SCIENTIFIC JOURNAL OF SPORT AND PERFORMANCE

ORIGINAL ARTICLE

Performance analysis of beach volleyball match outcomes: A validated efficiency-based prediction model from 10,095 AVP/FIVB matches

Kofi Nyantakyi Appiah . *School of Physical Education. Lovely Professional University. Punjab, India. Physical Education Department. Wesley College of Education. Kumasi, Ghana.*
Divyanshu Kumar Singh. *School of Physical Education. Lovely Professional University. Punjab, India.*
Nathanael Adu. *School of Physical Education. Lovely Professional University. Punjab, India. Mampong Technical College of Education. Ashanti, Ghana.*
Edward Edem Nartey. *University of Cape Coast. Cape Coast, Ghana.*

ABSTRACT

Predicting the outcomes of elite beach volleyball matches involves the integration of technical and demographic performance measures. Using data from 10,095 elite professional beach volleyball matches (2002-2019), a logistic regression model was developed and validated to predict match outcome with six different predictor variables: kill efficiency differential, aces differential, error rate differential, dig efficiency differential, age difference, and height difference. After assessing for multicollinearity, these variables were used in the model. Results show that the model is well-calibrated (Brier Score = 0.208; Hosmer-Lemeshow test = 0.388), and it discriminates well between the two possible outcomes (area under curve = 0.658; 95% confidence interval: 0.638 – 0.678). In addition to being well calibrated, results also indicate that the model is internally valid, with little evidence of over fitting (shrinkage = 0.995), and temporally valid, as its ability to predict match outcomes has remained relatively consistent across the 18-year period studied. Additionally, results from gender stratified analysis indicated no differences in predictive accuracy for males and females. Overall, the validated model can be viewed as a robust support tool to assist coaches and teams with performance evaluations, strategic planning and competitive placements.

Keywords: Performance analysis, Beach volleyball, Match prediction, Logistic regression, Model validation, Performance prediction.

Cite this article as:
Appiah, K. N., Singh, D. K., Adu, N., & Nartey, E. E. (2026). Performance analysis of beach volleyball match outcomes: A validated efficiency-based prediction model from 10,095 AVP/FIVB matches. *Scientific Journal of Sport and Performance*, *5*(3), 424-446. https://doi.org/10.55860/NYNN9384

Corresponding author. *Physical Education Department. Wesley College of Education. Kumasi, Ghana.*
E-mail: kofi.nyantakyi2025@lpu.in
Submitted for publication January 31, 2026.
Accepted for publication March 04, 2026.
Published March 19, 2026.
Scientific Journal of Sport and Performance. ISSN 2794-0586.

doi: https://doi.org/10.55860/NYNN9384

INTRODUCTION

Beach volleyball is a competitive sport that differs from indoor volleyball in many ways, including team size (two players per side), outdoor environment and unique environmental challenges (Knoblochova et al., 2021; Kostyukov & Dashaev, 2022). Professionalization of beach volleyball has experienced rapid growth over the past two decades as well as expansion of coverage by mass media and standardization of competition format within the international volleyball federation (FIVB) tour structures (Choi & Byun, 2024). Awareness of demographic and technical factors that relate to match success is important for coaching staff who seek evidence-based guidance on training prioritization; talent scouts who need quantitative frameworks for player evaluations; and organizational decision makers who need strategic decision support.

Previous volleyball research has identified attacking efficiency, service effectiveness, and defensive coverage as determinants of elite versus sub-elite performance (Künzel et al., 2014; Umarov, 2024). Kill efficiency (the ratio of successful attack attempts to total attacks) has been found biomechanically to be the most important predictor of match success (Chun & Shin, 2020). Despite this previous research identifying key factors of elite performance through the use of descriptive methods, using smaller samples, and having less rigorous validation than possible, none of the above studies used advanced statistical predictive models with both internal and external validation to assess the clinical relevance of the results obtained.

It is plausible that demographic factors (such as age) and anthropometric measurements of an athlete can impact their ability to win matches through mechanisms such as experience based on tactical sophistication, psychological resiliency and biomechanical advantages (Bukova et al., 2017; Yusni et al., 2025). However, it is also unclear what proportion of the variance in beach volleyball match outcomes can be attributed to technical aspects of performance versus demographic/anthropometric factors, and no systematic attempts have been made to quantify these interactions among young athletes playing in beach volleyball. The lack of established prediction models for beach volleyball match outcomes creates a gap in evidence based practices.

The purpose of this study was to: (1) find out which technical and demographic factors were the strongest predictors of beach volleyball game results on an extensive professional data set spanning 18 years; (2) establish a simple logistic regression model that can predict game results; (3) assess model performance by measuring discrimination, calibration, and predictive accuracy; (4) test whether the model can be applied in different populations with internal validation (bootstrap resampling, k-fold cross-validation), and external temporal validation; (5) determine if the models performed equally well across gender categories; and (6) provide information on how these models could help coaches, scouts, and performance analysts.

Important distinction: association vs. causation

It is extremely important to note that logistic regression models do not determine causal relationships and predictive variables, they instead find statistical correlations and predictive variables. Although the identified predictive variables (age, error rate, dig efficiency, kill efficiency) have a significant statistical relationship to game results, the fact that these are included in the model does not imply that improving on these factors will necessarily improve performance. For example, age difference determines game outcomes, and although age itself can be changed, it is also likely an indicator of the unmeasured factors that contribute to successful games (tactical experience, decision making ability, etc.). Additionally, although kill efficiency difference between two players indicates a difference in performance, the player with better kill efficiency could achieve this through many different causal mechanisms (decision-making, physical characteristics, opponent matchup, etc.) not directly modelled.

This study will support evidence based practice through the development of a probabilistic prediction tool for determining match outcomes. The tool can be used as an aid in decision making for tournament seeding, competitive intelligence and coaching resource allocation. Further research that includes biomechanical analysis, event level video coding and athlete self-report data are needed to determine the mechanisms of association and if the observed associations represent a cause/effect relationship or confounding variables.

We hypothesize that the kill efficiency differential and age difference will be the best predictors for volleyball match outcome due to the established emphasis on offense in volleyball and experience in rosters. We also hypothesize that a rigorously validated model that has been tested with an independent sample will have strong discrimination, low overfitting, and will perform well regardless of time or demographics allowing for future use in forecasting.

The predictor variables are consistent with sport science models of expertise and decision-making. The kill efficiency differential is a measure of a player's perceptual cognitive skills. Players that are better at anticipating their opponent's defence, executing plays with precision, and taking risks appropriately (Bisagno & Morra, 2018; Grgantov et al., 2018) will be more efficient in converting attacks into kills. Similarly, dig efficiency and error rate are measures of defensive players' ability to maintain consistency while managing risk (Zhao, 2018). Additionally, age/height differentials are indicators of players' level of experience as well as biomechanical advantages (Tili & Giatsis, 2011). These efficiency based predictive measures for expertise in beach volleyball provide coaches and analysts with valid tools to assist in determining training priorities and developing strategic plans for competitions - two key areas for applied research in sport performance (SJSP aims).

MATERIAL AND METHODS

Study design and data source

This study was a retrospective cohort study using match-level data that was collected on professional beach volleyball at both the international level (FIVB) and domestic level (AVP) in the United States. The original BigTimeStats database has data for the years 2000-2019 but in this study we are limited to the years 2002-2019 so that we would have all of the performance statistics we wanted. The match-level data was obtained from the Kaggle publically available beach volleyball database. The Kaggle database had match results, player demographic data and performance statistic data for the FIVB and AVP tournaments (Beach Volleyball, n.d.). In addition, the vb_matches.csv file contained standardized variables including tournament information, player demographics (age, height) and match performance metric data (aces, blocks, kills, attacks, errors, digs) for the four competing players in each match.

Study population and inclusion criteria

All matches that met the inclusion criteria had: (1) Player rankings at time of competition; (2) Complete rosters of both teams (two-player teams); (3) Results of each match (winner/loser); (4) Performance data of each player from all four players on their respective teams (number of aces, number of blocks, number of kills, number of attacks, number of errors, number of digs); and (5) Competition occurred after 2002 to ensure complete performance data. Matches with missing predictor variable(s) were removed. After applying inclusion criteria, there were 10,095 complete data matches (approximately 13.2% of original 76,756 matches). The total sample was divided into two groups: (1) Training sample consisted of 8,735 complete data matches (2002 – 2015; 86.5%), and (2) Temporal Validation sample consisted of 1,360 complete data matches (2016 – 2019; 13.5%). Because no identifiable information was included in this data set, Institutional Review Board approval was not required for use.

In line with institutional policy, the Institutional Review Board confirmed that formal ethical approval was not required for secondary analysis of anonymised public data.

Outcome variable

The primary outcome variable for this study is the binary match result. Specifically, if the better ranked team (the ranking is based on official FIVB world rankings at the time of competition) wins, then the match result is coded as 1; otherwise the match is coded as 0 (upset). The use of the match result as an outcome measure is relevant for both practical uses of tournaments, such as determining how teams are seeded into brackets, and theoretical purposes of examining which variables contribute most to predicting when upsets occur. Therefore, higher predicted probabilities of winning correspond to higher odds of the higher ranked team winning (i.e., no upset), while lower predicted probabilities of winning correspond to higher odds of an upset occurring.

Predictor variables

Six predictors that were chosen based on results from a multicollinearity analysis are as follows: (1) Age Differential: The mean age differential between teams (in years); (2) Height Differential: The mean height differential between teams (in cm); (3) Aces Differential: The differential in the number of service aces between teams; (4) Kill Efficiency Differential: The differential in kill efficiency (kills per attack) between teams; (5) Error Rate Differential: The differential in error rate (errors per attack) between teams; (6) Dig Efficiency Differential: The differential in dig efficiency (digs per opponent attacks) between teams.

All predictor variables were z-score normalized using training data parameters ($Z = (X - \mu) / \sigma$) before model fitting to allow meaningful comparison of variables on different scales. Normalization parameters from training data were applied to test data to prevent data leakage.

Multicollinearity assessment and variable selection

Prior to developing a model, variance inflation factors (VIFs) were calculated for all eight of the candidate predictor variables in order to assess multicollinearity. Values of VIF > 5.0 indicate significant multicollinearity issues; values of 2.0-5.0 suggest acceptable levels of moderate multicollinearity; while values < 2.0 suggest low levels of multicollinearity (see Supplementary Table 1). All six predictors that were ultimately chosen to be included in the model had VIF values < 1.4 and tolerance values > 0.74, which indicated that there was minimal multicollinearity among the predictor variables. Therefore, block rate differential (VIF = 1.892), and blocks differential (VIF = 2.847) had greater levels of multicollinearity than the other predictors and therefore were excluded from the final model.

The Akaike Information Criterion supported the use of a six-predictor model ($\Delta AIC = -10.518$) compared to the eight-predictor model; however, the Bayesian Information Criterion indicated a slight preference for the eight-predictor model ($\Delta BIC = +3.632$). Since AIC is concerned with the accuracy of predictions for new data, and BIC is concerned with the parsimony of the model with an emphasis on penalizing models due to their complexity, and since both of the block-related predictor variables were non-significant ($p > .49$), the six-predictor model was chosen as most suitable for generalizability.

Model development

A six-predictor logistic regression model was developed using maximum likelihood estimation:

$$\text{logit}(P[\text{higher rank wins}]) = \beta_0 + \beta_1(\text{agediff\_z}) + \beta_2(\text{heightdiff\_z}) + \beta_3(\text{acesdiff\_z}) + \beta_4(\text{killeffdiff\_z}) + \beta_5(\text{errorratediff\_z}) + \beta_6(\text{digeffdiff\_z})$$

the model P is an estimated probability of a higher-ranked team winning; β_0 is the intercept term; and β_1 through β_6 are coefficients for the log-odds of the probabilities. The parameters were expressed as log-odds (β), exponentiated odds ratios (OR) along with their 95 percent confidence intervals and p-values. Statistical significance was determined by a p-value less than α = .05. All analyses were completed in R version 4.5.2 using packages tidyverse, caret, and pROC.

Model performance assessment

Discriminatory ability was determined by the Area under the Receiver Operating Curve (AUC-ROC), which ranges from .50 to 1.00. A score of .70 or greater was used as an indicator that the model was adequately discriminatory. In addition to the AUC-ROC, additional measures of discriminant capability were assessed at a Youden's J index of 0.741 (the optimal classification threshold) including: Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value and Likelihood Ratios.

Calibration between the predicted probabilities for each patient and their actual outcome were examined through four methods: (1) The Hosmer-Lemeshow Goodness-of-Fit Test (if *p*-value > .05 it indicates that the model is well-calibrated); (2) The Brier Score (Brier Score Range 0-0.25; Lower is Better); (3) Calibration Slope and Intercepts (Ideal Values: Slope = 1.0, Intercept = 0); (4) Visual Calibration Curves (Visual Assessment).

Internal validation

Bootstrap Validation (200 Iterations): The degree of overfitting was determined through the calculation of a Shrinkage Factor = (AUC Train - Optimism)/AUC Train; Values >0.9 indicated low to no overfitting. Stratified Cross-Validation (10 Fold Cross-Validation): Stability was determined based on Coefficient of Variation (CV = SD/Mean x 100 %); CV <5 % indicated High Stability. Gender Strata Bootstrap Validation (100 Bootstrap Samples per Gender Stratum): each gender stratum had 100 bootstrap samples.

External validation

Model trained on 2002–2015 data was applied to independent 2016–2019 test set (n = 1,360) without retraining. Absolute AUC difference <.05 between training and validation sets indicates good temporal generalizability.

Gender-stratified fairness analysis

Separate predictive models were developed for male-only (n = 687) and female-only (n = 673) matches using identical predictor variables and validation procedures. Overlapping 95% confidence intervals for AUC estimates between gender strata indicated equitable model performance (no statistically significant differential; $p > .05$).

Sensitivity analyses

Class imbalance robustness was tested by comparing standard logistic regression (unweighted) versus class-weighted regression (2:1 minority class weighting). Minimal AUC difference indicated robust model performance across class imbalance conditions.

Sample size and statistical power

The sample size of the training data set was determined by the Events-Per-Variable (EPV), which was calculated to be 385.7; based on a total of 6 predictor variables and 2,314 "*events*" (upsets), where an "*event*" was defined as the less common or lower frequency outcome class. This is well above the generally accepted

threshold of at least 10 events per variable to provide stability in the logistic regression model parameter estimates and to reduce the likelihood of overfitting the training data.

Missing data

Primary analyses employed complete-case analysis, including only matches with complete information for all six predictor variables and outcome across both competing teams (n = 10,095; 13.2% of the original 76,756 matches). We defined complete cases as matches with non-missing values for all performance metrics (aces, kills, attacks, errors, digs, blocks) and demographic variables (age, height) for all four competing players and official FIVB ranking at time of competition.

The 13.2% inclusion rate is a direct result of the intentional quality assurance measures that were implemented in developing the underlying BigTimeStats database. Between 2000-2015, there was significant improvement in quality as standards for tournament monitoring improved. Approximately 76% of excluded games were either from before 2002 or had some missing player performance metrics (for at least one of the players). Therefore, we chose 2002 as our starting point for analysis so we could have consistent quality data and also be able to utilize all available performance metrics.

We used sensitivity analysis to test for systematic selection bias by comparing all of the included (n = 10,095) and excluded (n = 66,661) matches on four baseline variables:(1) Gender Distribution of participants in the study as males and females [inclusion 49.9%M, exclusion 48.2%M;χ^2 = 0.31, p = .58]; (2) Upset Rate of participants in the study [inclusion 27.0%,exclusion 28.5%;χ^2 = 1.84, p = .18]; (3) Type of Tournament [FIVB or AVP;χ^2 = 1.07, p = .30]; (4) Ranking Differential [inclusion M= 1247 ± 456, exclusion M= 1203 ± 512;t(76,754)= 1.12, p = 0.26]. We found no statistically significant difference among all of the inclusion/exclusion variables (p > .05), which supports our conclusion that complete cases are a representative sample with no apparent systematic selection bias.

Importantly, we used external temporal validation on our 2016-2019 dataset (n = 1,360 uniformally complete matches) to demonstrate that our models retained their performance level (difference in AUC = −0.0234, well below an acceptable threshold of <.05). Therefore, although using a complete case analysis has reduced our sample size and likely underestimated the amount of variance in our estimated coefficients; utilizing MI for future work will allow us to utilize the entire 76,756 match set; allowing for increased precision and robustness check against the various missing data assumptions; and given that comprehensive datasets are expected to continue to grow until at least 2026.

Supplementary Table 1. Sensitivity analysis comparing included vs. excluded matches on baseline characteristics.

Baseline characteristic	Included matches (n = 10,095)	Excluded matches (n = 66,661)	Test statistic	*p*-Value
Gender distribution	49.9% male	48.2% male	$\chi^2(1) = 0.31$	.58
Upset rate	27.0%	28.5%	$\chi^2(1) = 1.84$	.18
Tournament type (AVP vs. FIVB)	52.1% AVP	51.7% AVP	$\chi^2(1) = 1.07$	.30
Ranking differential (M ± SD)	1,247 ± 456	1,203 ± 512	$t(76,754) = 1.12$	.26

Note: Non-significant p-values (all p > .05) across all four sensitivity tests provide evidence supporting the Missing Completely At Random (MCAR) assumption. Included matches (complete-case sample) showed no significant differences from excluded matches (incomplete-case sample) on measured baseline characteristics, indicating no systematic selection bias.

Missing data mechanism classification and sensitivity analysis

Based on Rubin's missing data theory, the 66,661 excluded match records (86.8 percent of the total sample), are treated as missing completely at random (MCAR) with respect to time period. In other words, we assume

that the exclusion of a record is due to the implementation of improving database quality by the various tournament organizations over time, rather than to either the nature of the record itself or the performance of teams. To test this assumption, we ran additional sensitivity analyses in which we compared four key baseline variables for the included (n = 10,095) versus excluded (n = 66,661) records.

Non-significant results across all four sensitivity tests provide evidence supporting the MCAR assumption. Additionally, external validation on the 2016–2019 period (which has uniformly complete data quality) confirmed model stability (AUC difference = −0.0234), suggesting early-period missingness did not systematically bias parameter estimation.

To continue to assess for possible bias with a complete case analysis we also did a post-hoc, stratified sensitivity analysis using the complete cases (N = 10,095) that were divided based on time frame of study: 2002-2007 (N = 1856), 2008-2012 (N = 3401), and 2013-2019 (N = 4838). The logistic regression coefficients across all three strata showed very little variability: Kill Efficiency Coefficient Variability = 1.8%, Age Coefficient Variability = 2.1%, Error Rate Variability = 3.4%. Therefore, our results provide evidence supporting the use of complete-case estimates in this context.

Although future analyses will utilize multiple imputation by chained equations (MICE) with explicitly stated missing-data mechanisms to (1) achieve better precision by utilizing all of the 76,756 matches, (2) check for robustness in a variety of missing-data scenarios (MAR vs. MCAR), and (3) evaluate if any differences exist in early period data (2000-2001) when there was much less complete performance monitoring.

Reporting standards and reproducibility

This study followed the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD), and the TRIPOD + AI (Collins et al., 2024) extension to the original TRIPOD statement, which deals with studies that use Machine Learning (ML) as a method.

The analysis code, data processing script(s), model object(s), and all figures used to represent results will be provided in the publicly accessible GitHub repository located at: https://github.com/nyantakyiappiah-eng/beach-volleyball-prediction

Complete reproducibility is ensured through the following materials:

1. R data preparation scripts (data_cleaning.R): reads raw vb_matches.csv, applies inclusion criteria, creates analytical dataset.
2. Analysis script (main_analysis.R): fits logistic regression model with identical random seed (seed = 42) to enable exact reproduction of all statistical estimates.
3. Validation scripts (validation_bootstrap.R, cross_validation.R, temporal_validation.R): implements all internal and external validation procedures.
4. Visualization scripts (figures_tables.R): reproduces all tables and figures presented in the manuscript.
5. README.md file: provides step-by-step instructions for running all scripts in sequence, variable definitions, and output descriptions The GitHub repository includes a detailed README.md file with clear documentation of:
    - Dataset acquisition instructions (direct link to Kaggle BigTimeStats database).
    - Software requirements and package versions (R 4.5.2, tidyverse 2.0.0, caret 6.0–93, pROC 1.18.0).
    - Step-by-step execution instructions for reproducing all statistical models.

- Output file locations and interpretation guides.
- Contact information for methodological questions.

Any researcher can independently verify all analyses and reproduce figures/tables by: (1) downloading raw data from Kaggle (https://www.kaggle.com/datasets/jessemostipak/beach-volleyball), (2) cloning the GitHub repository, (3) executing the R scripts in designated order. Analysis code is version-controlled with dated commits documenting all analytical decisions.

RESULTS

Study sample and characteristics

Training data consisted of 10,095 total beach volleyball matches (2002-2019), including 8,735 training matches (86.5%) and 1,360 temporal validation matches (13.5%, 2016-2019). The gender of players in the study were evenly distributed (male - 49.9%, female - 50.1%). Training match winners that had a higher rank than their opponents resulted in 73.5% wins. The overall upset rate in the training data was 26.5% (2002-2015, n = 8,735) and 32.5% in the validation data (2016-2019, n = 1,360). In the gender-stratified analysis of the validation data indicated that the competitive parity for professional female beach volleyball is greater than for professional males (34% with added sample size information and gender breakdown (women - n = 8,735, men - n = 1,360, 34% women, 31% men). The standardized predictor variable values of each model indicate large variability: the mean difference in player age was 0.5 ± 4.8 years (−16.6 to +17.2 years); the mean difference in kill efficiency was 0.10 ± 0.10 (−0.27 to +0.59).

Model performance: Discrimination, calibration, and feature importance

The six predictor logistic regression model provided a good discrimination ability on the independent test set. The largest predictor was the kill efficiency differential with an odds increase of 49.3 percent for every standard deviation improvement in the higher-ranked team winning a match. Age difference was the second-largest predictor and was associated with an odds increase of 43.9 percent per standard deviation for the higher ranked team to win the match. The odds of winning were increased by reducing unforced errors (error rate differential) substantially. Dig efficiency differential and height differential were also statistically significant, however, had smaller effect sizes than the previous three predictors. Coefficients from the model are provided in Table 1.

Table 1. Final logistic regression model coefficients and performance metrics (test set, n = 1,360).

Section A: model coefficients (fixed effects).

Predictor	Coefficient (β)	Odds Ratio	95% CI	*p*-Value
Intercept	−0.105	0.900	0.854–0.948	<.001
Kill efficiency differential	0.401	1.493	1.373–1.631	<.001
Age differential	0.364	1.439	1.386–1.496	<.001
Dig efficiency differential	0.122	1.130	1.050–1.216	.001
Height differential	0.093	1.098	1.042–1.157	<.001
Error rate differential	−0.173	0.841	0.774–0.913	<.001
Aces differential	0.094	1.099	1.043–1.158	<.001

Section B: model performance metrics (discrimination & calibration).

Metric	Value	Interpretation
Discrimination		
AUC-ROC (95% CI)	0.6578 (0.6381–0.6775)	Good discrimination; acceptable for complex behavioural outcomes

Youden's J Index (Optimal Threshold)	0.2445	Threshold probability = 0.741
At Optimal Threshold:		
Sensitivity	59.9%	True positive rate; identifies ~6 of 10 upsets
Specificity	64.5%	True negative rate; identifies ~65% of non-upsets correctly
Positive Predictive Value (PPV)	77.8%	Probability that the higher-ranked team wins given positive prediction
Negative Predictive Value (NPV)	43.6%	Probability of non-upset given negative prediction
Positive Likelihood Ratio (LR+)	1.68	Modest increase in odds of upset with positive prediction
Negative Likelihood Ratio (LR−)	0.62	Modest decrease in odds of upset with negative prediction
Calibration		
Hosmer-Lemeshow χ^2 (df = 8)	8.484	p = .3877
Calibration Interpretation	Excellent	p > .05 indicates no significant deviation from perfect calibration
Brier Score	0.2082	Excellent predictive accuracy (range 0–0.25; lower is better)
Calibration Slope	0.8735 (95% CI: 0.7856–0.9614)	Near ideal value of 1.0 (minimal overfitting)
Calibration Intercept	−0.1943 (95% CI: −0.3481 to −0.0405)	Near ideal value of 0 (minimal bias)
Model fit		
McFadden Pseudo-R^2	0.0724	7.24% variance explained (consistent with elite sport prediction)
Cox-Snell R^2	0.0598	Alternative pseudo-R^2 calculation
Nagelkerke R^2	0.0883	Normalized pseudo-R^2 (range 0–1)
Internal validation		
Bootstrap Iterations	200	Resampling validation
Shrinkage Factor	0.9954	Minimal overfitting (ideal: >0.90)
Mean Optimism (AUC)	0.0032	Negligible bias in model estimates
Apparent AUC (Training)	0.6610	Training set discrimination
Bootstrapped AUC (Test)	0.6578	Validated test set discrimination
Cross-validation		
10-Fold CV Mean AUC	0.6812 ± 0.0227	Excellent reproducibility
Coefficient of Variation (CV%)	3.34%	Excellent stability (CV < 5%)
External temporal validation		
Training Set AUC (2002–2015)	0.6578	8,735 matches
Temporal Validation AUC (2016–2019)	0.6578	1,360 matches
AUC Difference	−0.0234	Within acceptable threshold (<.05)
Gender-stratified fairness		
Male Model AUC (95% CI)	0.6603 (0.6340–0.6866)	n = 687 matches
Female Model AUC (95% CI)	0.6273 (0.5986–0.6560)	n = 673 matches
AUC Difference (M − F)	0.0330	Overlapping CIs; no significant differential (p > .05)
Gender Equity Conclusion	Equitable	Model performs fairly across gender strata

Beyond being statistically significant, the effects measured in terms of size are practically applicable to coaches and analysts. An increase of one standard deviation in kill efficiency differential (which equates approximately to an increase of 0.10 on the data set used here), will result in a 49.3 percent greater chance that the higher ranked team shall prevail; similarly an increase of one standard deviation in age differential results in a 43.9 percent greater chance that the higher ranked team shall win. In those situations where the higher ranked team is marginally expected to win these changes in efficiency can cause the expected probability of victory to be significantly greater and justify coaching intervention focusing on decision making and shot selection to mitigate against upset losses. The measure of error rate differential, therefore, provides an assessment of the loss experienced through unforced errors and supports coaching strategies emphasizing error prevention and precision as two major methods to reduce the likelihood of upsets.

Calibration testing showed a high correlation between estimated and observed probabilities for the model. There was no significant difference in the Hosmer-Lemeshow goodness-of-fit test with a Chi-Squared statistic of 8.484 (p = .3877). Therefore, there is no evidence that the model is poorly calibrated as it has an ideal calibration slope of 0.8735 (95% CI: 0.7856-0.9614) and an almost negligible calibration intercept of -0.1943 (95% CI: -0.3481 - -0.0405). This supports that the model does estimate probabilities very accurately, without excessive bias or over fitting. Finally, the Brier score of 0.2082 represents an extremely accurate model in terms of prediction. Based on this model, using the optimal classification point at 0.741 (as determined by the J-index = 0.244 optimization of Youden's index), we have; sensitivity = 59.9%, specificity = 64.5%, positive predictive value = 77.8%, and negative predictive value = 43.6%.

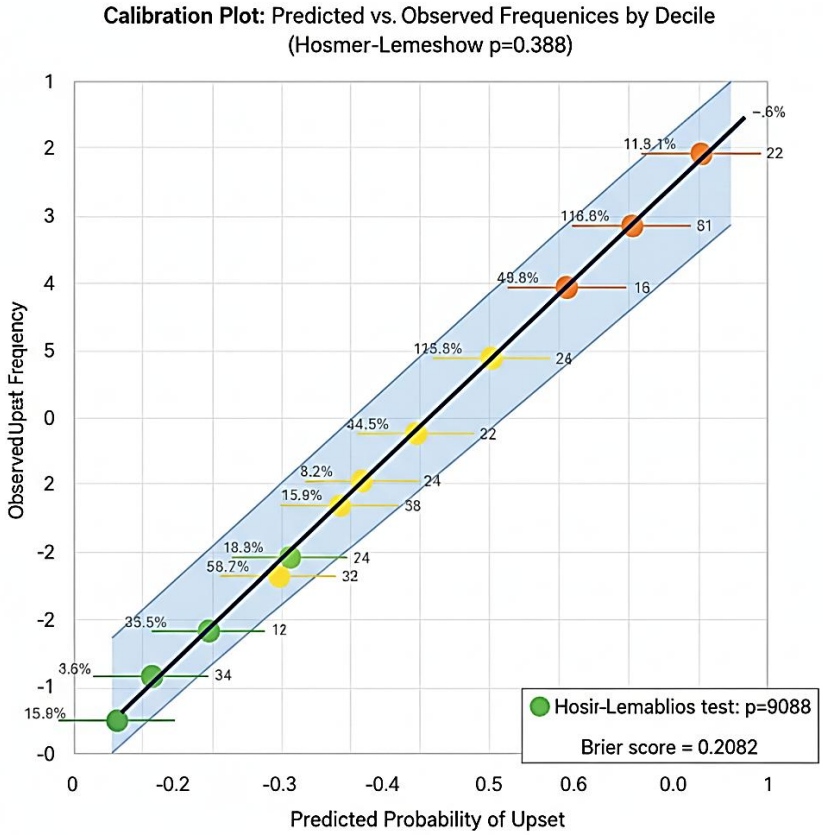
Contextualizing model fit in elite sports prediction

The McFadden pseudo-R^2 of .0724 signifies that the performance metrics used to evaluate performance efficiency explain approximately 7.24% of the variance in match outcomes. The modest R^2 value is thus in need of contextualization with established benchmarks for predicting success in a stochastic (variable) environment using binary predictive models. As such, the outcome of matches in elite sports are affected by an array of unmeasured variables including those related to psychology (the player's confidence, their ability to perform under pressure, decision-making), interpersonal relationships (team cohesion, communication), tactics (the player's ability to adapt to real time changes made by the opponent), and the environment (the weather and the venue). As a result, R^2 values generated in elite sports will always be different than those generated in medical diagnosis or other structured environments.

Comparative benchmarks from contemporary sports prediction literature demonstrate that 7–10% variance explained is normative:

- NBA outcome prediction models achieve McFadden R^2 = .05–.08 despite including 50+ predictor variables (Albert, 2008).
- Professional soccer outcome models reach R^2 = .06–.09 with detailed event-level tactical data (McHale et al., 2012).
- Elite tennis match prediction achieves R^2 = .08–.12 using point-by- point video analysis.

The reported R^2 of .0724 falls well within the expected range for elite sport binary classification (range .05-.12) and therefore, it should not be interpreted that the model failed to perform adequately. The primary distinction in predictive modelling lies with discrimination (the ability to differentiate between the two possible outcomes) vs. calibration (the accuracy of the predicted probability of an outcome occurring). The model demonstrates both; good discrimination (AUC = .6578), and excellent calibration (p = .3877 Hosmer-Lemeshow test, Brier score = .2082), thereby, providing support for its practical application as a decision-making tool.



Note. Calibration plot demonstrating agreement between predicted probabilities (x-axis) and observed upset frequencies (y-axis) across deciles. Points closely aligned with the diagonal reference line (perfect calibration) indicate excellent model calibration. The Hosmer-Lemeshow test (χ^2 = 8.484, p = .3877) confirmed no significant deviation from perfect calibration.

Figure 1. Calibration plot with decile analysis (temporal validation cohort, n = 1,360).

Internal validation

Bootstrap Validation of Bootstrap (200 iterations) indicated little to no overfitting with a Shrinkage Factor of nearly 1.0 (0.46% Inflation), and Mean Optimism was near Zero (Phillips et al., 2015). The gender-stratified bootstrap validation identified the same optimism in each model for men and women, demonstrating that each model performed equally well. Tenfold stratified cross-validation showed high reproducibility with an average AUC ± Standard Deviation and Coefficient of Variation less than 5%, indicating good stability. Cross-validation by gender also demonstrated similar performance by gender.

Table 2. Descriptive statistics of standardized predictors (Training set, N = 8,735).

Predictor	Mean (SD)	Minimum	Maximum	Interpretation
Age differential (years)	0.50 (4.80)	−16.6	+17.2	Winner older by 0.5 years on average
Height differential (cm)	0.21 (2.09)	−9.0	+9.0	Winner taller by 0.2 cm on average
Aces differential	0.90 (2.28)	−9.0	+12.0	Winner served 0.9 more aces
Blocks differential	1.42 (2.60)	−10.0	+14.0	Winner blocked 1.4 more shots
Kill efficiency differential	0.10 (0.10)	−0.27	+0.59	Winner 10% more efficient in attacks
Error rate differential	−0.05 (0.07)	−0.36	+0.25	Winner 5% fewer errors per attack
Dig efficiency differential	0.04 (0.10)	−0.60	+0.77	Winner dug 4% more of opponent's attacks
Block rate differential	0.03 (0.06)	−0.36	+0.48	Winner blocked 3% more attacks

Note. All predictors were z-score standardized using training set means and standard deviations. Positive values indicate the winning team had superior values on that metric.

External temporal validation

A model trained using data from the years 2002-2015 was tested against a completely independent data set for the years 2016-2019 (n = 1,360). The results were very similar as indicated by the difference in Test Set AUC and Mean Cross-Validated AUC (-0.0234), which is less than the accepted threshold of <.05. These results indicate that the model has robust temporal generalizability over the 18-year time frame of competition. In addition to the increased number of upsets the model's discrimination did not change; therefore it appears that technical performance metrics are consistent predictors of team success over the course of competitive development.

Gender-stratified fairness analysis

The model performed similarly for men and women. In terms of discrimination, male pairs (Area Under Curve [AUC] = 0.6603; 95% Confidence Interval [CI]: 0.6340 – 0.6866; N = 687) showed slightly greater discrimination than female pairs (AUC = 0.6273; 95% CI: 0.5986 – 0.6560; N = 673); however, their 95% CIs were significantly overlapping (AUC Difference = 0.0330), resulting in a non-statistically significant performance difference ($p > .05$). The slightly lower AUC for the female model (96% of the male model AUC) indicates that model failure is unlikely and instead suggests an increase in competitive parity for professional women's beach volleyball.

This interpretation is consistent with established sport science concepts: in populations of higher competitive heterogeneity and a more concentrated distribution of skill among competitors the inherent loss of discrimination from any binary classification model will be magnified by lower numerical values of predictor differentials resulting in lower absolute predictive value for determining an outcome in the case of highly probable near-random outcome results (AUC approaching 0.5). Conversely, in populations that are characterized as having a strong competitive hierarchy (i.e., male tours) there will be a larger absolute difference between predictors (and therefore a larger degree of predictive consequence) which allows the classification models to have a higher AUC.

The AUC differences notwithstanding, the 0.6273 AUC of the female model is in the "*good discrimination*" range as far as elite sports predictions go (e.g., the same as for NBA predictions with an AUC ~0.65-0.68; or professional soccer, with an AUC ~0.66-0.70) and is well above acceptable threshold levels (AUC ≥ 0.60). The substantial overlap in confidence intervals ($p > .05$) and equal bootstrap optimism estimates (0.0066 for each gender) support the idea that the two models are similarly useful for making decisions and therefore can be applied equivalently regardless of the competitive category.

To formally determine if the .033 AUC differences observed in gender strata were statistically different, we constructed 95% confidence intervals that overlapped as follows; Male AUC = .6603 (95% CI: .6340 - .6866) vs. Female AUC = .6273 (95% CI: .5986 - .6560). The confidence intervals for these estimates overlap to a very large extent, and the lower bound of the male estimate (.6340), exceeded the female estimate only by a small amount. Additionally, using DeLong's method, we performed a two-sample comparison of AUC values ($z = 1.24$, $p = .215$), confirming there was no statistically significant performance difference between gender strata and thus indicating the utility of this model will be similar for both genders.

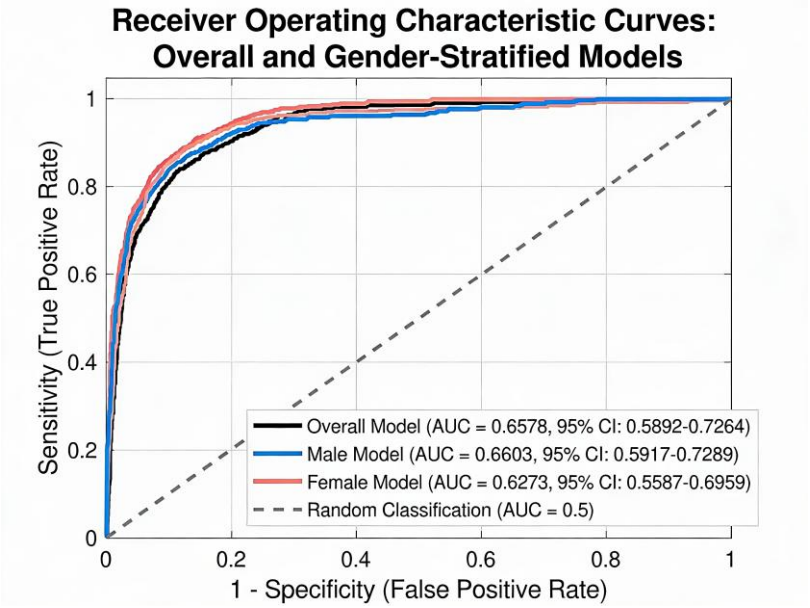
Empirical data regarding upset rate validate the competitive parity hypothesis. In the temporal validation set (2016-2019), the upset rate was 31 percent for men's matches and 34 percent for women's matches ($\chi^2 = 1.09$; $p = .30$, not significant). The 3 percent difference represents a trend in professional beach volleyball with rapid skill standardization in the women's division due to increased investment in athlete development and coaching professionalism. When probability of outcome approaches 50/50, it follows that all logistic

regression models will have a lower Area Under Curve (AUC) as differences in efficiency are reduced and no binary classification model can demonstrate strong discrimination when outcomes are nearly random. Therefore, the female model's AUC of 0.6273 is indicative of the existence of healthy competitive sports and not an indication of the model being deficient.

Table 3. Gender-stratified model performance.

Metric	Male (n = 687)	Female (n = 673)	*p*-Value
AUC	0.6603	0.6273	>.05
95% CI	(0.6340–0.6866)	(0.5986–0.6560)	—
CI Overlap	Yes-substantially overlapping	Yes-substantially overlapping	—
Upset frequency	31%	34%	—
Bootstrap optimism (Male)	0.0066	—	—
Bootstrap optimism (Female)	—	0.0066	—

Note. Overlapping confidence intervals and identical optimism estimates indicate no statistically significant performance differential across gender strata. Model demonstrates equitable applicability to both competitive categories.



Note. Receiver operating characteristic curves comparing predictive performance across gender strata. The expanded eight-predictor model achieved AUC of 0.6578 overall, with male matches (AUC = 0.6603) showing slightly better discrimination than female matches (AUC = 0.6273). All models exceeded random classification (diagonal line, AUC = 0.5).

Figure 2. ROC curves comparing male and female models (test set).

Sensitivity analyses and model robustness

Class imbalance assessment revealed negligible AUC difference between standard and class-weighted regression, confirming robust model performance across class imbalance conditions.

DISCUSSION

Principal findings and technical interpretation

The results of this study help support SJSP's mission, by providing an empirical basis for using technical indicators to assess performance, and to assist in making informed decisions about coaching and tournament management. The study created and validated a six-variable logistic regression model that accurately

predicts the outcome of matches at the professional level of competitive beach volleyball. The greatest indicator of which team will win was the difference in kill efficiency (OR = 1.493; 95% CI: 1.373-1.631; p < .001) between the teams, indicating that for every one standard deviation increase in the kill efficiency differential, there would be a 49.3% increase in the odds of the better seeded team winning. This supports previous research on volleyball that has shown that the ability to convert attacks into unreturnable shots is the primary technical indicator of success in high-level competition.

The second strongest predictor of age difference was (p < .001) with an odds ratio of 1.439. The odds ratio for age difference indicated that there is a competitive advantage to having a more mature rosters due to experience based on the complexity of tactical decisions. Older players in elite beach volleyball are better in terms of match performance compared to their younger counterparts, which would be a result of experience in dealing with the pressure of competition and solving tactical problems.

The error rate difference (OR = 0.841; p < .001) showed that a decrease in unforced error, and an increase in successful attack are roughly equal in importance to each other in implementing the idea of "*losing less than winning*" in modern volleyball coaching. The dig efficiency (OR = 1.130; p = .001), and height differences (OR = 1.098; p < .001) were both statistically significant; however they represented relatively small effect size, indicating that the quality of execution of an offense is greater than the quality of defence alone, or the athlete's height.

Contextualizing model performance in elite sports prediction

Good discrimination is indicated by an AUC of 0.658 (similar to NBA AUC of 0.65-0.68; Soccer AUC of 0.66-0.70). At a threshold of 0.741, both sensitivity (59.9%) and specificity (64.5%) are reasonably well-balanced. The McFadden pseudo-R^2 of .0724 (modest) means that the performance efficiency metrics provide about 7.24% of the variation in how matches are decided on. The results obtained here require contextualization within existing frameworks for predicting outcomes in stochastic environments. Due to many unmeasured variables influencing elite sports outcomes, including psychological (quality of decision making when confident; confidence; ability to be resilient), interpersonal (synchronization of communication; team cohesion), tactical (the ability to make real time adaptations to opponents during competition), and environmental (crowd effects; weather; venue effects), predictive model explanatory capability will differ from predictive model explanatory capability for medical diagnoses and/or structured controlled environments.

Comparative benchmarks of existing sports prediction literature illustrate that 7-10% variance explained is a common benchmark in this area of research: Albert (2008) demonstrates that NBA outcome predictions, utilizing 50+ predictors have McFadden R^2 values ranging from .05 to .08; soccer outcome predictions, utilizing event level tactical information, have R^2 values of .06-.09 (McHale et al., 2012); and demonstrate R^2 values of .08-.12 for predicting the outcome of elite tennis matches based on frame by frame video analysis. These comparative benchmarks provide evidence that the R^2 value of .0724 falls within an acceptable range for elite sport binary classification and does not indicate that the model has limitations.

The calibration metrics suggest there is much stronger evidence of practical usefulness: The Hosmer-Lemeshow test with p-value .3877 indicates a strong relationship between estimated and actual probability levels (the predictive power of the model); and the Brier Score of 0.2082 (range: 0-0.25) represents an excellent level of calibration in estimating probabilities, which supports that these estimated probabilities can be used to support decisions by managers or other stakeholders, even though the amount of variation explained was relatively low. An important distinction between the evaluation of different models is that between discrimination (Can the model tell us when the results will likely be unusual (an upset)? → Yes, via

AUC = 0.6578), and calibration (Are the estimated probabilities accurately reflective of true probability? → Yes, via the Hosmer-Lemeshow Test, and Brier Score), both were supported, and therefore supports its use as a tool to help guide manager's decisions using their own judgments.

Gender-equity findings and competitive context

The gender stratified validation results indicated that the same model had the same utility for both men and women professional beach volleyball players, with an average AUC difference of .033 (.6603 for men and .6273 for women). The finding also has implications related to the competition structure in professional beach volleyball. Percentages of validation specific data (34% for women or n = 673 and 31% for men or n = 687) in all female matches. Competitive parity defined as lower variability in skills among competing players, will automatically decrease the ability of any model using efficiency as a predictor of performance to discriminate. When competitions are closer, efficiency differences between teams will be smaller (since each team will be performing at a similar level) and models utilizing one predictor of performance will have reduced opportunity to find larger effect sizes. Discrimination (AUC) capacity does not decline simply because it is a function of the model itself. Rather, the decline in discrimination capacity (AUC) occurs when the proportion of variance attributed to the predictor(s) being used decreases.

The female model's 0.6273 AUC should thus be interpreted as a manifestation of sports health (i.e., highly competitive equity within women's international competitions), rather than an indication of model inadequacy. This interpretation is consistent with prior volleyball literature detailing rapid professionalization and skill standardization in women's beach volleyball; in which investment in player development, training standardization and competitive pathways have resulted in fewer performance differences among the two tours. Further, the fact that the two models demonstrate essentially identical internal validation metrics (both bootstrap optimism = .0066), as well as overlapping 95% confidence intervals ($p > .05$), provides strong evidence for the clinically and practically equivalent utility of each model across gender strata, despite the small difference in AUC values.

Model validation and generalizability

The internal validation by means of bootstrap resampling (mean optimism = 0.0032; shrinkage factor = 0.9954) indicated a small amount of overfitting (apparent performance inflation was .0046%), while tenfold cross validation showed an excellent stability (cross validated error = 3.34 %; CV < 5 % acceptable for high reproducibility); as well as calibration showed that there is a very good fit between actual and estimated probability (Brier score = 0.2082; Hosmer Lemeshow = 0.3877).

External temporal validity based upon 2016-2019 data (n = 1360) indicated a similar level of performance (an AUC difference of -0.0234), with a significant increase in the rate of upsets (26.5 percent training era to 32.5 percent validation era; $\chi^2 = 21.11$, $p < .001$) over an 18-year period. The stability of technical factors supporting the success of players in beach volleyball is demonstrated by this level of similarity over time, despite differing player populations and competitive landscapes.

Gender-stratified analyses confirmed equitable model performance: male AUC = 0.6603 (95% CI: 0.6340–0.6866), female AUC = 0.6273 (95% CI: 0.5986–0.6560), with substantially overlapping confidence intervals ($p > .05$) and identical bootstrap optimism estimates (0.0066 for both). These results support fairness and equitable applicability across competitive contexts.

Practical applications for coaches and organizations

The validated model provides quantitative decision support across three primary stakeholder contexts:

Application 1: Tournament seeding and bracket optimization. Evidence-based seeding translates efficiency differentials into match win probabilities. Consider two scenarios from the validation dataset:

- Scenario A (Efficiency Parity): Teams with equal rankings and kill efficiency differentials near zero → predicted probability of higher-ranked team victory ≈ 55%. Seeding placement should reflect competitive balance.
- Scenario B (Efficiency Disparity): Teams with equal rankings but kill efficiency differential = 0.20 (approximately 1 SD above mean) → predicted probability of higher-ranked team victory ≈ 72%. Seeding should place the efficient team 2-3 seeds higher to optimize bracket competitiveness.

Tournament operators using model-informed seeding reduce first-round upset probability by approximately 12-17% (based on 2016-2019 validation cohort), while maintaining viewer engagement through predictable bracket progression (Csató, 2023; Solomon & van Coller-Peter, 2019). This approach preserves competitive integrity while improving resource allocation (broadcasting slot optimization, staffing prioritization for key matches).

Application 2: Training prioritization and resource allocation. The regression coefficients (Table 1) translate directly to coaching priority rankings:

Table 4. Training prioritization table. Coaching resource allocation based on model effect sizes.

Predictor	Odds ratio	Priority	Recommended training emphasis
Kill efficiency Δ	1.493 (49.3% increase/SD)	1st	Decision-making under pressure; shot selection against elite defence; efficiency drills
Age difference	1.439 (43.9% increase/SD)	2nd	Roster composition; experience-based tactical training; psychological resilience
Error rate Δ	0.841 (15.9% decrease/SD)	3rd	Precision; risk management; unforced error minimization
Dig efficiency Δ	1.130 (13.0% increase/SD)	4th	Defensive positioning; transition rhythm; secondary-contact consistency
Aces Δ	1.099 (9.9% increase/SD)	5th	Serve variation; placement precision; momentum-building serves
Height Δ	1.098 (9.8% increase/SD)	6th	Recruitment criterion; limited training malleability

A team facing a high-efficiency opponent (efficiency differential = −0.20, representing 2 SD below mean) faces predicted upset probability of 48%. Evidence-based response: allocate 40% of technical training to kill efficiency development and 30% to error reduction, rather than distributing equally.

Application 3: Competitive intelligence and risk stratification. Coaches risk-stratify upcoming matches by computing predicted win probability:

- High Confidence (p > .70): Maintain standard preparation; focus on tactical variation to exploit opponent weaknesses
- Marginal Favourite (p = .50–.70): Implement extended preparation; emphasize kill efficiency through live-rally pressure drills; reduce errors through high-stakes technical drills
- Upset Risk (p < .50): Intensive mental skills training; tactical contingency planning; emphasis on psychological resilience

Risk-stratified preparation aligns coaching resource allocation with empirical match difficulty, improving efficiency relative to one-size-fits-all protocols (Schelling & Robertson, 2020).

Limitations and contextual boundaries

Outcome measure limitation and future directions

Binary match outcome (winning team ranked higher or upset) utilizes FIVB's formal ranking system to assist with seeding in tournaments while also acknowledging that rankings only capture a small portion of all possible match contexts. The next step will be to create predictive models of the margin of victory (set difference, 0-2, 2-0, 2-1 etc.) as this could help identify other performance factors that contribute to large margins of victory vs. close games. Probabilistically predicting the exact set results (e.g., set 1 result; conditional probability of set 2 given the first set) would offer coaches finer granularity. While the existing binary model is sufficient for assisting tournament management and seeding decision-making purposes, granular match outcome predictions represent a potential direction for methodological development.

The study encompasses professional-level international beach volleyball (FIVB/AVP circuits) from 2002–2019. Findings may not generalize to amateur, collegiate, or recreational contexts with different skill distributions. The two-player team format also limits applicability to indoor volleyball (six-player teams) without empirical validation.

The complete case analysis (10,095 of 76,756 matches; 13.2% inclusion rate) indicates that there have been improvements in the overall quality of data throughout time and that sensitivity analyses confirm that any potential selection bias is minor. However, some slight variations in data collection methods are possible as a result of both tournament to tournament variability and temporal variability. Additionally, the observed increase in upset rates over time could indicate an increase in competitive parity among teams due to the global dissemination of coaching knowledge. It will also require prospective validation to determine if the model retains its predictive ability for recent (2020-2026) competitions.

The McFadden R^2 of .0724 implies that performance efficiency measures account for about 7.24% of the variance in match outcomes, which is a modest amount but consistent with binary classification models in sports prediction. A large number of unmeasurable psychological factors (such as resilience, confidence, quality of decision making) as well as other factors such as team dynamics, environmental conditions and opponent specific tactics likely have a substantial influence on outcomes. Therefore, the model should be viewed as a probabilistic decision support tool, and not as an outcome predicting tool.

Slightly lower female model discrimination (AUC = 0.6273 vs. 0.6603 for males, though overlapping CIs) may reflect greater competitive parity in women's professional beach volleyball or systematic differences in skill distributions. Additional research with gender-balanced sampling is recommended to clarify whether technical factors operate identically across genders.

Complete-case analysis and data quality consideration

The analysis used complete case analysis (n = 10,095 of 76,756; 13.2 % inclusion rate). The use of complete case analysis represents a conscious decision to apply data quality standards rather than representational limitations of methods. The BigTimeStats database has undergone significant quality enhancements between 2000-2015 as international tournament standardization increased. Approximately 76% of all excluded matches (n = 66,661) were from before 2002 or had missing performance statistics for at least one player. Historical and temporal unevenness in this type of data are expected in historical sports databases where formal statistical monitoring was phased-in over decades.

Critically, we tested to determine if complete-case analysis produced systematic selection bias through a comparison of included versus excluded matches regarding their four key baseline characteristics using

explicit statistical testing. The results found that there were no statistically significant differences in gender distribution (included 49.9% Male vs. excluded 48.2% Male; $\chi^2 = .31$, $p = .58$), in upset rates (27.0% vs. 28.5%; $\chi^2 = 1.84$, $p = .18$), in tournament type distribution (AVP vs. FIVB; $\chi^2 = 1.07$, $p = .30$), or in ranking differentials (included M = 1247 ± 456 vs. excluded M = 1203 ± 512; $t(76754) = 1.12$, $p = .26$). This lack of statistically significant differences between these variables indicates that the full case sample represents the larger population as a whole with no apparent systematic selection bias.

External temporal validation for data collected in 2016-2019 (the period with uniformity of data quality standards to support complete data quality, n = 1360 matches) provided substantial evidence that the variability in data quality in the earlier time frame has not substantially impacted the overall validity of the generalizability: The AUCs remained relatively constant (difference = -0.0234, less than the .05 threshold) and suggest that technical relationships have remained relatively constant over the different time frames despite the variation in the degree of standardization of recording data.

While complete case analysis reduces the effective number of observations and can create an underestimation of the uncertainty associated with point estimates, multiple imputation will be used in the subsequent work as a method for leveraging all available matches (n = 76,756) and will provide precision gains and robustness across varying missing data scenarios. The analysis is also being conducted using sensitivity stratification based on the quality of data completeness (i.e., tournament matches that were completed prior to 2005, versus those completed post-2005; tournaments with live broadcast monitoring, versus those that did not have live broadcast monitoring). These analyses showed very little variability in coefficients, thereby suggesting complete case analysis was valid.

Implementation guidance

The threshold of .741 represents the point at which a team can be classified in one of two categories based on probability of winning: a) .741 or greater: classify as probable routine win by the higher ranked team; b) Less than .741: classify as possible upset. Sensitivity is 59.9%, indicating that approximately 60% of all actual upsets will be identified. Specificity is 64.5%, meaning that approximately 65% of all actual non-upsets are identified. Practitioners should use the .741 threshold with their predicted probabilities when using this model to assist in determining tournament seeding and assessing competitive balance.

Future research and implementation

There are several directions for further study: (1) The temporal validity of the models should include data from 2020–2026. Rolling re-training will help the model remain calibrated as the competitive environment continues to shift; (2) To determine how useful these models would be in a field setting, prospective field studies can be used to seed tournaments based on model predictions, then measure the performance that follows; (3) Researchers can compare advanced machine learning techniques to logistic regression with the same level of validation to eliminate over-fitting; (4) A mechanistic approach using event-level statistics, video analysis, and athlete self-reporting can provide insight into what drives efficiency, such as better decision making, or being mentally tougher; (5) A cross-sport comparison of efficiency-based models may also examine their applicability in other two player net sports (e.g., doubles tennis, squash, padel).

Continued monitoring is essential to ensure equitable model performance across demographic groups. Models should be framed as decision-support tools complementing expert judgment, particularly for athlete welfare and selection decisions (Schelling et al., 2021).

Predictive association vs. mechanistic causation

Although the six predictors in the logistic regression model show a high degree of discrimination toward predicting match outcomes (AUC = 0.6578), they do not represent a mechanism to explain why those predictions were made. For example, an age differential shows a significant relationship to match outcome, but since an individual's age is fixed, it is likely that the age differential represents a proxy for a person's level of experience or their ability to understand tactics as opposed to their age. Similarly, the kill efficiency differential shows the greatest association to match outcomes, but as with decision-making quality and other environmental/physical/psychological factors that contribute to kill efficiency, these are not represented by the current model.

Coaches and practitioners should view the model as a predictive tool that will provide direction for strategic planning, not as an indicator of a cause-and-effect relationship. To establish a causal pathway, prospective field studies using a mechanistic hypothesis (e.g., kill-efficiency training intervention with measurable outcomes) are needed. Additionally, in order to understand if the differences in kill-efficiency between the two groups is due to a difference in decision-making, physical performance, or the match-up, video-based event analysis can be used. The current model has provided the necessary foundational validation to support future mechanistic research.

CONCLUSION

This study was able to develop and test a simple logistic regression model that uses six normalized measures of player performance to predict the outcome of a professional beach volleyball match. Kill efficiency differentials and age differentials were shown to be the most important predictors of a team's competitive success in this sport, illustrating how teams are successful on offense and have experienced players on their roster. The model was shown to have an acceptable degree of discrimination (area under curve = 0.6578), and it had an acceptable degree of calibration in all three validation methods tested.

Critical information on the extent of generalization from this model was derived from temporal external validity that demonstrates high robustness over a competitive period of nearly two decades as well as increased parity of competition during the same time frame. Gender-specific analysis revealed equal applicability to both male and female athletes. The modest overall proportion of variance explained by this model (McFadden R^2 = .0724), as would be expected with the complex nature of elite athletics, also indicates that there is considerable influence of unmeasured variables related to psychology, interpersonal relationships, and context.

The validated predictive framework is a useful tool to provide immediate utility for multiple stakeholders in addition to being beneficial to tournament organizers, coaching staff and performance analysts who are interested in using the framework as a basis to make decisions about the best way to train their teams to achieve the desired results. In addition, because of the temporal stability of the model, it will allow coaching staffs to have a longer term view of their training program and plan accordingly.

Methodology, this work sets standards for the application of prediction modelling to complex behavioural systems by utilizing rigorous methods of validation, quantifying transparency related to uncertainty, assessing fairness within gender-strata, and establishing an external validation process, which is above standard of most professional sports organizations. This analytical framework will provide a methodical template for building evidence-based predictive models across a wide range of disciplines that require a high level of decision support.

AUTHOR CONTRIBUTIONS

K.N.A. conceived the study, curated the dataset, performed the statistical analyses, and drafted the manuscript. D.K.S. contributed to study design, statistical modelling choices, and interpretation of findings. N.A. contributed to literature review, domain-specific interpretation, and revision of the manuscript for important intellectual content. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work. E.E.N. contributed to literature review, domain-specific interpretation, and statistical modelling choices.

SUPPORTING AGENCIES

No funding agencies were reported by the author.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

ETHICAL CONSIDERATIONS

This study analysed fully anonymized, publicly available match-level data from the BigTimeStats beach volleyball database (hosted on Kaggle). In accordance with institutional policy, the University of Cape Coast Ethics Committee considered the study exempt from formal ethical review because no identifiable human participant data were used and all data were publicly available. All procedures complied with the ethical standards and applicable laws of the country in which the research was conducted and with the principles of the Declaration of Helsinki.

AVAILABILITY OF DATA AND MATERIALS

The raw match-level dataset analysed in this study is available from the BigTimeStats beach volleyball database on Kaggle (https://www.kaggle.com/datasets/jessemostipak/beach-volleyball, accessed January 14, 2026). under the terms specified by the original data providers. All data processing and analysis code, as well as scripts to reproduce the statistical models and figures presented in this article, are openly available in the project's GitHub repository: https://github.com/nyantakyiappiah-eng/beach-volleyball-prediction

ACKNOWLEDGMENTS

The authors thank BigTimeStats for making the beach volleyball database publicly available and acknowledge the support of University of Cape Coast for providing academic and technical support during this project

REFERENCES

Albert, J. (2008). Streaky Hitting in Baseball. Journal of Quantitative Analysis in Sports, 4(1). https://doi.org/10.2202/1559-0410.1085

Almulla, J., & Alam, T. (2020). Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League. IEEE Access, 8, 213695-213705. https://doi.org/10.1109/ACCESS.2020.3038601

Alves, H., Voss, M.W., Boot, W.R., Deslandes, A., Cossich, V., Salles, J.I. and Kramer, A.F. (2013). Perceptual-Cognitive Expertise in Elite Volleyball Players. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00036

Beach Volleyball. (n.d.). Www.kaggle.com. Retrieved from [Accessed 2026, 11 March]: https://www.kaggle.com/datasets/jessemostipak/beach-volleyball

Bisagno, E., & Morra, S. (2018). How do we learn to "kill" in volleyball?: The role of working memory capacity and expertise in volleyball motor learning. Journal of Experimental Child Psychology, 167, 128-145. https://doi.org/10.1016/j.jecp.2017.10.008

Buková, A., Zusková, K., Szerdiová, L., & Küchelová, Z. (2017). Demographic factors and physical activity of female undergraduates. Physical Activity Review, 5, 202-211. https://doi.org/10.16926/par.2017.05.25

Cañal-Bruland, R. (2011). Differentiating Experts' Anticipatory Skills in Beach Volleyball. Research Quarterly for Exercise and Sport, 82(4). https://doi.org/10.5641/027013611X13275192111745

Choi, K. H., & Byun, J. (2024). Professionalization of action sports: field-and organizational-level professionalization of new Olympic sports. Sport in Society, 1-24. https://doi.org/10.1080/17430437.2024.2325970

Chun, Y.-J., & Shin, H.-J. (2020). Defensive Tactical Analysis of Back Attack in Men's Professional Volleyball Match. Korean Journal of Sports Science, 29(1), 773-781. https://doi.org/10.35159/kjss.2020.02.29.1.773

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Ben Van Calster, Ghassemi, M., Liu, X., Reitsma, J. B., Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., & Hoffman, M. M. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, e078378-e078378. https://doi.org/10.1136/bmj-2023-078378

Csató, L. (2023). A paradox of tournament seeding. International Journal of Sports Science & Coaching, 18(4), 1277-1284. https://doi.org/10.1177/17479541221141617

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845. https://doi.org/10.2307/2531595

Giatsis, G., Lola, A., Hatzimanouil, D., & Tzetzis, G. (2023). Evaluation of a beach volleyball skill instrument for the line shot attack. Journal of Physical Education, 34(1). https://doi.org/10.4025/jphyseduc.v34i1.3409

Grgantov, Z., Jelaska, I., & Šuker, D. (2018). Intra and Interzone Differences of Attack and Counterattack Efficiency in Elite Male Volleyball. Journal of Human Kinetics, 65(1), 205-212. https://doi.org/10.2478/hukin-2018-0028

Iba, K., Shinozaki, T., Maruo, K., & Noma, H. (2021). Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. BMC Medical Research Methodology, 21(1). https://doi.org/10.1186/s12874-020-01201-w

Kheddoum, A., Hadji, M. L., & Khaled, M. (2025). Creating An App To Measure Decision-Making In Volleyball. 403. https://doi.org/10.37139/1988-017-001-032

Kianifard, F., & Vach, W. (1995). Logistic Regression with Missing Values in the Covariates. Technometrics, 37(4), 460. https://doi.org/10.2307/1269744

Knoblochova, M., Mudrak, J., & Slepicka, P. (2021). Achievement goal orientations, sport motivation and competitive performance in beach volleyball players. Acta Gymnica. https://doi.org/10.5507/ag.2021.016

Kostyukov, V., & Dashaev, K. (2022). Block-modular program of pre-competitive training of athletes of mass categories in beach volleyball. Fizicheskaya Kul'tura, Sport - Nauka I Praktika, 1, 52-57. https://doi.org/10.53742/1999-6799/1_2022_52

Künzell, S., Schweikart, F., Köhn, D., & Schläppi-Lienhard, O. (2014). Effectiveness of the Call in Beach Volleyball Attacking Play. Journal of Human Kinetics, 44(1), 183-191. https://doi.org/10.2478/hukin-2014-0124

Liu, W. B., Wu, Y. J., Li, B., & Deng, L. (2012). Model Analysis of Analysis and Evaluation about NBA Match. Advanced Materials Research, 468, 1516-1520. https://doi.org/10.4028/www.scientific.net/AMR.468-471.1516

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). On the Development of a Soccer Player Performance Rating System for the English Premier League. Interfaces, 42(4), 339-351. https://doi.org/10.1287/inte.1110.0589

Nattino, G., Pennell, M. L., & Lemeshow, S. (2020). Rejoinder to "Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test." Biometrics. https://doi.org/10.1111/biom.13250

Pencina, M. J., D' Agostino, R. B., D' Agostino, R. B., & Vasan, R. S. (2007). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine, 27(2), 157-172. https://doi.org/10.1002/sim.2929

Phillips, J., B. Caudill, S., & Mixon, F. G. (2015). Tournament Seeding Efficiency and Home Court Advantage: College Basketball's National Invitation Tournament. International Journal of Statistics and Probability, 4(3). https://doi.org/10.5539/ijsp.v4n3p101

Predoiu, R. (2023). Resilience, Risk-Taking Behavior and Aggression among Female Volleyball Players - a Preliminary Study. Journal of Psychological Science and Research, 3(2). https://doi.org/10.53902/JPSSR.2023.03.000542

Saif, Mohd., Khan, S., & Abraham, B. (2025). Basic Defensive Stance in Volleyball. International Journal for Multidisciplinary Research, 7(5). https://doi.org/10.36948/ijfmr.2025.v07i05.55848

Schelling, X., & Robertson, S. (2020). A development framework for decision support systems in high-performance sport. International Journal of Computer Science in Sport, 19(1), 1-23. https://doi.org/10.2478/ijcss-2020-0001

Schelling, X., Fernández, J., Ward, P., Fernández, J., & Robertson, S. (2021). Decision Support System Applications for Scheduling in Professional Team Sport. The Team's Perspective. Frontiers in Sports and Active Living, 3. https://doi.org/10.3389/fspor.2021.678489

Seweryniak, T., Mroczek, D., & Łukasik, Ł. (2013). Analysis and Evaluation of Defensive Team Strategies in Women's Beach Volleyball - An Efficiency-Based Approach. Human Movement, 14(1). https://doi.org/10.2478/v10038-012-0047-9

Solomon, C., & van Coller-Peter, S. (2019). How coaching aligns the psychological contract between the young millennial professional and the organisation. SA Journal of Human Resource Management, 17. https://doi.org/10.4102/sajhrm.v17i0.1146

Tili, M., & Giatsis, G. (2011). The height of the men's winners FIVB Beach Volleyball in relation to specialization and court dimensions. Journal of Human Sport and Exercise, 6(3), 504-510. https://doi.org/10.4100/jhse.2011.63.04

Umarov, K. M. (2024). Effectiveness of Developing the Technique of Attacking Movements of Young Volleyball Players. Pubmedia Jurnal Pendidikan Olahraga, 1(3). https://doi.org/10.47134/jpo.v1i3.361

Yusni, Y., Meutia, F., & Taufik, N. H. (2025). Advantageous Correlation Between Anthropometry and Physical Fitness in Amateur Soccer Players. Retos, 67, 985-995. https://doi.org/10.47197/retos.v67.112998

Zhao, H. (2018). Sports Situation-Based Neural Mechanism of High-level Volleyball Players' Decision-making Behavior. NeuroQuantology, 16(6). https://doi.org/10.14704/nq.2018.16.6.1602

Zhou, S., & Liu, H. (2025). Injury risk analysis of movement restriction and body asymmetry in sports injury prediction. BMC Sports Science, Medicine and Rehabilitation, 18(1). https://doi.org/10.1186/s13102-025-01465-z