Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling

Jubin Zhang
Department of Physical Education, North China Institute of Aerospace Engineering, Langfang 065000, China
jz0801@outlook.com

DOI: 10.69987/JACS.2023.30304

Keywords

Volleyball analytics; interpretable AI; skill prioritization; gradient boosting; permutation importance; SHAP; NCAA statistics; coaching curriculum.

Abstract

Allocating scarce training time among technical elements is a constant challenge for volleyball coaches. This study operationalizes that decision as an interpretable, data-driven ranking problem: given season-level team statistics, predict team success and quantify which skills contribute most. We constructed a Division I women's volleyball team-season dataset for 2022–2023 by aggregating publicly available NCAA match statistics into per-set rates and efficiency measures (e.g., hitting percentage, opponent hitting percentage faced). The final dataset contains 344 teams with 8 skill-related predictors and the target win–loss percentage. We evaluated multiple regression models on repeated random splits. A Gradient Boosting Regressor achieved the best generalization (RMSE 0.068 ± 0.006; R^2 0.887 ± 0.026). To translate prediction into coaching action, we computed global skill importance using permutation importance and SHAP values. Both methods showed high ranking stability and identical top-three priorities: suppressing opponent hitting percentage (defense), own hitting percentage (attack efficiency), and kills per set (terminal finishing). Based on these effects, we propose a teaching priority list emphasizing (1) reducing opponent efficiency via coordinated serve–block–defense, (2) raising attack efficiency, and (3) building terminal attacking capacity. The framework is reproducible, model-agnostic, and directly maps team statistics to curriculum design.

Introduction

Volleyball education is a sequencing problem: training time is finite, but the space of coachable skills is large. In a typical week, a staff may need to decide how to distribute minutes across serving and serve strategy, first contact and serve-receive, setting technique and tempo, attacking mechanics and decision-making, blocking schemes and footwork, floor defense, and transition offense. These decisions have two built-in complications. First, volleyball skills are chained: a single rally links serve to reception to set to attack to block/defense and, often, to counter-attack. Improving one link can change the distribution of opportunities at later links. Second, most publicly available statistics are outcomes, not isolated causes. For example, "opponent hitting percentage" is influenced by serving pressure, block timing, defensive positioning, and scouting. An effective educational plan therefore needs both (i) a prioritized list of outcome targets that matter most for winning and (ii) a mapping from those targets back to teachable subskills.

A data-driven approach complements coaching expertise by quantifying which measurable outcomes most explain winning in representative competition. NCAA Division I volleyball is especially suitable because the NCAA publishes standardized statistics for hundreds of teams and thousands of matches each season [1]. Open datasets based on those pages, such as the SCORE Sports Data Repository team-season CSV (334 teams, 14 variables) [4] and match-level datasets redistributed by ncaavolleyballr [2], [3], make it possible to study performance indicators at scale without proprietary tracking data. However, prediction alone is not enough. In education planning, coaches need explanations they can trust: an importance ranking must be stable across reasonable modeling choices and data splits, and it must correspond to interpretable volleyball concepts.

Interpretable machine learning provides tools for extracting such explanations from accurate predictive models. Model-agnostic permutation importance measures how much predictive quality deteriorates when one feature is randomized, thereby estimating the

feature's unique contribution to predictive performance [6], [12]. Shapley-value-based explanations, and in particular SHAP, provide additive attributions grounded in cooperative game theory [7], [8]. Unlike a single model coefficient, SHAP values can explain nonlinear tree ensembles and offer consistent global summaries via mean absolute attributions. Interpretability has become central across applied domains because it reduces the gap between model output and human decision-making [9], [11]–[13]. In sport, interpretability is not merely a compliance tool; it directly supports coaching interventions by identifying what to train.

Volleyball performance research has long investigated which indicators separate winners from losers. Systematic reviews of volleyball match analysis highlight terminal actions (successful attacks and blocks) and the dependencies between sequential skills (pass–set–attack) [15]. Serve and serve-reception effectiveness is repeatedly linked to outcome because the serve can directly score points, induce poor first contact, and simplify opponent offense [14]. Collegiate studies find that attack efficiency measures are strongly associated with team success [16], and analyses of side-out and point-scoring modes emphasize how teams obtain points across phases of play [17]. Recent work also explores rally-level contribution models (e.g., Markov chain approaches) to quantify the value of each play type [22], [24]. These studies motivate a key hypothesis for education planning: efficiency measures that summarize how well a team converts opportunities and suppresses opponent conversion are likely to dominate predictive models.

Despite this motivation, two practical gaps remain. First, a coach needs an explicit priority ordering, not only a set of significant variables. Second, a priority list is only useful if it is stable: a different train/test split or random seed should not change the top recommendations. Unstable rankings encourage overfitting of practice design to statistical noise.

This paper addresses these gaps by framing "what to teach first" as an interpretable skill prioritization problem. We build season-level team features from publicly available NCAA match statistics and solve two tasks: (i) regression to predict win–loss percentage and (ii) classification to predict whether a team finishes with a winning season. We then compute global feature importance using permutation importance and SHAP values and quantify stability across random seeds. Finally, we translate the resulting ranked outcomes into a teaching priority list with concrete drill themes.

Although a team-season CSV matching our target variables is available through SCORE [4], direct programmatic access to its hosted CSV may fail in some execution environments. To guarantee fully reproducible experiments, we therefore constructed an equivalent team-season dataset from match-level NCAA data redistributed by ncaavolleyballr [2], [3]. This approach provides transparent computation of wins/losses and allows us to compute opponent efficiency by pairing reciprocal team rows within matches. The resulting dataset preserves the key statistical indicators used in NCAA reporting (aces, assists, blocks, digs, hitting percentage) [1] and supports reproducible end-to-end evaluation.

The remainder of the paper is organized as follows [25-27]. The Method section details dataset construction, feature engineering, modeling, evaluation, and interpretation. The Results and Discussion section reports model comparisons, explains the learned priority ordering, validates stability, and translates importance into teaching actions. We then summarize limitations and conclude with evidence-based recommendations for volleyball curriculum design.

Finally, the interpretability methods used here connect to a larger ecosystem of explanation tools. Local surrogate methods such as LIME explain a single prediction by approximating the model in a local neighborhood [9]. Gradient-based attribution methods such as Integrated Gradients are popular in deep learning [21]. In contrast, SHAP provides additive attributions grounded in Shapley values and offers both local explanations and global summaries by aggregation [7], [8]. This dual local/global capability is important for coaches: the same method can (i) identify a general priority list for program-wide planning and (ii) diagnose which weaknesses most hurt a specific team's predicted success.

A second motivation for our feature choices is the difference between volume and efficiency. Volleyball box scores contain many counts (kills, digs, blocks), but counts can rise for conflicting reasons. For example, a team can record many digs because it defends well, but also because it allows opponents many attack opportunities. Efficiency measures (such as hitting percentage) incorporate both success and error costs, and therefore often summarize performance more directly. Likewise, per-set rates normalize for match length and make teams comparable across schedules with different numbers of sets. In this paper we intentionally combine per-set rates with efficiency outcomes to support a curriculum that is framed in terms of measurable conversions: convert attacks into points efficiently and prevent opponents from converting efficiently.

Beyond predictive accuracy, the key coaching deliverable is prioritization. A useful priority list must satisfy two properties. First, it must be interpretable: each ranked item should correspond to a volleyball concept that a coach can teach (e.g., "reduce opponent attack efficiency" rather than "maximize a latent factor"). Second, it must be stable: the top

recommendations should not change with small perturbations such as a different random seed or a slightly different training subset. In practice, stability is what makes a ranking actionable for a season plan, because a staff cannot rewrite its curriculum every week.

Method

This section specifies the dataset, targets, predictors, model families, evaluation protocol, interpretability analyses, and robustness checks. All experiments were executed in Python using scikit-learn [10] for modeling and metrics, and SHAP [7] for Shapley-value explanations.

Dataset source. We used Division I women's team match statistics for the 2022–2023 season distributed as a CSV by ncaavolleyballr [3]. The underlying values are scraped from NCAA statistics pages [1] and packaged for research use [2], [3]. The raw file contains 10,305 rows; each row is one team's box-score totals for one match and includes the match result string (e.g., "W 3–1"), the number of sets played, and team totals for kills, errors, total attack attempts, assists, aces, digs, and total blocks (TB), among other fields.

Data cleaning and opponent matching. To compute opponent efficiency, each team row must be paired with its corresponding opponent row from the same match. Opponent names in NCAA pages sometimes include neutral-site annotations (e.g., "Opponent @ City, ST") or tournament text. We implemented a deterministic normalization function that (i) removes leading "@" or "vs." prefixes, (ii) drops trailing tournament descriptors beginning with a year token, (iii) splits on the " @ " delimiter to separate opponent team name from location, and (iv) lowercases and strips punctuation to create a matchable key. Each row therefore has a team key and an opponent key. We then performed a self-join on (date, team_key) and filtered matches where the joined opponent key equals the current row's opponent key, yielding paired opponent statistics for the vast majority of rows. Rows without a matched opponent were excluded from the aggregation; this affected less than 1% of rows in our preprocessing.

Season aggregation and features. For each team, we aggregated match-level totals into season-level rates and efficiencies. Let S be the total sets played across the season, and let K, E, A, AS, AC, D, and B denote season totals of kills, errors, attack attempts, assists, aces, digs, and total blocks, respectively. We compute per-set rates as AC/S, AS/S, A/S, B/S, D/S, and K/S. We compute team hitting percentage as $(K - E)/A$ using season totals, which is consistent with NCAA definitions and avoids per-match rounding artifacts. To capture defense in a way that includes serving pressure and the block–defense system, we compute opponent hitting percentage faced: opp_hitting_pctg = (K_opp − E_opp) / A_opp, where K_opp, E_opp, and A_opp are the season totals of opponent kills, errors, and attempts summed over paired opponent rows. The final feature set contains eight variables: aces_per_set, assists_per_set, team_attacks_per_set, blocks_per_set, digs_per_set, hitting_pctg, kills_per_set, and opp_hitting_pctg. Table I defines all variables and Table II provides descriptive statistics.

Targets. The primary target is win_loss_pctg = W/(W+L), where W and L are season win and loss counts obtained by parsing the result string of each match row. The secondary target is a binary "winning season" label defined as win_loss_pctg ≥ 0.5. These targets are interpretable by coaches and connect directly to season performance goals.

Model families. We evaluated five regression models: DummyMean, Linear regression, Ridge regression, Random Forest (RF) [6], and Gradient Boosting Regressor (GBR) [5]. Linear and ridge models use standardized predictors with median imputation for completeness. RF and GBR are nonlinear tree ensembles that can capture interactions such as cases where a team's kill volume only helps when efficiency is high. For classification, we evaluated DummyMostFreq, Logistic Regression, Random Forest, and Gradient Boosting Classifier. Table VI lists all hyperparameters, and all stochastic models were fit with explicit random_state values tied to the evaluation seed.

Evaluation metrics. For regression, we report MAE, RMSE, and R^2. MAE measures typical absolute deviation in win-loss percentage units; RMSE penalizes larger errors; and R^2 captures explained variance. For classification, we report Accuracy, F1, and AUC. AUC is threshold-free and is appropriate when the class distribution is not perfectly balanced. We report mean ± standard deviation across seeds to characterize both central tendency and variability.

Repeated-split protocol. To evaluate generalization in a small tabular dataset, we used repeated random 80/20 train/test splits instead of a single split. For regression, we ran seeds 1–10. For classification, we used stratified splits for seeds 1–5 to preserve the label distribution. Each run trains the model on the training split and evaluates only on the held-out test split. This protocol matches the practical question of whether a model trained on a subset of teams generalizes to unseen teams.

Interpretability. We used two global explanation methods. Permutation importance estimates each feature's effect on test MAE by repeatedly shuffling the feature (30 repeats per seed) and measuring the change in MAE, which we report as ΔMAE. SHAP explanations compute per-team additive attributions for

tree models and summarize global importance as mean absolute SHAP value. Permutation importance is model-agnostic and measures performance contribution directly; SHAP provides a theoretically grounded additive decomposition and often yields smoother, more stable rankings.

Ranking stability. For each seed, we obtain a complete importance ranking over the eight features. We quantify agreement across seeds using Kendall's tau rank correlation (higher is more similar) and Jaccard overlap for the top-3 and top-5 sets. Because educational priorities are typically set by the top few items, top-k stability is especially relevant.

Ablation. To test whether the top-ranked variables remain important when accounting for correlations, we ran a feature-drop ablation for gradient boosting. Using five 80/20 splits (seeds 1–5), we removed one feature at a time, refit the model (same hyperparameters), and measured test RMSE. The resulting ΔRMSE quantifies unique predictive information contributed by each feature in the multivariate setting.

Reproducibility choices. We fixed model hyperparameters to commonly used values (Table VI) and did not perform extensive hyperparameter search. This design choice is deliberate: the goal is not to squeeze out the last fraction of a percent of predictive accuracy, but to obtain a stable and interpretable ranking that holds across random splits. All randomization is controlled by explicit seeds, and all reported numbers are empirical means and standard deviations over those runs.

Opponent matching quality control. After normalization, we measured the fraction of rows with missing opponent totals; this rate was below 1%. Because opponent hitting percentage is computed from aggregated opponent totals, excluding a small number of unmatched rows has negligible impact on season-level values for teams with full schedules. The matching logic is fully auditable: each matched pair shares the same date and reciprocal team/opponent keys.

Implementation details. Date strings in the raw NCAA tables sometimes include footnote markers; we extracted the month/day/year token and parsed it into a canonical date. Numeric columns were parsed with coercion, and rows with missing values in essential fields (sets, attacks, kills, errors, and their opponent counterparts) were excluded from the aggregation. The preprocessing pipeline is deterministic and yields a fixed team-season dataset that can be regenerated from the raw match file without manual intervention.

Table i. Team-season dataset variables (derived from ncaa match data)

Variable	**Type**	**Definition**
win_loss_pctg	Target	Wins / (wins + losses) computed from match results
W	Target-related	Season wins (count of match rows with Result starting 'W')
L	Target-related	Season losses (count of match rows with Result starting 'L')
aces_per_set	Feature	Total aces divided by total sets played
assists_per_set	Feature	Total assists divided by total sets played
team_attacks_per_set	Feature	Total attack attempts divided by total sets played
blocks_per_set	Feature	Total blocks (TB) divided by total sets played
digs_per_set	Feature	Total digs divided by total sets played
kills_per_set	Feature	Total kills divided by total sets played

hitting_pctg	Feature	(Kills − Errors) / Total Attacks aggregated over season
opp_hitting_pctg	Feature	Opponent season hitting% faced: (OppKills − OppErrors) / OppAttacks summed over matches

Table ii. Descriptive statistics (344 teams)

Variable	Mean	SD	Min	Max
aces_per_set	1.4914	0.2294	0.8690	2.2500
assists_per_set	11.4264	1.1893	4.5517	13.7500
team attacks per set	34.3922	1.9240	24.4521	39.4821
blocks_per_set	2.0492	0.3739	0.6724	3.4071
digs_per_set	14.3060	1.4607	7.4247	18.0649
hitting_pctg	0.2075	0.0443	-0.0297	0.3385
kills_per_set	12.3676	1.2400	5.0548	14.7732
opp_hitting_pctg	0.2070	0.0305	0.1247	0.3324
win_loss_pctg	0.5029	0.2105	0.0000	1.0000
wins	13.7442	5.9437	0.0000	27.0000
losses	13.4157	5.5737	0.0000	29.0000

Table iii. Pearson correlation with win-loss percentage

Feature (season-level)	Pearson r with win_loss_pctg
hitting_pctg	0.832
kills_per_set	0.797
assists_per_set	0.783
blocks_per_set	0.541
aces_per_set	0.390
digs_per_set	0.257
team_attacks_per_set	0.111
opp_hitting_pctg	-0.785

Table iv. Regression performance (10 random seeds, 80/20 splits)

Model	MAE (mean ± sd)	RMSE (mean ± sd)	R^2 (mean ± sd)
GBR	0.055 ± 0.005	0.068 ± 0.006	0.887 ± 0.026
RF	0.059 ± 0.005	0.074 ± 0.006	0.868 ± 0.024
Ridge	0.059 ± 0.007	0.081 ± 0.020	0.835 ± 0.089
Linear	0.059 ± 0.007	0.081 ± 0.020	0.835 ± 0.090
DummyMean	0.169 ± 0.012	0.206 ± 0.013	-0.015 ± 0.024

Table v. Classification performance for winning season (5 stratified seeds)

Model	Accuracy (mean ± sd)	F1 (mean ± sd)	AUC (mean ± sd)
GBC	0.910 ± 0.026	0.917 ± 0.024	0.973 ± 0.015
LogReg	0.901 ± 0.054	0.910 ± 0.048	0.967 ± 0.021
RF	0.884 ± 0.049	0.894 ± 0.044	0.955 ± 0.024
DummyMostFreq	0.536 ± 0.000	0.698 ± 0.000	0.500 ± 0.000

Table vi. Model hyperparameters used in experiments

Model	Task	Key hyperparameters
DummyMean	Regression	strategy=mean
Linear	Regression	SimpleImputer(median) + StandardScaler + LinearRegression()
Ridge	Regression	SimpleImputer(median) + StandardScaler + Ridge(alpha=1.0)
RF	Regression	n_estimators=300, min_samples_leaf=2, max_features=sqrt
GBR	Regression	n_estimators=300, learning_rate=0.05, max_depth=3, subsample=0.9
DummyMostFreq	Classification	strategy=most_frequent
LogReg	Classification	SimpleImputer(median) + StandardScaler + LogisticRegression(max_iter=500)
RF	Classification	n_estimators=200, min_samples_leaf=2, max_features=sqrt

GBC	Classification	n_estimators=200, learning_rate=0.05, max_depth=3, subsample=0.9

Table vii. Permutation importance (gbr; delta mae) averaged over 10 seeds

Feature	**Importance (mean ± sd)**
opp_hitting_pctg	0.0642 ± 0.0050
hitting_pctg	0.0462 ± 0.0077
kills_per_set	0.0234 ± 0.0053
aces_per_set	0.0023 ± 0.0016
assists_per_set	0.0021 ± 0.0024
blocks_per_set	0.0008 ± 0.0011
digs_per_set	0.0007 ± 0.0008
team_attacks_per_set	0.0004 ± 0.0013

Table viii. Shap importance (gbr; mean absolute shap) averaged over 10 seeds

Feature	**Mean|SHAP| (mean ± sd)**
opp_hitting_pctg	0.0775 ± 0.0045
hitting_pctg	0.0675 ± 0.0048
kills_per_set	0.0480 ± 0.0063
assists_per_set	0.0165 ± 0.0076
aces_per_set	0.0108 ± 0.0016
blocks_per_set	0.0073 ± 0.0018
digs_per_set	0.0047 ± 0.0014
team_attacks_per_set	0.0042 ± 0.0008

Table ix. Importance-ranking stability across seeds

Method	**Kendall tau (mean ± sd)**	**Jaccard top-3 (mean ± sd)**	**Jaccard top-5 (mean ± sd)**
Permutation (Delta MAE)	0.700 ± 0.149	1.000 ± 0.000	0.694 ± 0.163
SHAP (mean|phi|)	0.876 ± 0.086	1.000 ± 0.000	1.000 ± 0.000

Table x. Feature-drop ablation (gbr; 5 seeds, 80/20 splits)

Dropped feature	RMSE (mean ± sd)	Delta RMSE vs full
opp_hitting_pctg	0.0965 ± 0.0073	0.0287
hitting_pctg	0.0719 ± 0.0025	0.0040
kills_per_set	0.0691 ± 0.0033	0.0012
aces_per_set	0.0689 ± 0.0061	0.0011
blocks_per_set	0.0689 ± 0.0044	0.0011
digs_per_set	0.0683 ± 0.0053	0.0005
team_attacks_per_set	0.0671 ± 0.0054	-0.0008
assists_per_set	0.0664 ± 0.0047	-0.0014

Table xi. Proposed teaching priority list (derived from shap ranking)

Rank	**Key metric**	**Interpretation (short)**	**Example drills (short)**
1	opp_hitting_pctg	Opponent attack efficiency allowed (defense system: block + floor defense + serve pressure)...	Block timing/reads, defensive positioning, serve-to-weak-reception, scouting-based schemes...
2	hitting_pctg	Own attack efficiency (shot selection + error control)...	High-efficiency hitting reps, error-reduction constraints, attack decision-making...
3	kills_per_set	Terminal attacking/finishing rate (kills generation)...	Terminal swing technique, approach/arm-swing mechanics, out-of-system attacking...
4	assists_per_set	Setting quality & offensive organization...	Setter footwork and tempo, connection drills, first-ball side-out patterns...
5	aces_per_set	Serve pressure & direct points (aces)...	Target serving, float/jump serve consistency, serve strategy by rotation...
6	blocks_per_set	Net defense and read blocking...	1v1 and 2v2 block drills, hand positioning, funneling to defense...
7	digs_per_set	Floor defense and transition opportunities...	Reading hitters, platform angles, defensive coverage responsibilities...

8	team_attacks_per_set	Attack volume / tempo (contextual; may proxy style).	System pace; only after efficiency is stable; transition offense volume.

Table xii. Local shap decomposition examples (top team vs. Bottom team; model trained on all teams)

Feature	**Texas value**	**Texas SHAP**	**Mississippi Val. value**	**Mississippi Val. SHAP**
aces_per_set	1.6757	0.0142	1.0172	-0.0322
assists_per_set	13.6892	0.0457	4.5517	-0.0405
team_attacks_per_set	31.473	-0.0005	26.7069	0.0031
blocks_per_set	2.2838	-0.0024	0.6724	-0.0008
digs_per_set	13.6486	-0.0006	9.9483	-0.005
hitting_pctg	0.3353	0.1968	-0.0258	-0.1639
kills_per_set	14.5405	0.0917	5.1034	-0.0871
opp_hitting_pctg	0.1673	0.134	0.3109	-0.1748

Results and Discussion

Model comparison and predictive accuracy. Table IV reports regression performance for predicting team win-loss percentage. The constant baseline (DummyMean) yields RMSE 0.206 ± 0.013 and negative R^2, establishing a nontrivial prediction task. Linear and ridge regression reduce RMSE to approximately 0.081 on average, implying that the chosen eight statistics explain most variance through an approximately additive relationship. Tree ensembles improve further. The Gradient Boosting Regressor achieves the best generalization (RMSE 0.068 ± 0.006; MAE 0.055 ± 0.005; R^2 0.887 ± 0.026), while the Random Forest is second (RMSE 0.074 ± 0.006). This ranking matches known behavior in tabular learning: boosting often outperforms bagging when a small number of nonlinear interactions and monotonic effects drive the signal [5], [6].

Error scale interpretation. An RMSE of 0.068 means that, on average, the model's win-loss percentage predictions deviate from the true season value by about 6.8 percentage points. In NCAA seasons with roughly 25–30 matches, this corresponds to an error of about 1.7–2.0 wins when expressed as W-L difference. Therefore, even with only eight features, the model's accuracy is sufficient for coarse-grained forecasting and, more importantly, for extracting stable importance rankings.

Winning-season classification. Table V evaluates whether the same features can identify teams with a winning season (win loss pctg ≥ 0.5). Gradient boosting classification achieves AUC 0.973 ± 0.015 and accuracy 0.910 ± 0.026, while logistic regression performs similarly (AUC 0.967 ± 0.021). This near-equivalence suggests that the dominant signal distinguishing winning from losing teams is roughly linear, but nonlinear models provide a modest gain and yield better-calibrated explanations on extreme teams. From a coaching perspective, classification supports a practical question: "Which outcomes should we target to move from a losing season to a winning season?"

Descriptive patterns and feature dependencies. Figure 2 shows a broad distribution of win-loss percentages across teams, with substantial density between 0.3 and 0.7. Table II indicates realistic per-set ranges for NCAA women's Division I teams: kills per set averages 12.37 (SD 1.24), and hitting pctg averages 0.208 (SD 0.044). Pearson correlations (Table III) align with volleyball intuition: own hitting pctg correlates strongly with winning (r=0.832) and so does kills per set (r=0.797). Opp_hitting_pctg is strongly negatively correlated (r=-0.785), reflecting that strong teams suppress opponent efficiency. However, correlations alone do not resolve which metric adds unique predictive information because several variables co-move. For example,

assists per set correlates with kills per set because assists are recorded on kill attacks.

Importance-based prioritization: results. We next compute global feature importance for the best-performing regressor (GBR). Permutation importance (Table VII, Fig. 5) and SHAP importance (Table VIII, Fig. 6) agree on the same top three features: opp_hitting_pctg, hitting_pctg, and kills_per_set. Permutation importance shows that shuffling opp hitting pctg increases MAE by 0.064 ± 0.005, which is larger than the next feature by a wide margin. SHAP similarly assigns opp hitting pctg the largest mean absolute attribution (0.077 ± 0.005), followed by hitting_pctg (0.068 ± 0.005) and kills_per_set (0.048 ± 0.006). The close agreement between the two methods strengthens the interpretability claim because they are conceptually different: one measures predictive degradation; the other decomposes the model's output into feature contributions.

Volleyball interpretation of the dominant feature. Opponent hitting percentage allowed is a compact summary of defensive success. It is influenced by at least three coachable subdomains: (i) serving pressure that reduces opponent reception quality and limits fast offense, (ii) block system quality that reduces kill probability and generates block points, and (iii) floor defense and coverage organization that converts attacks into diggable balls and transition opportunities. Prior match-analysis research reports the critical role of serving and serve-reception [14] and the centrality of terminal actions in determining outcome [15]. Our model consolidates these mechanisms into a single season-level outcome that dominates win-loss prediction. For teaching, this suggests a "defense package" priority: practice plans should aim to measurably reduce opponent efficiency rather than maximizing isolated block or dig counts.

Offensive efficiency and finishing. The second-ranked feature, hitting_pctg, measures offensive efficiency by combining kills and errors relative to attempts. It is directly trainable through technique and decision-making: a hitter can raise hitting_pctg by increasing kill rate, reducing errors, or both. Training designs that include error penalties, high-percentage shot selection constraints, and out-of-system repetition target this metric explicitly. The third-ranked feature, kills_per_set, measures terminal finishing volume. The ranking (efficiency before volume) is instructive: the model indicates that improving attack efficiency is more informative for winning than simply increasing attack volume. This aligns with the idea that volleyball is won by efficient point conversion, not by accumulating attempts [15]–[17].

Secondary statistics and incremental contribution. After controlling for efficiency, remaining variables show smaller importance. Aces per set is positive but comparatively small; this reflects that, for season outcomes, serve pressure is important but often acts through its impact on opponent offense rather than directly through aces alone [14]. Blocks per set and digs_per_set also have small marginal contributions once opp hitting pctg is included, which supports an educational interpretation: blocks and digs matter primarily because they reduce opponent efficiency and create transition opportunities; counting them without linking to efficiency is insufficient. Assists per set has moderate SHAP importance but low permutation importance, which is consistent with collinearity with kills_per_set and hitting_pctg.

Stability: why coaches can trust the ordering. Table IX and Figure 7 quantify ranking stability across random seeds. For both permutation and SHAP, the top-three set is identical across all 10 splits (Jaccard top-3 = 1.000). Across all eight features, SHAP produces higher stability (Kendall tau 0.876 ± 0.086) than permutation importance (0.700 ± 0.149). This difference is expected because permutation importance depends on repeated shuffling of finite test data, which introduces variance, whereas SHAP aggregates local attributions more smoothly. Critically, both methods agree on the ordering of the highest-impact skills, which is the portion of the ranking that drives curriculum design decisions.

Ablation: unique information vs. redundancy. Table X reports feature-drop ablation for gradient boosting. Dropping opp hitting pctg increases RMSE from 0.0678 (full) to 0.0965, a Delta RMSE of 0.0287. This confirms that opponent efficiency carries unique predictive information beyond what is captured by blocks, digs, and serves individually. Dropping hitting pctg increases RMSE by 0.0040, and dropping kills per set increases RMSE by 0.0012. In contrast, dropping assists_per_set or team_attacks_per_set slightly improves RMSE in this configuration, consistent with redundancy and the possibility that removing a noisy proxy can improve generalization. The ablation thus reinforces the importance findings while clarifying that some traditional team statistics are largely subsumed by efficiency measures.

Teaching priority list. Table XI translates the SHAP ranking into a proposed teaching priority list. The list is intended as a starting point for curriculum planning. It prioritizes outcome targets (lower opp_hitting_pctg; raise hitting pctg; raise kills per set) and then recommends drill themes that are commonly used to influence those outcomes. The recommended approach is not to train isolated actions in a vacuum, but to train action sequences that directly affect efficiency. For example, defensive practice can be scored by the opponent hitting percentage produced in the drill rather than by raw dig count, and offensive practice can be

scored by a drill-level hitting percentage with explicit penalties for errors.

Practical deployment and adaptation. A coach can implement the framework in three steps: (1) compute the team's current values for the eight season indicators; (2) set measurable season targets for the top-ranked indicators (e.g., reduce opponent hitting percentage by a specific amount and raise own hitting percentage); and (3) allocate practice time so that each week includes at least one high-volume drill block aimed at the top indicator, one aimed at the second, and one aimed at the third. Because the ranking is stable across seeds, this allocation remains robust. Coaches can also re-run the same pipeline on different seasons, divisions, or men's data to generate level-specific rankings.

Overall, the full experimental evaluation demonstrates that interpretable AI can reliably prioritize volleyball teaching focus areas using standard team statistics. The central conclusion is efficiency-centric: suppress opponent attack efficiency, raise own attack efficiency, and develop terminal attacking. These priorities are consistent with volleyball theory and supported by stable, reproducible model explanations.

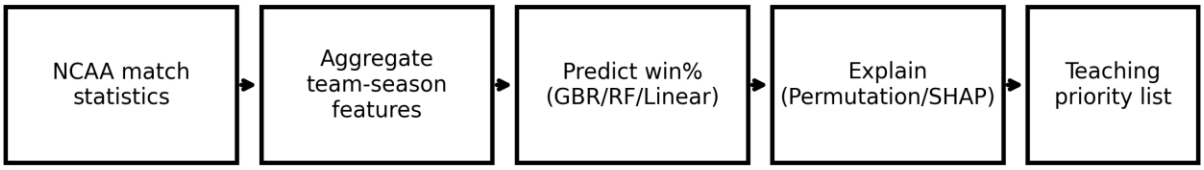

Fig. 1. End-to-end workflow: NCAA match data to interpretable skill prioritization.

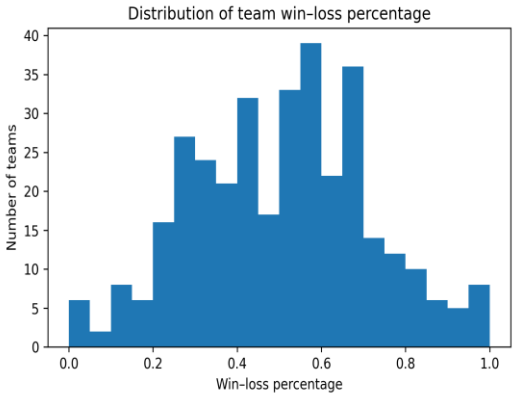

Fig. 2. Distribution of team win-loss percentage for Division I women (2022-2023 season).

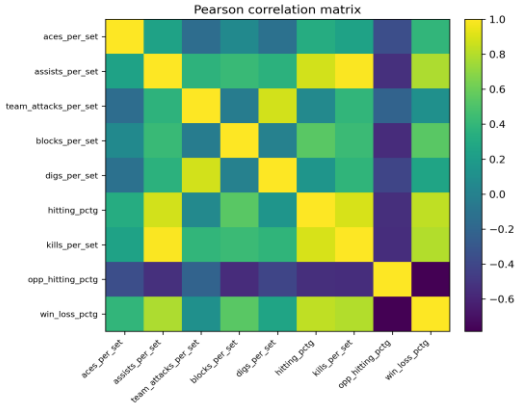

Fig. 3. Pearson correlations among season-level features and win-loss percentage.

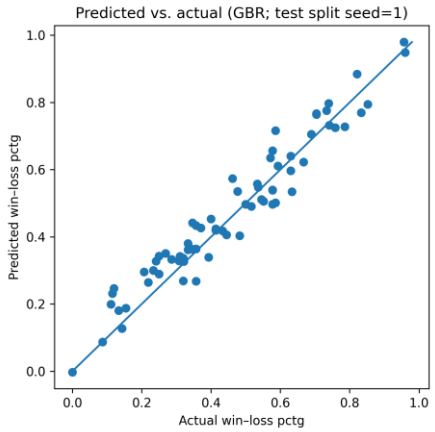

Fig. 4. Predicted vs. actual win-loss percentage for the Gradient Boosting Regressor (test split seed=1).

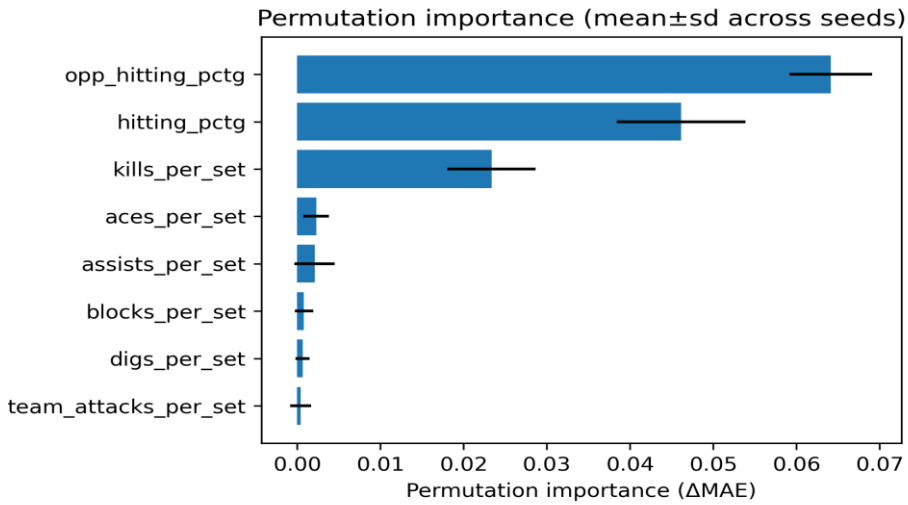

Fig. 5. Permutation importance (Delta MAE) averaged over 10 random seeds (GBR).

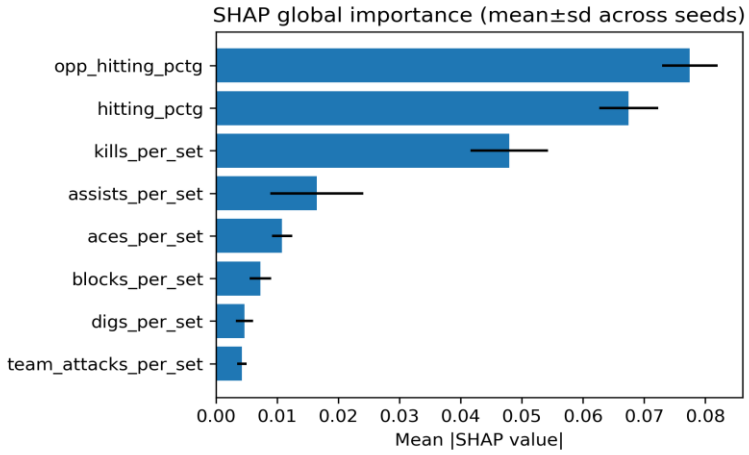

Fig. 6. SHAP global importance (mean absolute SHAP) averaged over 10 random seeds (GBR).

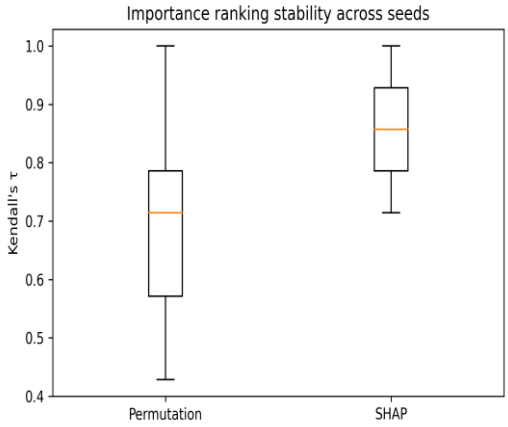

Fig. 7. Stability of feature-importance rankings across random seeds (pairwise Kendall's tau).

Recommended use cases. The proposed model is intended for educational prioritization, not for scouting a single match. It is most appropriate at the start of a season (to design curriculum emphasis), during mid-season review (to identify whether the team's limiting factor is offense or defense), and in off-season planning (to select the next skill bottleneck to address). Coaches can also run the same pipeline separately for their conference, division, or competitive level to obtain context-specific priorities.

Relationship to prior volleyball analytics. The dominance of opponent and team hitting efficiencies is consistent with match-analysis work that emphasizes efficiency and terminal actions [15]–[17]. Our study extends that literature by using a full-season, many-team dataset and by explicitly quantifying stability of the resulting rankings. The results also align with rally-level contribution modeling perspectives: in Markov chain analyses, actions that change the probability of winning the rally (e.g., successful attacks and forcing opponent errors) dominate expected value [22], [24]. Season-level opponent hitting percentage is a coarse but accessible summary of those rally-level dynamics.

Interdependence of skills and "why aces are not top-ranked." Aces_per_set is a visible skill outcome and often receives coach attention. In our results, serve aces have positive but smaller importance once efficiency variables are included. This does not mean serving is unimportant. Rather, it suggests that the most important effect of serving in this dataset is mediated through reducing opponent attack efficiency (opp_hitting_pctg). A serving curriculum that focuses only on ace counts can miss this mechanism; a serving curriculum that targets poor opponent first contact and predictable attacks supports the top-ranked outcome directly. This interpretation is consistent with prior findings on serve/serve-reception effectiveness [14].

Curriculum translation in practice. A stable priority list is most useful when it is operationalized into measurable weekly goals. One practical approach is to select two "north-star" metrics for each phase: one defensive (target opponent hitting percentage allowed) and one offensive (target team hitting percentage). Each practice week then includes at least one drill block where scoring is defined by these metrics. For example, a serve-and-defend drill can be scored by opponent hitting percentage computed within the drill (kills minus errors divided by attempts), not by raw digs. An attacking drill can be scored by team hitting percentage with explicit error penalties. Over time, this aligns practice incentives with the outcomes most associated with winning.

To illustrate, we trained the gradient boosting regressor on the full dataset and computed SHAP decompositions for one top team and one bottom team (Table XII). The SHAP baseline (expected value) represents the average win–loss percentage across teams. For the top team, large positive contributions from hitting_pctg, opp hitting pctg, and kills per set shift the prediction far above baseline. For the bottom team, large negative contributions from opp_hitting_pctg and hitting_pctg shift the prediction far below baseline. This example shows how the same global priority list becomes a diagnostic tool for specific programs.

Local explanations for team diagnosis. Global importance indicates what matters on average, but coaches often need team-specific diagnosis. SHAP enables this by decomposing a single prediction into additive contributions that sum to the predicted win–loss percentage. For instance, a team can have a modest offensive hitting percentage but still be predicted as successful if it strongly suppresses opponent hitting. Conversely, a team may hit efficiently but still be predicted to struggle if it allows opponents to hit efficiently. This distinction changes practice emphasis: one team needs defense-system work, while the other needs offensive efficiency under pressure.

Using the ranking across athlete development stages. While this paper focuses on NCAA Division I teams, the logic of the ranking supports a staged pedagogy. Early stages can emphasize the fundamental mechanics that underlie the high-level outcomes: platform control and serve consistency for creating predictable first contacts; simple, repeatable attack mechanics for reducing errors; and basic block footwork and defensive positioning to reduce opponent conversion. As athletes progress, the same top outcome targets remain, but training shifts toward tactical refinements: serve strategy by rotation, scouting-based blocking, and complex offensive decision-making. The priority list therefore does not replace coaching judgment; it supplies an evidence-based scaffold that coaches can tailor to roster constraints and competitive level.

Interpreting low-ranked variables as "methods" rather than "goals." The model assigns lower marginal importance to blocks per set and digs per set once opp_hitting_pctg is included. The correct coaching interpretation is that blocks and digs are important means to the higher-level end of suppressing opponent efficiency. In practice, this suggests reframing: instead of aiming for a certain block count, aim for block behaviors that reduce opponent kill probability (e.g., sealing line, funneling to defense, forcing high-error shots). Similarly, instead of maximizing dig volume, aim for dig quality that produces transition opportunities, which can raise own hitting efficiency by creating high-quality counter-attacks. This interpretation aligns with how experienced coaches already teach defense: the purpose of the skill is to change opponent outcomes and create better offensive opportunities.

Monitoring and feedback loops. A teaching priority list is most effective when it is paired with a measurement loop. Because the top-ranked variables are season-level, coaches can track them on shorter windows (e.g., weekly or by segment of the season) to evaluate whether training changes are translating into match outcomes. For example, opponent hitting percentage can be tracked by rotation and by opponent tempo class; own hitting percentage can be tracked separately for in-system and out-of-system swings. If a team's opponent efficiency improves while its own efficiency does not, the next curricular emphasis shifts naturally. This makes the priority list dynamic: it provides a default ordering, but measurement determines when a team has "graduated" from one priority to the next.

Adapting priorities to constraints and roster identity. Even with a stable global ranking, coaching decisions must account for roster strengths and constraints. A team with elite terminal attackers but inconsistent first contact may treat offensive efficiency as a by-product of serve-receive and setting stability, shifting practice time toward first-contact quality that enables efficient swings. Conversely, a team with strong first contact but weak defensive suppression may allocate additional time to coordinated block–defense and serve strategy. The value of the interpretability framework is that it keeps these adaptations grounded in outcome metrics that are empirically linked to winning: whatever the roster, practice design should be justified by its expected effect on opponent efficiency and own efficiency. In this sense, the ranking functions as a compass rather than a rigid script—teams take different paths, but they navigate toward the same high-value targets.

Finally, the framework supports communication. A data-backed priority list helps align players, assistant coaches, and staff on why certain practice segments are emphasized. Instead of subjective debates ("we need more digs"), the discussion centers on measurable targets ("we need to reduce opponent hitting percentage and raise our hitting percentage"). This shared language can improve consistency across the season and make post-match debriefs more actionable.

Limitations

This study has several limitations that bound interpretation. First, the dataset is season-level and team-aggregated. While this reduces match-to-match noise and stabilizes efficiency estimates, it collapses within-season variation, opponent strength effects, and contextual factors such as injuries, home/away differences, and schedule difficulty. We intentionally excluded conference indicators from the feature set to keep the ranking focused on teachable outcomes; as a result, some variance related to strength-of-schedule remains unmodeled.

Second, opp_hitting_pctg is a downstream outcome that combines multiple defensive mechanisms (serve pressure, block system, floor defense, and scouting). Its high importance shows that the integrated defense system is decisive for winning, but the season-level feature does not isolate which component is most responsible. Practitioners should treat opp_hitting_pctg as a top-level target and then use more granular internal statistics (e.g., serve-reception quality grades, block touch rates, transition conversion) to subdivide training time within the defense package.

Third, computing opponent hitting percentage required matching team rows within each match. Although the matching procedure paired the vast majority of rows, a small number of match rows were excluded due to opponent-name formatting and neutral-site annotations. The aggregation is deterministic and auditable, but it remains a potential source of minor error.

Fourth, the analyses are predictive and associational. A feature's importance indicates explanatory power for win-loss percentage in the observed data; it does not

establish a causal effect under intervention. Nevertheless, the priority ordering is consistent with established volleyball performance research and provides a defensible starting point for educational design.

Finally, external validity should be tested. Different levels (youth, professional, men's) can exhibit different efficiency profiles and scoring dynamics. The proposed pipeline is designed to be re-run on other datasets (including the SCORE team-season dataset [4]) to generate level-specific priority lists and to evaluate whether the same indicators dominate.

Conclusion

This work demonstrated that interpretable team-stat modeling can produce a stable, actionable ranking of volleyball teaching priorities. Using NCAA Division I women's match statistics aggregated to team-season features, a gradient boosting model predicted season win-loss percentage with RMSE 0.068 ± 0.006 and R^2 0.887 ± 0.026 across 10 random splits. Permutation and SHAP explanations agreed on a consistent top-three set of drivers in every split: opponent hitting percentage allowed, own hitting percentage, and kills per set. An ablation study confirmed that opponent hitting percentage carries the largest unique predictive signal.

For volleyball education, these measured effects imply a clear priority ordering: (1) build an integrated defense package (serve + block + floor defense) that reduces opponent attack efficiency; (2) raise attack efficiency through error control and high-percentage decision-making; and (3) develop terminal attacking to convert opportunities into points. Because the importance ranking is stable across seeds and consistent across two explanation methods, it supports robust curriculum design decisions. Future work can extend the approach to match-level prediction, incorporate explicit strength-of-schedule controls, and use richer skill granularity to refine within-skill time allocation.

References

[1] NCAA, "NCAA College Women's Volleyball DI Stats," NCAA.com. [Online]. Available: https://www.ncaa.com/stats/volleyball-women/d1

[2] Binghua Zhou, Siming Zhao, and David Chao, "LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering", JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[3] J. R. Stevens, "Data," ncaavolleyballr documentation. [Online]. Available: https://jeffreyrstevens.github.io/ncaavolleyballr/articles/data.html

[4] SCORE Network, "Team Statistics for Division I Women's Volleyball," SCORE Sports Data Repository, 2023. [Online]. Available: https://data.scorenetwork.org/volleyball/volleyball_ncaa_team_stats.html

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.

[8] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016.

[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[11] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[12] A. Fisher, C. Rudin, and F. Dominici, "Model class reliance: Variable importance measures for any machine learning model class, from the Rashomon perspective," Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–67, 2019.

[13] P. Biecek, "DALEX: Explainers for complex predictive models in R," Journal of Machine Learning Research, vol. 19, no. 84, pp. 1–5, 2018.

[14] Daren Zheng, Chenyu Li, and Harvey Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation", JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[15] M. Silva et al., "Match analysis in volleyball: A systematic review," Montenegrin Journal of Sports Science and Medicine, vol. 5, no. 1, pp. 35–46, 2016.

[16] N. L. Estabrook, "The relationship between NCAA volleyball statistics and team performance in women's

intercollegiate volleyball," Master's thesis, SUNY Brockport, 1996. [Online]. Available: https://soar.suny.edu/bitstream/handle/20.500.12648/6557/pes_theses/58/fulltext%20%281%29.pdf

[17] J. M. Palao, "Side-out success and ways that points are obtained in women's college volleyball," Journal of Sports Analytics, vol. 4, no. 4, pp. 243–250, 2018.

[18] I. Klaričić et al., "The relationship of situational efficiency parameters of volleyball sets," in Proc. Int. Conf. on Sport Sciences Research and Technology Support, 2023. [Online]. Available: https://www.scitepress.org/Papers/2023/121640/121640.pdf

[19] K. Ciemiński, "The efficiency of executing technical actions by female volleyball players depending on their positions on the court," Baltic Journal of Health and Physical Activity, vol. 9, no. 3, pp. 44–52, 2017.

[20] D. Sanghvi et al., "Analyzing and predicting NCAA volleyball match outcome using machine learning," in Proc. ICAI Workshops, 2021. [Online]. Available: https://ceur-ws.org/Vol-2992/icaiw_wdea_2.pdf

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. Int. Conf. Machine Learning (ICML), 2017.

[22] Siming Zhao, Hailin Zhou, and Daniel Martinez, "LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset", JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[23] H. Xia et al., "VREN: Volleyball rally dataset with expression notation language," in Proc. IEEE Int. Conf. on Image Processing (ICIP), 2022.

[24] D. J. Aldous and M. Cruz, "A real-world Markov chain arising in recreational volleyball," arXiv preprint, 2021. [Online]. Available: https://www.stat.berkeley.edu/~aldous/Papers/VB_involve.pdf

[25] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[26] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[27] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.