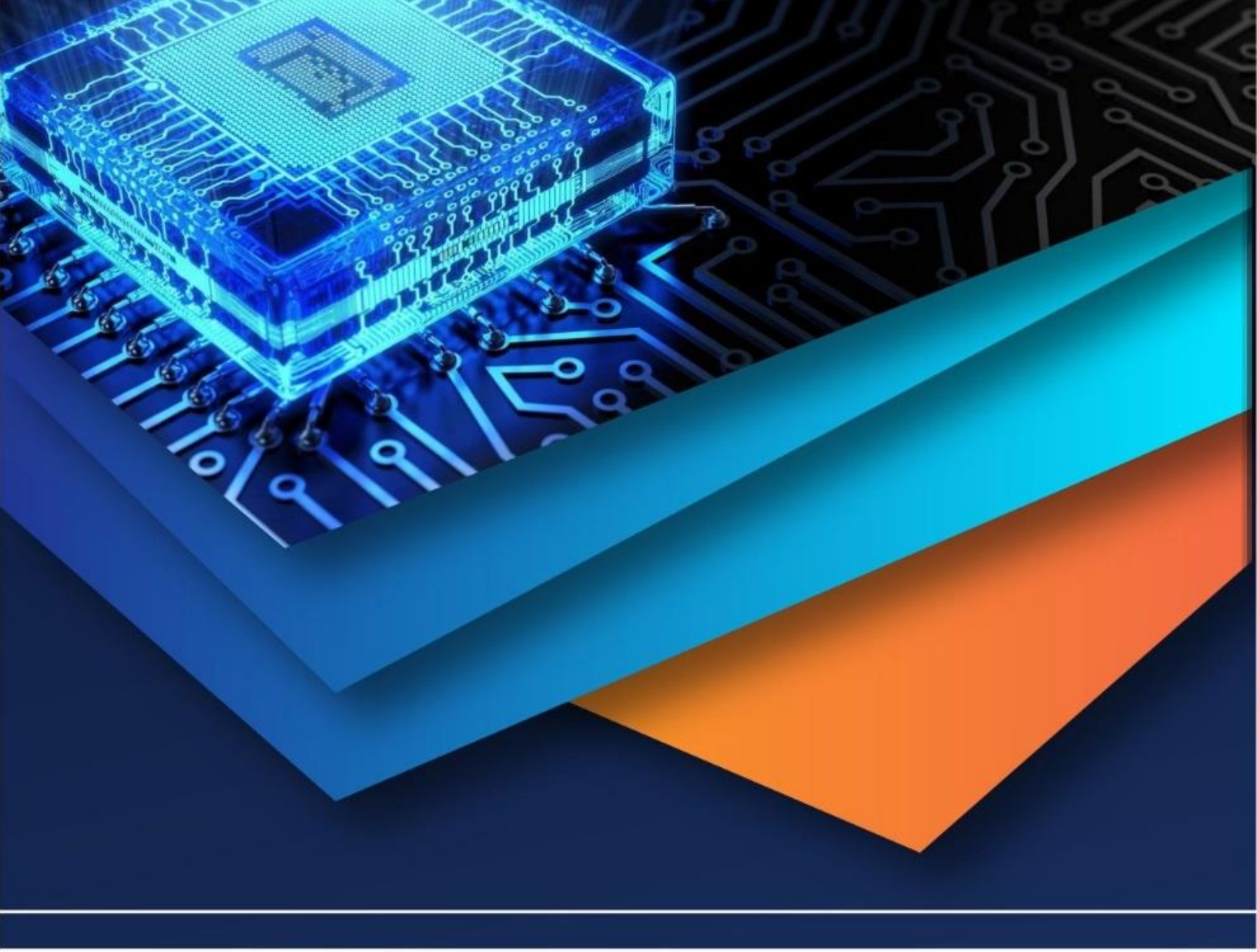
[8] P. Hassard and D. Kerr, "Predicting Football Match Outcomes Using Event Data and Machine Learning Algorithms," in Proc. 35th Irish Systems and Signals Conference (ISSC 2024), Derry/Londonderry, UK, Jun. 2024. doi: 10.1049/icp.2024.1567

[9] R. P. Bunker and F. Thabtah, "A Machine Learning Framework for Sport Result Prediction," Appl. Comput. Inform., vol. 15, no. 1, pp. 27–33, 2019. doi: 10.1016/j.aci.2017.09.005

[10] Z. Khan, M. Ali, and R. Patel, "Logistic Regression and Artificial Neural Networks for English Premier League Match Result Prediction," Int. J. Comput. Sci. Inf. Technol., vol. 16, no. 3, pp. 45–57, 2024.

[11] J. Štemberk, O. Přibyl, and V. Markovič, "Comparative Analysis of Machine Learning Methods for Football Match Outcome Prediction," in Proc. Int. Conf. Intelligent Systems and Applications (ISA 2023), Prague, Czech Republic, pp. 88–95, 2023.

[12] S. Almalki, A. Al-Harbi, and M. Al-Otaibi, "Deep Neural Networks for Football Match Outcome Prediction Using Historical Match Data," J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 8, p. 101724, 2023. doi: 10.1016/j.jksuci.2023.101724
[13] C. Yeung, R. Sit, and K. Fujii, "Evaluating Soccer Match Prediction Models: A Deep Learning Approach and Feature Optimization for Gradient-Boosted Trees," arXiv preprint arXiv:2309.14807, Sep. 2023.
[14] S. Vanithas, "Forecasting Premier League Match Outcomes for the 22/23 Season Using Deep Learning," Medium / Towards AI, 2023. [Online]. Available: https://towardsai.net
[15] S. Chakraborty, A. Mondal, A. Bhattacharjee, A. Mallick, R. Santra, S. Maity, and L. Dey, "Cricket Data Analytics: Forecasting T20 Match Winners Through Machine Learning," Int. J. Knowl.-Based Intell. Eng. Syst., vol. 28, no. 1, pp. 85–102, 2024. doi: 10.3233/KES-230060
[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, pp. 785–794, 2016. doi: 10.1145/2939672.2939785
[17] C. Yeung, R. Sit, and K. Fujii, "A Generalizable Machine Learning Approach for Match Outcome Prediction in Football," arXiv preprint arXiv:2505.01902, May 2025.
[18] M. J. Dixon and P. F. Pope, "The Value of Statistical Forecasts in the UK Association Football Betting Market," Int. J. Forecasting, vol. 20, no. 4, pp. 697–711, 2004. doi: 10.1016/j.ijforecast.2003.12.011
[19] O. Hubáček, G. Šourek, and F. Železný, "Exploiting Sports-Reference Data for Soccer Match Outcome Prediction," in Proc. ECML-PKDD Workshop on Machine Learning and Data Mining for Sports Analytics, Würzburg, Germany, 2019.
[20] D. Berrar, P. Lopes, and W. Dubitzky, "Incorporating Domain Knowledge in Machine Learning for Soccer Outcome Prediction," Mach. Learn., vol. 108, no. 1, pp. 97–126, 2019. doi: 10.1007/s10994-018-5747-8
[21] I. Wickramasinghe, "Applications of Machine Learning in Cricket: A Systematic Review," Mach. Learn. Appl., vol. 10, p. 100435, Dec. 2022. doi: 10.1016/j.mlwa.2022.100435
[22] P. N. Gour and M. F. Khan, "Ensemble-Based IPL Match Winner Prediction Using Multi-Model Machine Learning Approaches," Int. J. Adv. Res. Comput. Commun. Eng., vol. 14, no. 11, pp. 89–97, Nov. 2025.
[23] A. Singh and P. Kumar, "Real-Time Win Probability Estimation in T20 Cricket Using Gradient Boosting Models," Sports Analytics Review, vol. 4, no. 1, pp. 22–35, 2022.
[24] D. Shah and A. Sharma, "Effect of Run Rate, Required Run Rate, and Wickets on IPL Match Prediction Using Regression Analysis," Int. J. Sports Sci., vol. 11, no. 2, pp. 45–58, 2021.
[25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
[26] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in Proc. 34th Int. Conf. Machine Learning (ICML 2017), Sydney, Australia, pp. 3145–3153, 2017.
[27] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS 2017), vol. 30, pp. 4765–4774, 2017.
[28] J. M. Oliva-Lozano, M. Vidal, F. Yousefian, R. Cost, and T. J. Gabbett, "Predicting the Match Outcome in the 2023 FIFA Women's World Cup and Analysis of Influential Features," J. Hum. Kinet., vol. 93, pp. 45–58, 2025. doi: 10.5114/jhk/195563
[29] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.
[30] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324
[31] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[32] StatsBomb, "StatsBomb Open Data," GitHub, Dec. 2023. [Online]. Available: https://github.com/statsbomb/open-data
[33] football-data.org, "Football Data API," 2024. [Online]. Available: https://www.football-data.org
[34] cricsheet.org, "Cricsheet: Ball-by-Ball Cricket Data," 2024. [Online]. Available: https://cricsheet.org
[35] Kaggle, "European Soccer Database," Kaggle Datasets, 2016. [Online]. Available: https://www.kaggle.com/datasets/hugomathien/soccer
[36] RapidAPI, "Cricket Live Data API," RapidAPI Hub, 2024. [Online]. Available: https://rapidapi.com/hub/cricket
[37] Precedence Research, "Sports Analytics Market Size, Share, Growth Report, 2025–2034," Precedence Research, Ottawa, Canada, 2024.
[38] V. Patel and R. Ghosh, "Dynamic Win Prediction in Cricket Using Ball-by-Ball Data and Machine Learning Fusion Techniques," J. Predict. Model. Sports, vol. 6, no. 2, pp. 77–91, 2023.
[39] M. Waskom, "Seaborn: Statistical Data Visualization," J. Open Source Softw., vol. 6, no. 60, p. 3021, Apr. 2021. doi: 10.21105/joss.03021
[40] L. Biewald, "Experiment Tracking with Weights and Biases," Weights & Biases, 2020. [Online]. Available: https://wandb.ai
[41] S. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, Jan. 2020. doi: 10.1038/s42256-019-0138-9
[42] N. Kumar and A. Mishra, "Machine Learning Approaches for Cricket Winner Prediction Using Match Context Features," J. Data Analytics AI, vol. 5, no. 3, pp. 101–115, 2020.
[43] S. Jha and R. Verma, "Score Prediction and Win Probability Modeling in IPL Using XGBoost and Random Forest," in Proc. IEEE Sports Analytics Conference, 2022, pp. 34–41.
[44] A. A. Constantinou, "Dolores: A Model That Predicts Football Match Outcomes from All Over the World," Mach. Learn., vol. 108, no. 1, pp. 49–75, 2019. doi: 10.1007/s10994-018-5703-7
[45] K. A. S. Kaluarachchi and A. Aparna, "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket," in Proc. 5th Int. Conf. Information and Automation for Sustainability (ICIAfS 2010), Colombo, Sri Lanka, pp. 250–255, 2010. doi: 10.1109/ICIAFS.2010.5715681
[46] S. Salaboyn W. Wieckowski, and J. Watrobski, "Swimmer Assessment Model (SWAM): Expert System Supporting Sport Potential Measurement," IEEE Access, vol. 10, pp. 5051–5068, 2022. doi: 10.1109/ACCESS.2021.3140392
[47] F. Nasim, M. A. Yousaf, S. Masood, A. Jaffar, and M. Rashid, "Data-Driven Probabilistic Score Prediction for Batsman Performance in a Cricket Match," Intell. Autom. Soft Comput., vol. 36, no. 3, pp. 2965–2982, 2023. doi: 10.32604/iasc.2023.035401
[48] S. Chakraborty, L. Dey, A. Kairi, and S. Maity, "Prediction of Winning Team in Soccer Game: A Supervised Machine Learning-Based Approach," in Advances on Mathematical Modeling and Optimization with Its Applications, CRC Press, Taylor and Francis, 2023, pp. 145–162. ISBN: 9781032479613



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY