Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Interpretable Classification Model for Air Pistol Shot Execution Quality Based on SCATT Aiming Trajectories

Huiying Feng
Wuhan Sports University
Tong Zhou
Wuhan Sports University
Shanrui Diao
Wuhan Sports University
Yunyun Du
whsuhyh@163.com

Wuhan Sports University

Article

Keywords: air pistol, SCATT, aiming trajectory, random forest, SHAP

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10952837/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective evaluation of single-shot execution in air pistol shooting is limited by the lack of interpretable methods that reproduce coaches' technical judgments. We developed a random forest classifier using SCATT aiming trajectories and virtual shot locations from 2,857 shots by two elite athletes, labeled as good or non-good executions by professional coaches. Candidate features were organized into coach-informed groups, screened at group and variable levels, and interpreted using Shapley Additive Explanations (SHAP). External consistency was assessed in 1,566 shots from two additional athletes without retraining or threshold adjustment. On the internal test set, the final model achieved balanced accuracy of 0.8450, good-execution F1 of 0.7532, and area under the receiver operating characteristic curve (ROC-AUC) of 0.9175. The corresponding values in the additional athletes were 0.8187, 0.7513, and 0.9349. The most influential features described the relationship of the virtual impact point and terminal blue trajectory to the yellow stability region, together with the green-to-yellow transition. These findings indicate that SCATT trajectory features can reproduce coach-defined execution-quality labels and provide technically meaningful explanations for post-shot review. Because only four athletes were included, broader validation is required.

Introduction

10 m air pistol shooting is a competitive sport that requires exceptionally high levels of stability and precision[1–3]. Although competition performance is ultimately reflected by the score achieved, single-shot execution quality during training analysis primarily represents whether the pre-shot aiming process develops a stable and repeatable technical pattern, rather than being directly determined by the score of an individual shot. A high score reflects the shooting outcome, whereas single-shot execution quality also involves whether muzzle movement gradually converges before the shot, whether stability is maintained immediately before triggering, and whether abrupt deviations occur at the moment of discharge. In this study, among the 726 shots labeled as "good shots" by experts, 376 shots (51.8%) achieved scores of 10 rings or higher, whereas 350 shots (48.2%) did not reach this scoring level. This distribution indicates that the score is associated with the process quality evaluated by experts but cannot replace coaches' judgments of the complete shot execution process.

The SCATT system can continuously record the two-dimensional movement trajectory of the optical axis of the pistol barrel on a virtual target surface and represent different stages before shot release using different colors. In training practice, coaches typically evaluate shot quality by considering whether the yellow trajectory remains concentrated and stable, whether the green trajectory appropriately surrounds the yellow trajectory, whether the blue trajectory exhibits sudden deviation near shot release, and whether the virtual shot location is positioned within or close to the stable region of the yellow trajectory[1, 2, 4, 5]. Therefore, the definition of "good shot" in this study represents an operational process label assigned by coaches based on the complete SCATT trajectory profile and virtual shot location.

With the advancement of digital technologies in sports training, machine learning has been increasingly applied to performance prediction, training monitoring, and injury risk assessment[6–8]. In shooting sports, previous studies have demonstrated that aiming accuracy, holding stability, trigger control quality, postural control, and aiming-point fluctuations are closely associated with shooting performance[1–5, 9–17]. Other studies have begun to utilize SCATT data, multisensor measurements, and machine learning models to identify key factors underlying shooting performance and movement patterns[18–21]. These findings indicate that trajectory data can objectively characterize shooting technique; however, how to transform coaches' experiential judgments of "good shots" into quantitative criteria that are modelable, reproducible, and interpretable remains to be further investigated.

Random forests are well suited to SCATT trajectory data because they capture nonlinear relationships and feature interactions with few distributional assumptions[22]. However, models intended for training support must explain why a prediction was made rather than operate as opaque 'black boxes'[23, 24]. SHAP (Shapley Additive Explanations) was therefore used to quantify global and shot-specific feature contributions[25]. Previous shooting studies have mainly predicted score, performance level, or high/low score from aiming accuracy, stability, trigger control, posture, and trajectory fluctuation[1–3, 10–15, 17, 19]. Relatively little work has modeled coaches' judgments of single-shot process quality or organized SCATT variables into coach-readable semantic groups. Accordingly, this study evaluates whether a compact set of trajectory relationships can reproduce expert labels while retaining explanations that correspond to practical coaching criteria.

Based on this background, this study used SCATT aiming trajectories and expert annotations from two elite air pistol athletes to develop a binary classification model distinguishing "good shots" from "non-good shots." The objectives were: (1) to explicitly represent coaches' judgments of single-shot process quality and determine whether trajectory features could reproduce these judgments; and (2) to identify the minimal feasible feature groups at the expert semantic level, determine the final input variables for the random forest model, characterize the relationship between model decisions and coaching experience through global and individual SHAP interpretations, and conduct preliminary external consistency validation using data from two athletes who were not involved in model development. The contribution of this study is the integration of coaching expertise, feature-group evaluation, and interpretable machine learning to establish a modeling framework for single-shot execution quality assessment in training analysis, while providing an initial evaluation of the model's consistency with reference labels in new athletes.

Materials and Methods

Study Design

The study first collected SCATT single-shot aiming trajectory data from elite air pistol athletes. Criteria for identifying "good shots" were established through expert interviews, and professional coaches

assigned single-shot quality labels based on trajectory profiles and virtual shot locations provided by the SCATT system. Based on these data, a candidate feature set was constructed, and the samples were divided into training and testing sets. Subsequently, multiple baseline models were compared to assess the learnability of the classification task. A random forest model was selected as the primary model for feature selection and interpretability analysis. Within the training set, feature-group combination evaluation, variable selection, and model optimization were performed, and the final model performance was evaluated on the internal test set. The developed model was then directly applied to data from two athletes who were not involved in model development to conduct preliminary athlete-level external consistency validation. Finally, SHAP analysis was applied to the final model to reveal the underlying decision-making factors and characterize their correspondence with expert knowledge. The overall workflow is presented in Fig. 1.

Participants

The study participants were two professional air pistol athletes, including one male and one female athlete. The inclusion criteria were as follows: athletes who had won a championship in the air pistol event at a world-level official competition organized by the International Shooting Sport Federation (ISSF) within the previous five years, or who had achieved a top-three ranking in the air pistol event at a national-level elite competition organized by the National Games of the People's Republic of China or the Chinese Shooting Association. For external consistency validation, two additional air pistol athletes who were not involved in model development were included. All participants received both written and verbal information regarding the study objectives, data collection procedures, and data management protocols. All collected data were treated confidentially and anonymized. The study protocol was approved by the Medical Ethics Committee of Wuhan Sports University (approval no. 2025AFD619). The study was conducted in accordance with the Declaration of Helsinki and all relevant institutional guidelines and regulations, and all all athlete participants, including those in the external validation cohort, provided written informed consent before data collection.

Data Acquisition Equipment

The SCATT laser training system (SCATT Electronic Ltd., Moscow, Russia) was used for both model development and external consistency validation to collect athletes' aiming trajectories. The system recorded the two-dimensional position of the pistol barrel's optical axis within the virtual target coordinate system and divided the trajectory into different phases according to the moment of shot release. The exported data had an interval of approximately 0.01 s between adjacent sampling points, corresponding to a sampling frequency of approximately 100 Hz.

Data Preprocessing

Data for model development were obtained from two athletes. After outlier removal, coordinate normalization, and data quality assessment, a total of 2,857 valid single-shot samples were retained. Data for external consistency validation were collected from two additional athletes and processed using the same procedure. A total of 1,566 shots were included, consisting of 422 reference-labeled good shots and 1,144 reference-labeled non-good shots.

Data extraction and format conversion were performed using the official SCATT-associated script samples.vbs (Copyright © 2002 ZAO Scatt). The script reads raw SCATT data files and extracts the horizontal (X) and vertical (Y) coordinates of the aiming point frame by frame at each time point according to the fixed system sampling frequency. The extracted data are then exported in plain-text format and used as the data source for subsequent statistical analyses.

The X and Y coordinates generated by samples.vbs represent the real-time position of the aiming point of the pistol barrel's optical axis within the SCATT virtual target coordinate system, with units of millimeters (mm). These coordinates differ from the ballistically corrected virtual shot-hole coordinates. In this coordinate system, the center of the 10-ring target is defined as the origin, with the positive direction of the X-axis toward the right side of the target and the positive direction of the Y-axis toward the lower side of the target. Each pair of X and Y coordinates corresponds to a specific sampling time point and can be used to reconstruct the dynamic changes in the aiming point before and after shot release. Because the Y-axis direction of the coordinates exported by the script is opposite to that displayed in the SCATT visualization trajectory plot, the Y-axis coordinates were inverted in this study to ensure consistent coordinate orientation before subsequent analyses.

Explicit Representation of Expert Knowledge

Semi-Structured Interviews

Semi-structured interviews were conducted to establish the criteria for evaluating single-shot execution quality. Before the interviews, the meanings of trajectories displayed in the SCATT visualization interface were consistently explained: the green trajectory represented movement of the pistol barrel from target acquisition to 1.0 s before shot release; the yellow trajectory represented the stable aiming phase from 1.0 s to 0.3 s before shot release; the blue trajectory represented the trigger-release preparation phase from 0.3 s before shot release to the moment of discharge; and the virtual shot location represented the simulated impact position calculated and displayed by the system[1, 2, 4, 5]. The interview participants included two professional coaches and three elite athletes. All five expert participants provided informed consent to participate in the semi-structured interviews. The interviews focused on the characteristics of good shots observed in SCATT trajectory images, with particular emphasis on the concentration of the yellow trajectory, the extent to which the green trajectory surrounded the yellow trajectory, the presence of sudden deviations during the blue trajectory phase, and the relationship between the virtual shot location and the yellow trajectory region.

After the interviews, the researchers organized the collected information and summarized experience-based descriptions into operational decision-making themes based on representative trajectory patterns: gradual stabilization of pistol-barrel movement before shot release, relatively concentrated yellow trajectories, appropriate surrounding of the yellow trajectory by the green trajectory, absence of obvious abrupt deviations during the blue trajectory phase, and virtual shot locations positioned within or close to the stable region of the yellow trajectory. A representative SCATT trajectory is shown in Fig. 2.

Dual-Coach Annotation Procedure

Based on the above decision-making criteria, two professional coaches independently annotated the 2,857 single-shot samples as either “good shots” or “non-good shots.” Informed consent was obtained from both coaches for their participation in the shot-quality annotation and disagreement-review procedures During annotation, the coaches simultaneously reviewed the complete trajectory profile and the SCATT virtual shot location, while the shooting score was not considered as a criterion for label assignment. Samples with inconsistent initial annotations between the two coaches were jointly reviewed to reach consensus labels, and these consensus labels were used for model training. Ultimately, 726 samples were labeled as good shots and 2,131 samples were labeled as non-good shots.

Cohen’s Kappa Agreement Assessment

The agreement between the two coaches’ initial annotations was assessed using Cohen’s kappa coefficient. The kappa coefficient was 0.92, indicating a high level of agreement between the annotations. This result supports the use of expert annotations as reference labels for subsequent model training; however, the classification of borderline samples may still be influenced by coaches’ experience and specific training objectives.

Feature Engineering

Trajectory Segmentation Strategy

The pre-shot trajectory was divided into three phases: the green trajectory, yellow trajectory, and blue trajectory. Features describing stability, surrounding relationships, blue trajectory endpoint position, and the relationship between trajectory characteristics and shot location were extracted from different time windows accordingly. This segmentation strategy ensured that subsequent feature engineering was aligned with the sequence of observations used by coaches when evaluating SCATT trajectory images.

Construction of Candidate Feature Groups

Based on the coordinate-normalized SCATT green, yellow, and blue trajectories, together with the virtual shot location provided by the system, 44 candidate features were constructed and categorized into seven feature groups (A–G) according to the decision-making logic of coaches. The A–G feature groups

corresponded to: (A) the relationship between the shot location and the yellow trajectory; (B) the relationship between the blue trajectory endpoint and the yellow trajectory; (C) yellow trajectory stability; (D) blue trajectory phase movement characteristics; (E) the surrounding relationship between the green and yellow trajectories; (F) shot location characteristics and its relationships with the green and blue trajectories; and (G) data quality and sampling-point-related indicators. Among these groups, feature group G was primarily used to assess trajectory data completeness and was not included in feature-group combination evaluation. Two intermediate coordinate variables, Blue_End_X and Blue_End_Y, were additionally retained in the modeling dataset. These variables were used only for calculating blue trajectory endpoint-related features and were not included among the 44 candidate features or used as final model inputs. The seven feature groups and their definitions are summarized in Table 1; the names, definitions, calculation formulas, and units of the 44 candidate features are provided in Supplementary Table S10.

Table 1
Seven candidate feature groups and definitions

Category	Feature group	Definition
A	Impact point–yellow trajectory relation	Characterizes the location, containment, and distance of the virtual impact point relative to the yellow-trajectory stability region.
B	Terminal blue point–yellow trajectory relation	Characterizes the location, containment, and distance of the terminal blue point relative to the yellow-trajectory stability region.
C	Yellow-phase stability	Characterizes the yellow-phase trajectory's spatial extent, shape, dispersion, drift, speed, and acceleration.
D	Blue-phase motion	Characterizes pre-shot blue-phase direction, directional concentration, stability, displacement, path length, speed, and acceleration.
E	Green–yellow enclosure relation	Characterizes green- and yellow-trajectory coverage, center separation, and the proportion of yellow points inside the green convex hull.
F	Impact position and relations to green/blue trajectories	Characterizes the virtual impact-point coordinates and its positional relations to the green convex hull and terminal blue point.
G	Data quality and sampling-point count	Records valid sampling-point counts in the green, yellow, and blue phases to assess trajectory-data completeness.

Data Standardization and Dataset Partitioning

The 2,857 samples were stratified according to expert-derived labels and divided into a training set (n = 2,285) and a testing set (n = 572). The training set was used for feature selection, model training, and cross-validation, whereas the testing set was reserved for final model performance evaluation. All features were standardized using Z-score normalization based on the mean and standard deviation calculated from the training set. The same normalization parameters were then applied to the testing

set, and no information from the testing set was used during parameter estimation. For binary classification labels, “good shots” were encoded as 1 and “non-good shots” were encoded as 0.

After candidate feature-group construction and dataset partitioning, the analysis was conducted from two complementary perspectives. Feature-group combination evaluation was used to identify a small number of technical information categories that retained the primary discriminative capability and were easier for coaches to interpret. Variable selection was used to determine the specific inputs of the final random forest model and served as the basis for performance evaluation and SHAP interpretation. These two analyses were performed independently. The feature-group combination results did not constrain the scope of variable selection, and the final variables were selected from all candidate features that passed data quality assessment.

Feature-Group Combination Evaluation

Feature-group combination evaluation included the six technical feature groups (A–F), whereas feature group G, consisting of data quality and sampling-point-related indicators, was excluded from combination enumeration and minimal combination selection. Different combinations of technical feature groups were enumerated within the training set, and stratified five-fold cross-validation was used to calculate performance metrics, including ROC-AUC and the F1-score for good shots. Feature-group combinations were selected according to three principles: performance close to the optimum, fewer feature groups, and clearer sport-specific interpretability. The internal test set was not involved in feature-group selection and was used only once to evaluate the selected parsimonious combination.

Variable-Level Feature Selection Procedure

In addition to feature-group combination evaluation, variable-level feature selection was further performed to determine the input variables of the final random forest classification model. First, data quality screening was conducted within the training set to remove features with substantial missing values and constant or near-constant features. Second, absolute coordinate variables and sampling-point-count variables that were not considered core technical descriptors were excluded to reduce the risk that the model learned individual positional preferences or sampling-related variations. Subsequently, permutation importance was calculated using the random forest model within the training set, and features with stable contributions to model discrimination were retained[26]. A Spearman correlation matrix was then calculated among the retained features. Feature pairs with $|r| \geq 0.90$ were considered highly redundant, and when redundancy was identified, the feature with higher model importance and clearer sport-specific interpretation was preferentially retained. Finally, cross-validation within the training set was used to compare the performance of Top-5, Top-10, Top-15, Top-20, Top-25, Top-30, and all candidate feature sets. The final variable set was determined by jointly considering balanced accuracy, F1-score for good shots, PR-AUC, Top-10% precision, the number of features, and the cost of sport-specific interpretation.

Model Development and Validation

Model Benchmarking Matrix and Rationale for Primary Model Selection

To evaluate the ability of different algorithms to learn expert-derived labels using the complete set of 44 candidate features, a multi-model benchmarking matrix was established. The evaluated models included a majority-class baseline model, logistic regression, linear support vector machine (SVM), radial basis function (RBF) kernel SVM, k-nearest neighbors (KNN), a single decision tree, gradient boosting, LightGBM, CatBoost, Explainable Boosting Machine (EBM), and random forest. The multi-model comparison was primarily conducted to determine whether the classification task was learnable and to provide evidence for selecting the primary model. The model with the highest value on a single metric was not considered the sole selection criterion. The random forest model was ultimately selected as the primary model for subsequent variable selection and SHAP interpretation, mainly because it can effectively capture potential nonlinear relationships and feature interactions in trajectory-based data. In addition, it is well suited for heterogeneous features, including distances, areas, ratios, velocities, accelerations, and binary spatial relationship indicators.

Random Forest Parameter Settings and Hyperparameter Tuning Strategy

The random forest model was implemented using the RandomForestClassifier algorithm. To improve the model's ability to identify the minority class of "good shots," the class weight was set to class_weight = "balanced"[27]. The main parameters of the final fixed-parameter model were: n_estimators = 300, max_depth = 8, min_samples_split = 5, min_samples_leaf = 3, max_features = "sqrt", and random_state = 42.

To examine whether hyperparameter optimization could further improve model performance, random search-based hyperparameter tuning was performed within the training set using five-fold stratified cross-validation[28]. The tuning space included parameters related to the number of trees, maximum tree depth, minimum samples required for node splitting, minimum samples required for leaf nodes, maximum feature selection strategy, and splitting criterion. A total of 40 parameter combinations were evaluated through random search. The F1-score for good shots was used as the primary optimization metric, while balanced accuracy, precision, recall, ROC-AUC, and PR-AUC were also recorded. Hyperparameter tuning was conducted exclusively within the training set, and the testing set was not involved in parameter selection. The fixed model settings and tuning overview is provided in Supplementary Table S1.

Stratified Five-Fold Cross-Validation

Model training, feature-group combination evaluation, and Top-K feature selection were all performed within the training set using stratified five-fold cross-validation based on class labels. Each fold maintained a relatively consistent proportion of good and non-good shots. The training folds were used for model fitting, while the validation folds were used to calculate Accuracy, Balanced Accuracy,

Precision, Recall, F1-score for good shots, ROC-AUC, and PR-AUC. Given that good shots accounted for 25.4% of all samples, PR-AUC was additionally reported because it provides a more sensitive evaluation of minority-class recognition performance[29]. All candidate approaches were compared based on the mean results obtained from five-fold cross-validation. After determining the Top-10 features and model parameters, the classification threshold was selected using out-of-fold predicted probabilities generated from five-fold cross-validation on the training set. The threshold was determined primarily by maximizing the F1-score for good shots, with Balanced Accuracy used as a secondary criterion[30–32]. Subsequently, the model was retrained using the complete training set and evaluated once on the internal testing set.

Uncertainty Assessment of Final Model Performance

To assess the uncertainty of performance estimates for the final random forest classification model on the internal testing set, bootstrap confidence intervals were calculated. Within the internal testing set, 2,000 bootstrap resamples were generated with replacement. For each resample, a classification threshold of 0.519 was applied, and Accuracy, Balanced Accuracy, Precision, Recall, F1-score for good shots, ROC-AUC, and PR-AUC were recalculated. The 2.5th and 97.5th percentiles were used to define the 95% confidence intervals for each metric. This procedure was used to evaluate the stability of performance estimates on the internal testing set and was not involved in model training, feature-group combination selection, variable selection, or threshold optimization.

Athlete-Level External Consistency Validation

To evaluate model performance on new athletes, data from two additional air pistol athletes who were not involved in model development were included, comprising a total of 1,566 shots. The external dataset was processed using the feature calculation procedures, missing-value handling rules, and standardization parameters established from the development training set. During validation, the trained Top-10-feature random forest model and the classification threshold of 0.519 were directly applied, without additional feature selection, model retraining, or threshold adjustment. Coaches annotated each shot according to the predefined criteria, and these annotations were used as reference labels. Model performance was evaluated using Accuracy, Balanced Accuracy, Precision, Recall, F1-score for good shots, ROC-AUC, and PR-AUC. In addition, clustered bootstrap resampling was performed 2,000 times using the training data file as the sampling unit to calculate 95% confidence intervals.

SHAP-Based Interpretability Analysis

The final model was interpreted using SHAP (Shapley Additive Explanations). At the global level, the mean absolute SHAP values and feature effects were reported to identify the features that contributed most to model discrimination. At the individual level, predictions were categorized into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) based on the threshold of 0.519. Representative cases from each category were selected to analyze which features drove the model toward the corresponding prediction outcome for individual shots.

Results

Comparison of Baseline Performance Across Multiple Models

The results showed that the majority-class baseline model was unable to identify good-shot samples, with a Balanced Accuracy of 0.5000 and an F1-score for good shots of 0, indicating that the task could not be solved by relying solely on class distribution. Compared with the majority-class baseline, all machine learning models substantially improved the ability to identify good shots. EBM achieved relatively high Balanced Accuracy and ROC-AUC values of 0.8411 and 0.918, respectively. CatBoost achieved the highest F1-score for good shots (0.738), whereas LightGBM achieved the highest PR-AUC (0.809). The random forest model achieved a ROC-AUC of 0.917 and a PR-AUC of 0.803, demonstrating overall stable performance. Complete benchmark results are presented in Supplementary Table S2.

The performance of different models indicates that nonlinear relationships exist between SCATT trajectory features and expert-derived shot quality labels, which can be effectively learned by machine learning algorithms. Although linear models achieved relatively high Recall, their Precision was comparatively lower. Gradient boosting-based models performed strongly on certain metrics but involved greater complexity for subsequent feature selection and interpretation. Random forest did not achieve the best performance on every individual metric; however, it was selected as the primary model because of its stable overall performance and its compatibility with permutation importance-based feature selection and SHAP-based interpretation within a coherent analytical framework.

Systematic Evaluation of Feature Groups

Based on the predefined feature-group framework and combination evaluation strategy, a systematic combination analysis of the six technical feature groups (A–F) was conducted within the training set to identify the most parsimonious feature-group combination capable of reproducing expert decision-making logic. This analysis focused on semantic-level parsimony and did not directly determine specific variable selection. Variable-level selection was performed independently from all qualified candidate features and was used to construct the final model. The data quality indicators in feature group G were used only to describe data completeness and were not included in feature-group combination selection.

AUC Heatmap of Individual and Pairwise Feature-Group Combinations

Figure 3 presents the five-fold cross-validation ROC-AUC values on the training set for models constructed using individual technical feature groups (A–F) and pairwise combinations of these groups. The diagonal elements represent the performance of models based on individual feature groups, whereas the off-diagonal elements represent the performance of models based on pairwise feature-group combinations. Darker colors indicate higher ROC-AUC values. Among individual technical feature

groups, group B, representing the relationship between the blue trajectory endpoint and the yellow trajectory, achieved the best performance, with a ROC-AUC of 0.902.

This indicates that whether the blue trajectory endpoint remains within the stable yellow trajectory region at the moment of shot release is an important source of information for expert identification of good shots. Among pairwise combinations, the B + E combination, consisting of the relationship between the blue trajectory endpoint and the yellow trajectory together with the surrounding relationship between the green and yellow trajectories, achieved the best performance, with a ROC-AUC of 0.9223. This combination included seven features and primarily represented stability maintenance immediately before shot release and the coordination relationship between different aiming phases.

Identification of the Parsimonious Feature-Group Combination

Supplementary Table S3 shows that model ROC-AUC values remained generally high as the number of feature groups increased; however, the additional improvement gained from incorporating more feature groups was limited. The five-group combination A + B+C + E+F achieved the highest ROC-AUC (0.9251), followed by the three-group combination A + B+E (ROC-AUC = 0.9248) and the two-group combination B + E (ROC-AUC = 0.9223). Compared with the best-performing combination, the B + E combination showed only a 0.0028 decrease in ROC-AUC, while reducing the number of features from 29 to 7 and substantially decreasing interpretability complexity.

Regarding the F1-score for good shots, the four-group combination A + C+E + F achieved the highest value (0.7464), while the B + E combination achieved 0.7428, showing only a small difference. Because the B + E combination maintained strong discriminative performance while involving fewer feature groups with clearer technical meanings, it was selected as the most parsimonious feasible combination at the technical feature-group level and was evaluated once on the internal testing set. The evaluation results showed an Accuracy of 0.8339, Balanced Accuracy of 0.8045, Precision of 0.6506, Recall of 0.7448, F1-score of 0.6945 for good shots, ROC-AUC of 0.9043, and PR-AUC of 0.7503.

The B + E combination was used to represent the most parsimonious feature-group combination at the expert semantic level. The final random forest model still selected its input variables from all qualified candidate features based on permutation importance, correlation redundancy analysis, and Top-K sensitivity analysis. Subsequently, SHAP was applied to interpret the contribution of these variables within the final model.

Results of Variable-Level Feature Selection

Beyond feature-group combination evaluation, variable-level feature selection was further conducted to determine the input variables of the final random forest model. The procedure consisted of data quality assessment, removal of non-core variables, random forest permutation importance-based screening, correlation redundancy control, and Top-K feature-number sensitivity analysis. The entire procedure was

performed exclusively within the training set, and the testing set was not involved in feature selection. The complete workflow is summarized in Supplementary Table S4.

Data quality assessment showed that none of the candidate variables exhibited substantial missingness. However, N_Blue_Points was a constant variable in the training set and provided no discriminative information; therefore, it was removed. Subsequently, non-core variables including Impact_X, Impact_Y, N_Green_Points, and N_Yellow_Points were removed. These variables primarily represented absolute spatial coordinates or sampling counts and could be influenced by individual aiming preferences, coordinate calibration differences, or data recording variations. Therefore, they were not included as inputs for the final random forest model.

After the initial model-based importance screening, Spearman correlation coefficients were calculated among the 37 retained candidate features. When the absolute correlation coefficient between two features was ≥ 0.90, the two features were considered to contain highly redundant information. In such cases, the feature with higher random forest permutation importance and clearer sport-specific meaning was preferentially retained. The removed features are summarized in Supplementary Table S5.

After correlation redundancy control, 31 candidate features remained. These features were further ranked according to random forest permutation importance, and the performance of models with different Top-K feature numbers was compared using cross-validation within the training set. The results showed that after increasing the number of features to Top-10, further inclusion of additional features resulted in only limited performance improvements despite slight fluctuations in some metrics. For example, the Top-25 model achieved a Balanced Accuracy of 0.8471 and an F1-score for good shots of 0.7474; however, the improvement compared with Top-10 was marginal, while model complexity and interpretation cost increased. Therefore, considering model performance, the number of input features, and sport-specific interpretability, the Top-10 features were selected as the final input variables for the primary model. Cross-validated results across feature counts are presented in Supplementary Table S6.

Random Forest Threshold Analysis

A random forest model was trained using the Top-10 features identified in Supplementary Table S7 with the fixed parameter settings. After evaluating all valid threshold values based on out-of-fold prediction probabilities from the training set, the highest F1-score for good shots was achieved at a threshold of 0.519, reaching 0.7524, with a Balanced Accuracy of 0.8455. Using this threshold, the model achieved an Accuracy of 0.8671, Balanced Accuracy of 0.8450, Precision of 0.7117, Recall of 0.8000, F1-score of 0.7532 for good shots, ROC-AUC of 0.9175, and PR-AUC of 0.7836 on the testing set. The corresponding confusion matrix values were TN = 380, FP = 47, FN = 29, and TP = 116. The out-of-fold performance results under representative thresholds on the training set are presented in Supplementary Table S8. The random-search-tuned model achieved a good-shot F1-score of 0.7498 and a Balanced Accuracy of 0.8466 during training-set cross-validation. On the internal testing set, it achieved an Accuracy of 0.8566,

Balanced Accuracy of 0.8402, F1-score for good shots of 0.7405, ROC-AUC of 0.9164, and PR-AUC of 0.7791.

Final random-forest test performance

Using the ten selected features and the 0.519 threshold fixed from training-set out-of-fold predictions, the final random forest achieved an accuracy of 0.8671, balanced accuracy of 0.8450, good-shot precision of 0.7117, recall of 0.8000, good-shot F1 of 0.7532, ROC-AUC of 0.9175 and PR-AUC of 0.7836 on the internal test set. Bootstrap 95% CIs are reported in Table 2.

Table 2
Final random-forest test performance with bootstrap 95% confidence intervals

Metric	Estimate	95% CI
Accuracy	0.8671	0.8374–0.8951
Balanced Accuracy	0.8450	0.8079–0.8806
Good-execution precision	0.7117	0.6400-0.7799
Good-execution recall	0.8000	0.7319–0.8652
Good-execution F1	0.7532	0.6953–0.8027
ROC-AUC	0.9175	0.8913–0.9403
PR-AUC	0.7836	0.7149–0.8448

Ranking-based Top-K analysis

Good-execution probabilities can also rank shots for prioritized training review. In the internal test set, 93.10% of the highest-ranked 5% of shots were good executions; the corresponding proportions were 82.76% for the top 10% and 74.78% for the top 20%. Thus, beyond binary classification, the model could help coaches locate high-probability good executions within large training sets for exemplar-based review. Detailed ranking results are reported in Supplementary Table S9.

Athlete-level external consistency

The external analysis included 1,566 shots: 422 good and 1,144 non-good reference labels. The model agreed with the labels for 1,372 shots, yielding accuracy of 0.8761, balanced accuracy of 0.8187, good-shot precision of 0.8184, recall of 0.6943, good-shot F1 of 0.7513, ROC-AUC of 0.9349, and PR-AUC of 0.8413. The confusion matrix was TN = 1,079, FP = 65, FN = 129 and TP = 293. Cluster-bootstrap 95% CIs for accuracy, balanced accuracy, precision, recall, F1, ROC-AUC and PR-AUC were 0.8644–0.8892, 0.7945–0.8445, 0.7701–0.8579, 0.6407–0.7520, 0.7162–0.7837, 0.9234–0.9460 and 0.8093–0.8705, respectively.

Athlete 1 contributed 929 shots (270 good and 659 non-good), with TN = 631, FP = 28, FN = 93, and TP = 177. Athlete 2 contributed 637 shots (152 good and 485 non-good), with TN = 448, FP = 37, FN = 36, and TP = 116. ROC-AUC was approximately 0.937 for both athletes, whereas good-execution recall was 0.6556 and 0.7632, respectively (Table 3).

Table 3
Athlete-level external consistency of the final random forest

Data source	Shots	Accuracy	Balanced accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Overall	1566	0.8761	0.8187	0.8184	0.6943	0.7513	0.9349	0.8413
Athlete 1	929	0.8698	0.8065	0.8634	0.6556	0.7453	0.9370	0.8638
Athlete 2	637	0.8854	0.8434	0.7582	0.7632	0.7607	0.9373	0.8258

Global and local SHAP explanations

Global feature importance

Global SHAP analysis showed that Impact_To_YellowHull_SignedDistance and Impact_To_Yellow_Center had the largest contributions, accounting for approximately 60.70% of the total contribution. This indicates that the spatial relationship between the shot location and the yellow trajectory stability region was the most important basis for model discrimination. Blue_End_Inside_YellowHull ranked third, indicating that whether the blue trajectory endpoint remained within the yellow trajectory stability region at the moment of shot release was also an important source of information. The top five features accounted for 86.09% of the total contribution. The SHAP beeswarm plot further illustrated how different feature values influenced prediction direction. Higher Impact_To_YellowHull_SignedDistance values and a value of 1 for Blue_End_Inside_YellowHull were generally associated with positive SHAP values. Lower values of Impact_To_Yellow_Center and Green_Yellow_Center_Distance also primarily contributed to predictions of good shots. These findings indicate that a shot location close to the yellow trajectory stability region, a blue trajectory endpoint remaining within this region, and effective coordination between the green and yellow trajectories were the primary factors supporting the model's identification of good shots. The global importance and beeswarm distributions are shown in Fig. 4.

Individual SHAP Analysis of Representative Cases

To further illustrate the decision-making logic of the model at the single-shot level, test-set predictions were categorized using the threshold of 0.519, and representative true positive (TP), true negative (TN), false positive (FP), and false negative (FN) cases were selected for individual SHAP waterfall plots. The results showed that, for correctly identified good-shot samples, Impact_To_YellowHull_SignedDistance, Impact_To_Yellow_Center, and Blue_End_Inside_YellowHull provided the major positive contributions to

the model output, increasing the predicted probability of a good shot from the baseline value of 0.498 to 0.977. This finding indicates that when the shot location has a favorable spatial relationship with the yellow trajectory stability region and the blue trajectory endpoint remains within this region, the model tends to classify the shot as a good shot. Representative waterfall plots are shown in Fig. 5.

For correctly identified non-good-shot samples, the aforementioned spatial relationship features contributed negatively to the model output, reducing the predicted probability of a good shot to 0.001. This indicates that the model could identify non-good shots based on inconsistent relationships between shot location and the yellow trajectory stability region, as well as unfavorable blue trajectory endpoint positioning relative to the yellow trajectory. Misclassified samples further revealed the decision boundaries of the model.

For samples with high-confidence false-positive predictions, the model was still strongly influenced by the positive contributions of Impact_To_YellowHull_SignedDistance, Impact_To_Yellow_Center, and Blue_End_Inside_YellowHull, resulting in a predicted probability of 0.963 for a good shot. This suggests that although the sample exhibited a feature pattern close to that of good shots in the spatial relationships emphasized by the model, experts may have classified it as a non-good shot based on more subtle trajectory characteristics or shot execution process information. For false-negative samples, these core spatial relationship features showed negative contributions, reducing the predicted probability of a good shot to 0.010. This suggests that the model may rely heavily on the consistency between shot location and the yellow trajectory stability region, while insufficiently capturing some process-based judgments incorporated into expert evaluations.

Coach-readable interpretation rules

The SHAP results can be translated into four coach-readable review cues. An impact point within or near the stable yellow region favors a good-execution prediction. A terminal blue point that remains within that region indicates that the muzzle did not depart markedly from the stable area near firing. A small distance between green- and yellow-line centers indicates a coherent transition into the final stable region. Low dispersion in the final pre-shot window indicates better stability as trigger release approaches.

Consistency between model explanations and expert criteria

Comparison of the global and local SHAP results with the interview-derived criteria showed broad descriptive agreement. The highest-contributing features concerned the impact-to-yellow relation, terminal-blue-to-yellow relation, green-to-yellow enclosure, and pre-shot stability, mirroring the experts' emphasis on a compact yellow line, coherent green-line enclosure, absence of a marked blue-line disruption, and an impact point within or near the stable yellow region.

More specifically, Impact_To_YellowHull_SignedDistance and Impact_To_Yellow_Center represented the spatial relation between the impact point and the stable yellow region; Blue_End_Inside_YellowHull indicated whether the terminal blue point remained within that region; Green_Yellow_Center_Distance and Yellow_Points_Inside_GreenHull_Ratio described green-to-yellow alignment and enclosure; and Steady_Std_500ms quantified stability in the final pre-shot interval. These features closely corresponded to the cues used in SCATT-based coaching review.

Discussion

SCATT trajectory features contained learnable information associated with expert-defined execution-quality labels: EBM, CatBoost, LightGBM, and random forest all achieved strong discrimination. This extends prior use of machine learning for sports performance and risk monitoring[6–8] by targeting coaches' judgments of the shot process rather than score alone. The result is technically plausible because aiming accuracy, holding stability, trigger control, timing, and aiming-point trajectories are established determinants of shooting performance[1–3, 5]. Random forest was not the best model on every metric—EBM had strong balanced accuracy and ROC-AUC, CatBoost a higher good-execution F1, and LightGBM the highest PR-AUC—but it offered stable overall performance, nonlinear interaction modeling[22], and direct integration with permutation importance and SHAP. SHAP provides both global and individual explanations[25], making random forest a practical primary model for feature selection and interpretation.

The final random forest model obtained through variable-level feature selection achieved a Balanced Accuracy of 0.8450, an F1-score of 0.7532 for good shots, and a ROC-AUC of 0.9175 on the internal testing set, indicating that a small number of key features retained sufficient capability to reproduce expert judgments after training-set-based feature selection. These findings are consistent with previous studies that used optoelectronic training systems to quantify aiming stability, trajectory fluctuations, and key technical factors in shooting performance[1, 2, 4, 5]. Nevertheless, model outputs should still be regarded as auxiliary information for expert-based post-shot analysis[33]. Particularly for borderline cases and high-confidence misclassified samples, coaches should incorporate training context, daily performance status, and detailed technical observations for further review.

Group B, the terminal-blue-to-yellow relation, was the strongest individual semantic group. This supports the coaching emphasis on whether muzzle position remains within the stable yellow region immediately before firing. Prior air-pistol and air-rifle research has repeatedly linked aiming accuracy, hold stability, and trigger control with performance[1–3], and the aiming-point trajectory itself can quantify aiming stability[5]. The advantage of group B is therefore technically plausible rather than merely statistical.

The B + E combination retained near-optimal cross-validated performance with only seven features and maintained useful discrimination on the internal test set. These findings suggest that coaches consider not only the final impact point but also whether the muzzle remains within the stable yellow region at

release and whether the earlier green trajectory encloses and connects smoothly with that region. Group B represents stability near trigger release, whereas group E represents convergence from a broader aiming phase into the final stable region. Their combination captures both immediate control and process convergence, consistent with prior descriptions of stability, aiming accuracy, and firing timing as key technical components[2, 3].

The feature-group and SHAP results address different questions and are not contradictory. Group analysis tested whether a small set of coach-defined semantic groups could discriminate the labels. SHAP explained the contribution of individual variables in the final random forest. Impact-to-yellow variables dominated the final model, whereas B + E showed that terminal-blue and green-to-yellow relations alone still retained much of the expert decision structure.

B + E should therefore not replace the final model; its value is to identify a minimal coach-readable representation. In practical review, coaches may first ask whether the terminal blue point remains within the stable yellow region and whether the green line forms a coherent enclosure and transition around the yellow line. Shots satisfying both conditions are more likely to match the expert definition of a high-quality execution.

Without retraining or threshold adjustment, the model achieved balanced accuracy of 0.8187, good-execution F1 of 0.7513, and ROC-AUC of 0.9349 against coach labels in two additional athletes. The principal spatial relations learned during development therefore continued to reproduce the coach-assigned reference labels in these athletes. At the fixed 0.519 threshold, external data showed lower recall and higher precision, which may reflect differences in class prevalence, athlete characteristics, or feature distributions. Because only two additional athletes were studied, these results do not constitute strong evidence of population-level cross-athlete generalization.

The principal limitation is the small number of athletes: two contributed to development and two to external consistency analysis. A large number of repeated shots from the same athlete does not provide the same inferential breadth as a large athlete sample. The external results show consistency only in the two included athletes and may not extend to other training environments, devices, or athlete populations. Future studies should repeat feature selection, tuning, and threshold selection within each training fold, apply athlete- or session-grouped nested cross-validation, and evaluate the locked workflow in independent multicenter cohorts[34–38]. In addition, independent impact-blind labeling would help separate process quality from virtual impact-point information embedded in the current expert reference.

Conclusion

This study developed an interpretable classifier of single-shot trigger-execution quality from SCATT trajectories and coach labels. The terminal-blue-to-yellow relation was the most informative individual semantic group, and its combination with green-to-yellow enclosure formed a parsimonious representation. The final Top-10 random forest achieved balanced accuracy of 0.8450, good-execution

F1 of 0.7532, and ROC-AUC of 0.9175 on the internal test set; the corresponding performance values in two additional athletes were 0.8187, 0.7513, and 0.9349. SHAP explanations showed descriptive agreement with the coach-derived criteria. The approach can support technically grounded training review, but broader claims require a larger athlete sample and independent, multicenter validation.

Declarations

Competing interests

The authors declare no competing interests.

Author Contribution

Huiying Feng: Model development, experimental work, data analysis, and manuscript preparation; Tong Zhou: Research design, provision of resources, and manuscript revision; Shanrui Diao: Experimental work and data collection; Yunyun Du: Research design, supervision of the manuscript, and manuscript revision. Huiying Feng and Tong Zhou made equivalent contributions and are listed as co-first authors. All authors reviewed the manuscript and approved the final version.

Acknowledgement

The authors would like to thank the professional coaches and elite athletes who participated in the expert interviews and shot-quality annotation process for their valuable time, expertise, and constructive feedback, which contributed to the development of the evaluation criteria used in this study.

Data Availability

Non-identifying data can be shared upon reasonable request to the corresponding author

Funding

This research was funded by Hubei Provincial Natural Science Foundation Key Project: “Research on a Prediction Model for Shooting Trigger Timing Based on Biofeedback and Machine Learning” (Project Number: 2025AFD619).

References

1. Mon-López, D. et al. Optoelectronic analysis of technical factors and performance of elite-level air pistol shooting. *PLoS ONE*. **17**, e0262276 (2022).

2. Olsson, E. & Laaksonen, M. S. Key technical components for air pistol shooting performance. *Int. J. Perform. Anal. Sport*. **21**, 348–360 (2021).
3. Sundaram, V., Sundar, V. & Middleton, K. Technical determinants of air rifle and pistol shooting performance: a systematic review and meta-analysis. *Int. J. Sports Sci. Coach*. **19**, 1844–1862 (2024).
4. Bale, J. H. & Wilkinson, M. Validity and reliability of an opto-electric training system in elite and national level ISSF air rifle shooters. *Sports Eng.* **26**, 27 (2023).
5. Zanevskyy, I., Korostylova, Y. & Mykhaylov, V. Aiming point trajectory as an assessment parameter of shooting performance. *Hum. Mov.* **13**, 211–217 (2012).
6. Taber, C. B. et al. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Sci. Rep. 14, 1162 (2024).
7. Claudino, J. G. et al. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports Med. Open.* **5**, 28 (2019).
8. Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R. & Tischer, T. Machine learning methods in sport injury prediction and prevention: a systematic review. *J. Exp. Orthop.* **8**, 27 (2021).
9. Ball, K., Best, R. & Wrigley, T. Body sway, aim point fluctuation and performance in rifle shooters: inter- and intra-individual analysis. *J. Sports Sci.* **21**, 559–566 (2003).
10. Jin, R. et al. Unstable trunk core strength training improves shooting performance and holding stability in elite 10 m air pistol athletes: a pilot randomized controlled trial. *Sci. Rep.* **16**, 23617 (2026).
11. Hawkins, R. & Bertrand, P. Relationship between twelve mechanic measures and score for national-level pistol shooters. *Int. J. Perform. Anal. Sport*. **15**, 332–342 (2015).
12. Hawkins, R. N. & Sefton, J. M. Effects of stance width on performance and postural stability in national-standard pistol shooters. *J. Sports Sci.* **29**, 1381–1387 (2011).
13. Ihalainen, S., Kuitunen, S., Mononen, K. & Linnamo, V. Determinants of elite-level air rifle shooting performance. *Scand. J. Med. Sci. Sports*. **26**, 266–274 (2016).
14. Ihalainen, S., Mononen, K., Linnamo, V. & Kuitunen, S. Which technical factors explain competition performance in air rifle shooting? *Int. J. Sports Sci. Coach*. **13**, 78–85 (2018).
15. Lang, D. & Zhou, A. Relationships between postural balance, aiming technique and performance in elite rifle shooters. *Eur. J. Sport Sci.* **22**, 1493–1498 (2022).
16. Liu, Y., Hu, N., Sun, M., Qu, F. & Zhou, X. The effects of hand tremors on the shooting performance of air pistol shooters with different skill levels. *Sensors* **24**, 2438 (2024).
17. Mononen, K., Konttinen, N., Viitasalo, J. & Era, P. Relationships between postural balance, rifle stability and shooting accuracy among novice rifle shooters. *Scand. J. Med. Sci. Sports*. **17**, 180–185 (2007).

18. Bale, J. H. et al. An ecological dynamics interpretation of match-wide performance outcomes in elite ISSF air rifle shooting competitions. *Int J. Sports Sci. Coach* (2026).
19. Diao, S., Zhou, T. & Du, Y. Personalized training model for 10 m air pistol through machine learning: a pilot study. *Sci. Rep.* **16**, 19971 (2026).
20. Flyckt, J., Andersson, F., Westphal, F., Månsson, A. & Lavesson, N. Explaining rifle shooting factors through multi-sensor body tracking: using transformers and attention to mine actionable patterns from skeleton graphs. *Intell. Data Anal.* **27**, 535–554 (2023).
21. Guo, J. et al. A random forest-based accuracy prediction model for augmented biofeedback in a precision shooting training system. *Sensors* **20**, 4512 (2020).
22. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
23. Barredo Arrieta, A. B. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*. **58**, 82–115 (2020).
24. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
25. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
26. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).
27. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
28. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev. Data Min. Knowl. Discov*. **9**, e1301 (2019).
29. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. **10**, e0118432 (2015).
30. Hernández-Orallo, J., Flach, P. & Ferri, C. A unified view of performance metrics: translating threshold choice into expected classification loss. *J. Mach. Learn. Res.* **13**, 2813–2869 (2012).
31. Thiele, C. & Hirschfeld, G. cutpointr: improved estimation and validation of optimal cutpoints in R. *J. Stat. Softw.* **98**, 1–27 (2021).
32. Shahare, P., Mahalwar, A. & Shahade, A. K. A distributed fusion framework for breast cancer recurrence prediction using MapReduce. *Sci. Rep.* **16**, 16316 (2026).
33. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health*. **3**, e745–e750 (2021).
34. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).
35. Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

36. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS ONE*. **14**, e0224365 (2019).
37. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 91 (2006).
38. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit. Med.* **5**, 48 (2022).

Figures

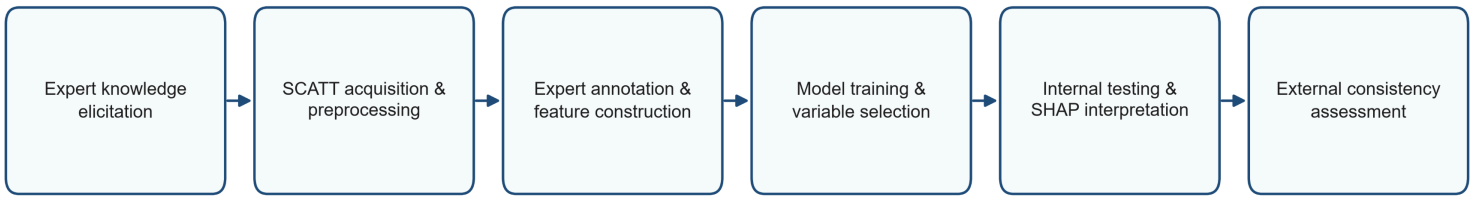

Figure 1

Study design and analysis workflow

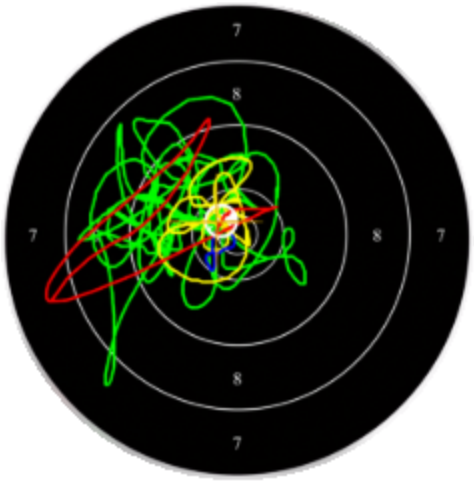

Figure 2

Example of a single-shot SCATT aiming trajectory

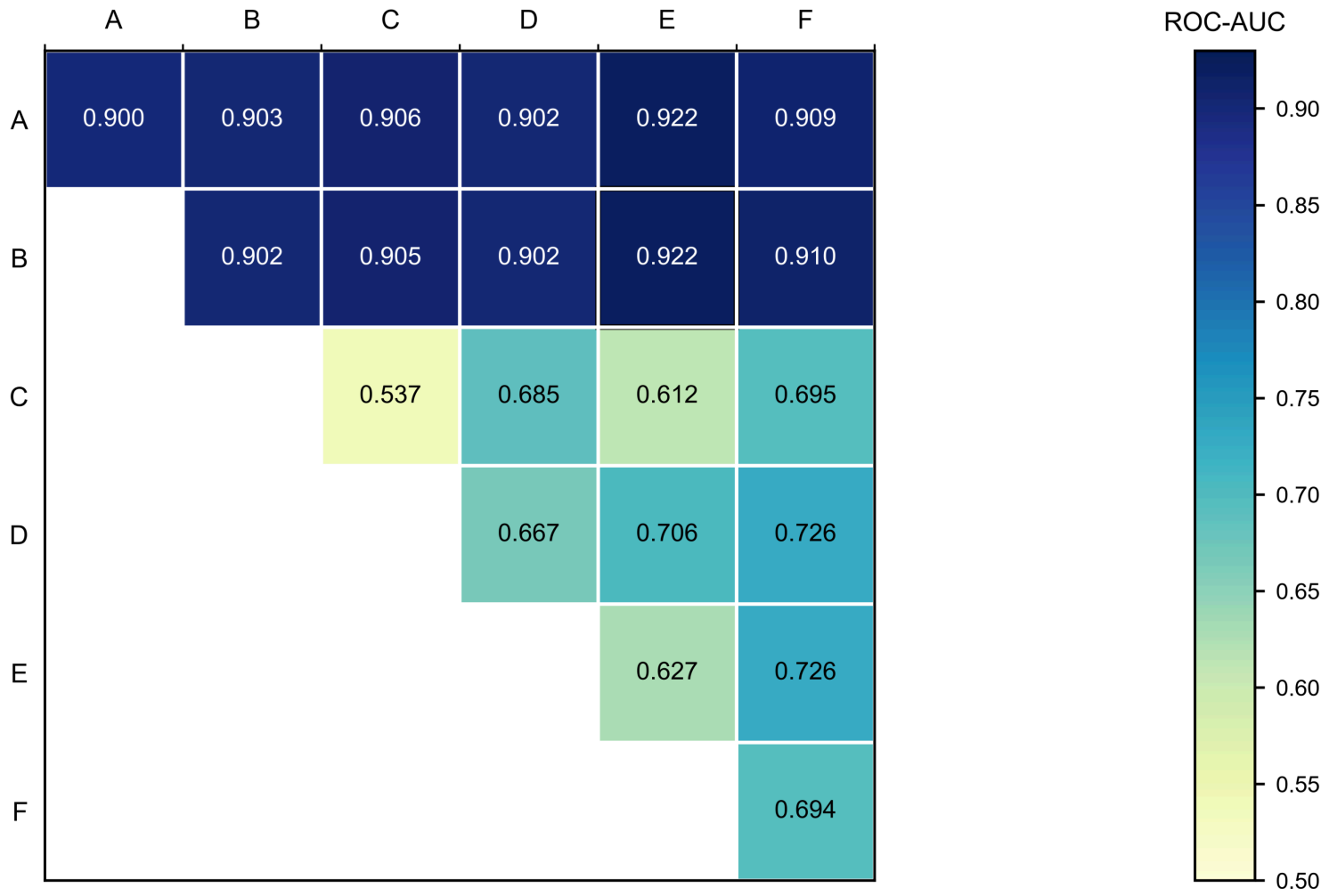

Figure 3

Cross-validated ROC-AUC heat map for single feature groups and pairwise combinations

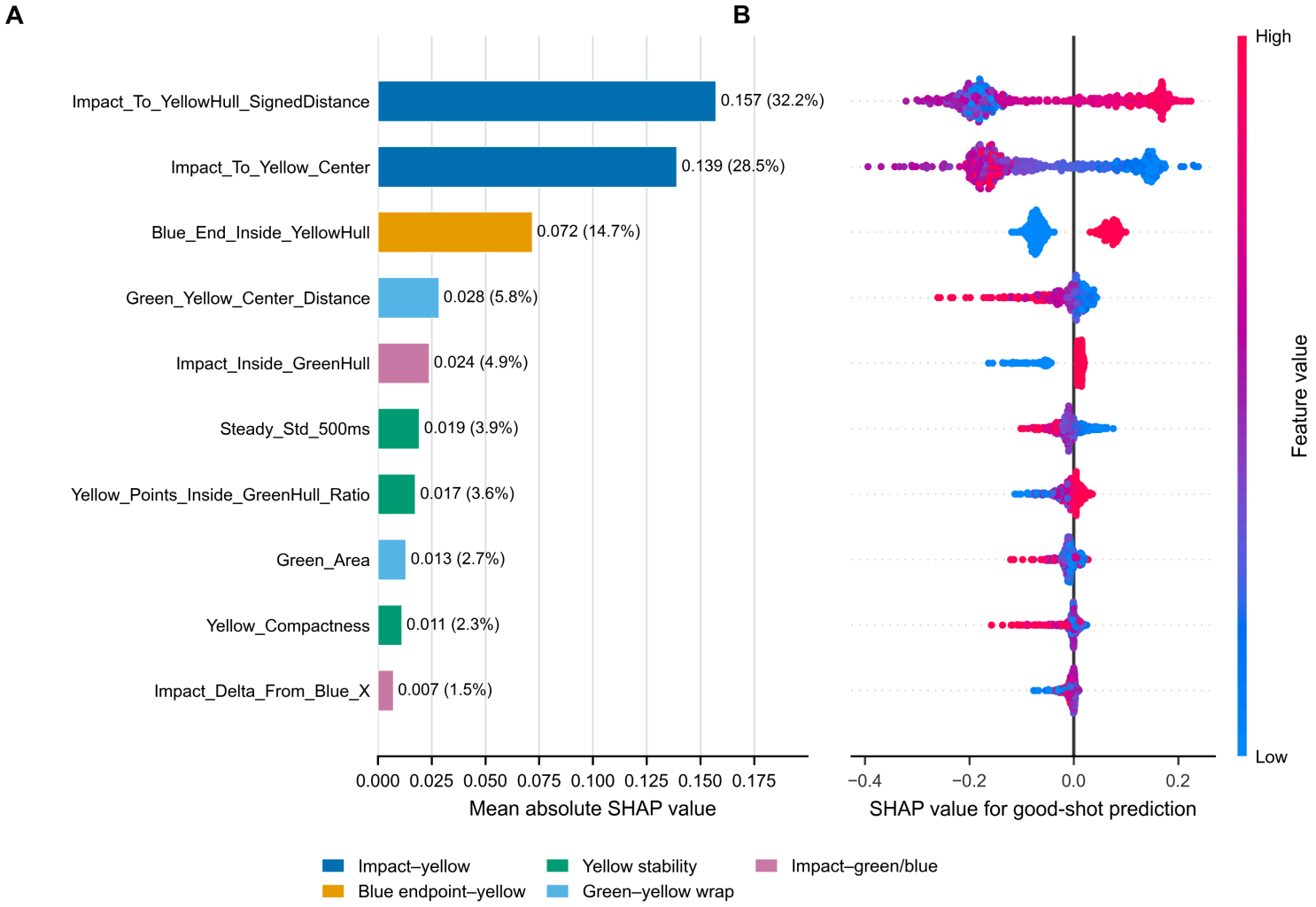

Figure 4

Global SHAP explanations: (a) mean absolute SHAP importance and (b) beeswarm distribution.

Note: Bar length denotes mean absolute SHAP value; percentages denote each feature's contribution to the total importance of the Top-10 features.

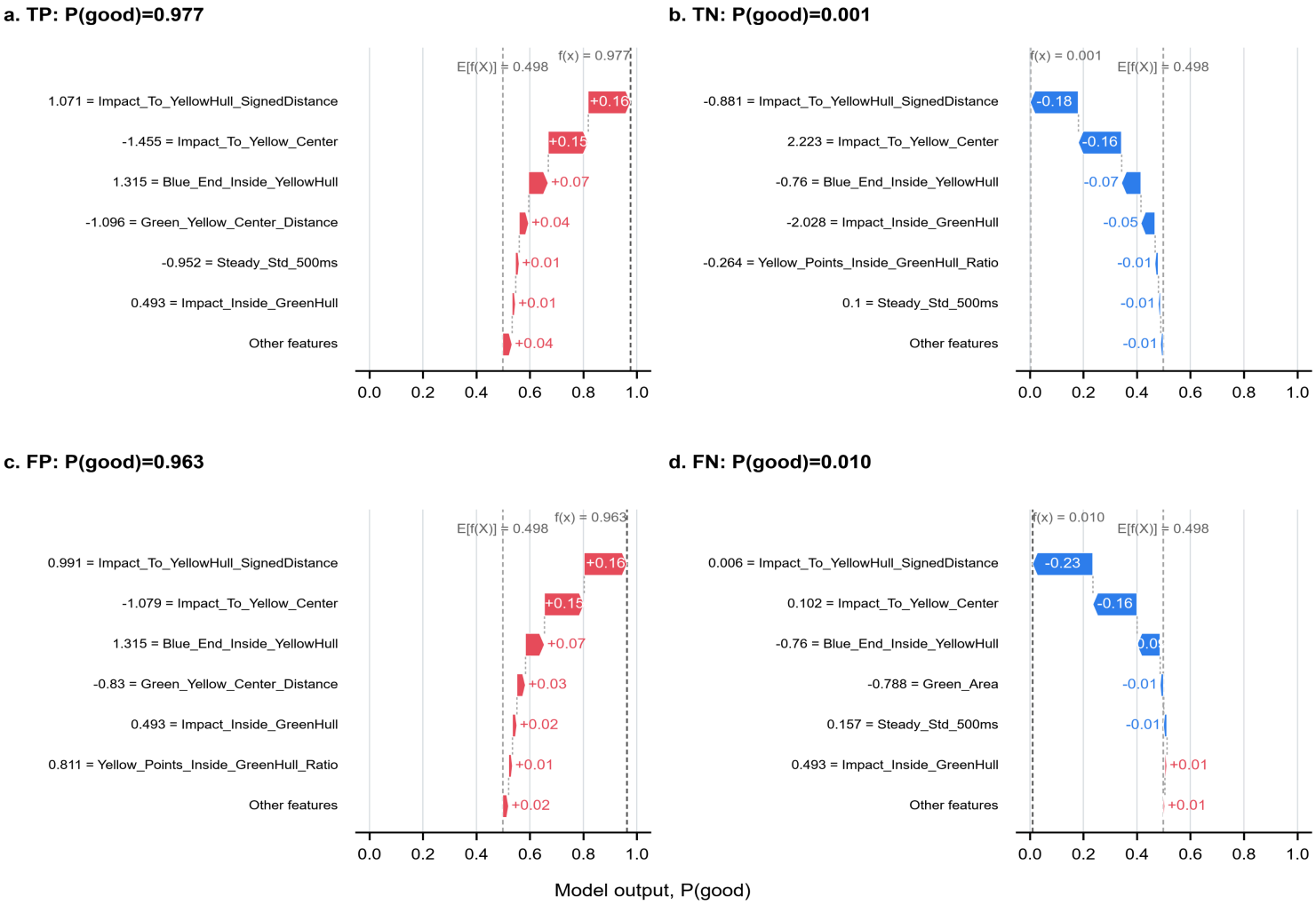

Figure 5

Local SHAP explanations for representative (a) TP, (b) TN, (c) FP and (d) FN cases

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx