Interpretable Prediction and Large-Scale Analysis of Judging in Professional Boxing

Paper Track: Other Sports
Paper ID: 145

1. Introduction

In professional boxing, nearly half (46.5%) of matches are decided by judges' scorecards rather than knockout or stoppage [1]. In scoring each round, judges are instructed to consider four factors: clean and effective punching, effective aggression, ring generalship, and defense [2]. However, the definitions of these terms are vague, and guidelines provide little guidance on how to weigh these factors relative to one another. This stands in stark contrast to judged sports like gymnastics [3], figure skating [4], and freestyle skiing [5], which rely on highly codified scoring systems with clearly defined skills, difficulty ratings, and deductions. With so much subjectivity in boxing's scoring guidelines it isn't surprising to see that the sport is rife with judging controversy. Many fans and athletes express frustration and mistrust in judging. Indeed, suspicions of corruption and partisanship have long followed the sport [6].

This ambiguity in scoring criteria has contributed to a general lack of interpretability as researchers and performance analysts struggle to identify the most important aspects of a boxer's performance, limiting athletes' potential for improvement. Existing research identifying the most important factors of a boxing performance has relied on manually annotating fight footage, severely limiting sample sizes. Studies typically analyze fewer than 50 bouts, restricting researchers' ability to identify subtle judging patterns or develop robust predictive models [7, 8].

Large-scale repositories of combat sports statistics exist, though they come with considerable limitations. Since the 1980s, CompuBox has been the standard for professional boxing statistics, relying on two trained operators with clickers to count thrown and landed punches in real time. However, all landed punches are categorized as either a "jab" or a "power punch" with no further differentiation; a glancing blow that barely connects is counted identically to a clean, powerful shot that visibly staggers an opponent [9]. This crude binary classification fails to capture the qualitative differences that are central to judges' determination of "clean and effective punching" [2]. Furthermore, CompuBox doesn't track positional data or contextual information such as whether a punch was thrown as part of a combination, while moving forward, or with the opponent on the ropes, critical details for assessing ring generalship and effective aggression [10]. These shortcomings, combined with documented concerns about the accuracy of the real-time clicker-system [11], have led analysts to caution against relying solely on CompuBox for fight evaluation [10]. Consequently, most academic researchers have opted for custom stat-tracking methodologies driven by manual annotation. A similar clicker-based system is used in the Ultimate Fighting Championship (UFC). Though post-fight review provides a layer of validation, the UFC system also struggles to differentiate between strike types, relying on a subjective definition of "significant" versus "non-significant" strikes [12, 13]. Nevertheless, UFC Fight Stats have proven useful for academic research. James et al. (2016) used 11 metrics from UFC Fight Stats to analyze winning performances, providing a valuable framework for determining the relative importance of performance metrics in combat sports, though relying on overall fight statistics rather than round-by-round breakdowns [14].

Computer vision systems trained on professionally annotated data can gather fight statistics on a far larger scale than would be feasible with manual annotation alone. With larger samples, we can analyze trends in judging data to uncover what, in practice, judges are taking into account when making their decisions, a critical first step toward addressing scoring inconsistencies. Other sports have demonstrated the value that computer-vision systems can bring to officiation: gymnastics' Judging Support System uses 3D pose estimation to verify technical elements and improve scoring consistency [3, 15, 16], while soccer's Video Assistant Referee (VAR) increased the accuracy of offsides decisions from 92.1% to 98.3% across 2,195 matches [17].

In recent years, computer vision systems have been introduced to boxing through two primary approaches:

a) Most systems use pose-estimation, overlaying skeletal models onto images of athletes, before classifying actions. While widely used in sports like baseball and gymnastics [18], these systems struggle with combat sports due to occlusion issues [19]. 3D pose estimation can mitigate occlusion but requires multiple calibrated camera perspectives [20], making deployment impractical in many settings. Pose-estimation accuracy and temporal resolution requirements are particularly demanding in sports requiring fine-grained action classification. Research in baseball has shown that while pose estimation can capture general throwing mechanics, subtle variations that differentiate pitch types remain difficult to reliably detect [21]. Punch tracking in boxing presents analogous challenges, as differentiating punch quality and power requires capturing subtle kinematic details that may be lost with insufficient pose estimation accuracy.
b) Action-recognition models built on raw video data have shown considerable accuracy improvements over pose-based systems in combat sports [22]. DeepStrike, a system designed for combat sports, has a neural network trained to derive patterns from full images rather than just skeletal data [23]. From boxing footage, DeepStrike outputs 47 distinct metrics per fighter per round, from punch cleanness and power to stance time and positional data (see Appendix Table 7). While used to generate live stats for professional broadcasts [24, 25, 26], DeepStrike has not yet been applied in academic performance analysis.

DeepStrike enables us to generate detailed statistics on thousands of rounds of boxing with ease. We want to investigate the use of gradient descent and neural networks to map DeepStrike's statistics onto judges' scores. In doing so, we aim to construct a point-style scoring system that is both transparent and interpretable, while achieving accuracy comparable to professional judges using the traditional 10-point must system.

We also use gradient descent, logistic regression and correlation analysis to determine which aspects of a performance most strongly influence judges' decisions. In analyzing judges' decision-making, we hypothesize that:

1) Differentiating punches by impact is very important. Judging models that differentiate punches by impact will significantly outperform models that rely solely on the number of landed punches.

2) Pressure and aggression indicators will emerge as particularly important factors, ranking within the top half of features when assessed through correlation analysis and feature selection.

2. Methods

2.1 Sample

HD videos of professional boxing matches were sourced from YouTube and Dailymotion, excluding videos with artifacts or upscaled standard definition footage [27]. After removing duplicates from 5,427 initial videos, 5,348 unique matches remained. Of these, 1,003 bouts (19%) had complete round-by-round scorecard data available from BoxRec [28], totaling 7,864 rounds.
The dataset consists primarily of men's matches (96.7%) with standard 3-minute rounds. The remaining 3.3% are women's matches, all featuring 2-minute rounds.

A distribution of all scores in the dataset revealed that 73.1% of rounds (5,745) were unanimous decisions where all three judges agreed on the winner, while 26.9% (2,119) were split decisions. The vast majority of rounds (93.1%) feature only 10-9 and 9-10 scores. Most other scores result from knockdowns or point deductions, which DeepStrike is not trained to identify. While 10-10 and 10-8 scores can occur without knockdowns or deductions, they are rare [2, 29]. So, for the sake of simplicity, the remaining 6.8% of rounds were excluded from analysis, resulting in a final sample of 7,323 rounds with binary scoring.

2.2 Procedure

All data processing, modeling, and analysis was conducted using Python with the numpy, pandas, scikit-learn, and TensorFlow libraries. Analysis code, model weights and anonymized data are available open-source on GitHub [31].

Data Collection: We manually identified HD videos and cross-referenced BoxRec for the corresponding scorecards. Videos were downloaded in full resolution and uploaded to the Jabbr closed beta web-application to which we were granted access. Each upload required manual input of bout metadata and anchor image tagging for fighter identification. Jabbr's annotation team then marked round boundaries and verified video quality before the DeepStrike computer vision system generated round-by-round performance metrics.

Data Validation: Multiple validation steps ensured data quality: automated removal of duplicate bouts, manual review of video-scorecard pairing (with particular attention to repeat match-ups), anchor image verification during tagging, analyst review of video quality, and post-analysis flagging of statistical outliers indicating potential pairing errors.

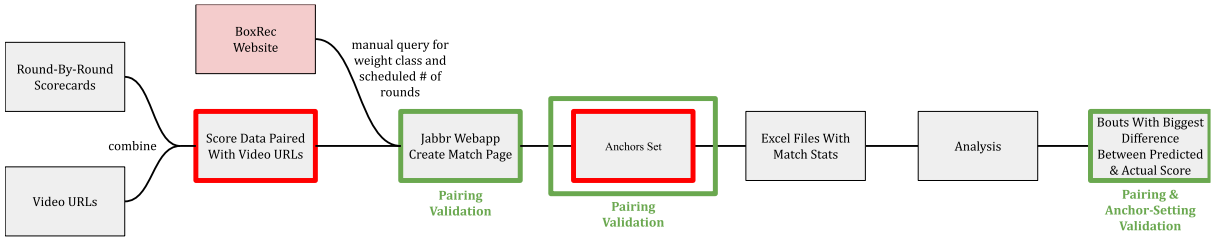

Figure 1: *Data validation process. Steps where labeling errors can be introduced are outlined in red. Validation steps are outlined in green.*

2.3 Material and Measures

Performance metrics were generated using DeepStrike, an action-recognition computer vision system for combat sports [23], trained on 194 fully annotated bouts (171,301 punches). DeepStrike generates metrics at two levels: individual punch characteristics and round-level aggregate statistics. Each punch thrown is characterized by the following attributes.

Table 1: *Metrics generated for each punch thrown*

Metric	Description	Range
Power Commit (PC)	Determined in terms of type and effort committed to the punch. A value of 1 is assigned to punches thrown with so little intent that it could be debatable if they should be counted as a punch thrown. Max are highlight worthy shots with full rotation and body-weight committed to the punch.	$[1, 5]$
Landed Quality (LQ)	How cleanly a punch lands based on imperfections (glancing, blocked, parried, deflected, off-target, etc.). Scored 0 (miss) to 5 (clean contact with large reaction), with intermediate values based on number of imperfections (1=3+, 2=2, 3=1, 4=0 with reaction).	$[0, 5]$
Impact (I)	Combination of landed quality and power commit: $I = \min(\text{LQ}, \text{PC} + 1)$ for landed punches ($\text{LQ} > 0$), otherwise $I = 0$.	$[0, 5]$
Hand	Which hand the punch was thrown with, either left or right. We use stance data to infer whether this is likely to be the lead or rear hand.	{left, right}
Punch Type	The type of punch thrown.	{straight, hook, uppercut, overhand}
Target Area	Area of the opponent's body targeted.	{head, body}

Punch-level data is aggregated into some of our round-level performance measures. There are 47 round-level performance measures, spanning five categories:

Volume Statistics: Includes basic punch counts (thrown, landed, missed) and accuracy rates. Landed punches are enumerated for each impact level (min, low, mid, high, max) and target area (head, body). Miss counts are similarly categorized by the power commit of the attempted punch.

Combination Metrics: Quantify the percentage of punches thrown as singles, doubles, triples, or combinations of four or more punches.

Positional Metrics: Includes stance preferences (orthodox, southpaw, squared), distance management (outside, midrange, inside, clinch), and balance distribution (backfoot, neutral, frontfoot) measured during punch events.

Pressure Indicators: Three metrics which characterize a boxer's ring control: distance pressure (time spent in close proximity to opponent), movement pressure (time spent advancing/forcing opponent backward), and positional pressure (time spent with opponent on ropes/corners).

Aggression Indicators: Three metrics which characterize a boxer's offensive initiative: combination aggression (time spent throwing punches in combinations), exchange aggression (time spent initiating or ending exchanges), and power aggression (time spent throwing high-commitment punches).

Complete operational definitions for all 47 round-level metrics are available (see Appendix Table 7).

Professional judge scores serve as our dependent variable. Each round in our sample is scored by three licensed judges using the 10-point must system. After filtering uncommon scores, individual

decisions were treated as binary (win = 1, loss = 0). Averaging across judges yielded four outcomes: unanimous loss (0), split decision loss (1/3), split decision win (2/3), and unanimous win (1).

2.4 Scorecard Prediction

We implement two methods for predicting scorecards: a points-based model (PB) optimized using gradient descent and a multi-layer perceptron (MLP). To avoid overfitting we limited both models to 39 non-redundant performance metrics, removing features that are mathematical combinations of other included metrics.

We also experimented with constructing a point-based model with very few input parameters and, in doing so, explore the robustness of minimal, lightweight scoring models that are far easier to interpret than maximally detailed models.

2.4.1 Points-Based Modeling

We formalize an interpretable scoring system that assigns each boxer a score between 0 and 1 based on a ratio of points earned.

Each boxer accumulates points based on their round performance:

$$R_{\text{points}} = aR_1 + bR_2 + cR_3 + \dots \tag{1}$$

$$B_{\text{points}} = aB_1 + bB_2 + cB_3 + \dots \tag{2}$$

where $\{R_1, R_2, R_3, \dots\}$ and $\{B_1, B_2, B_3, \dots\}$ represent performance statistics for the red and blue corners. These stats could, for example, be how many punches they landed, the quality of the punches landed, the amount of time their opponent was on ropes, etc. $\{a, b, c, \dots\}$ are the weights determining how the system values each statistic. Critically, identical weights are applied to both boxers' statistics, ensuring the system cannot learn to favor one corner over the other. This is a serious concern, seeing as corner assignment is not random in pro boxing. The red corner is traditionally assigned to the fighter who is more popular, higher-ranked or more experienced.

We convert point totals to predictions as follows:

$$R_{\varphi} = \frac{R_{\text{points}} + D}{B_{\text{points}} + D} \tag{3}$$

$$R_{\Theta} = \frac{(R_{\varphi})^S}{(R_{\varphi})^S + 1} \tag{4}$$

Training Details: Data for individual rounds is randomly partitioned into training (80%) and testing (20%) sets. Weights $\{a, b, c, \dots\}$ and parameters {*D*, *S*} were optimized via gradient descent (learning rate: 0.00001, momentum: 0.9996, 20,000 iterations, gradient clipping: ±0.5) to minimize MSE between predictions and judge scores

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(R_{\Theta,i} - y_i)^2 \tag{5}$$

Where *N* is the number of rounds, $R_{\Theta,i}$ is the predicted score and $y_i \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ is the average score across the three judges.

2.4.2 Multi-Layer Perceptron Network

An MLP with two hidden layers (32 and 16 neurons, ReLU activation, L2 regularization: 0.001) and sigmoid output is implemented as an alternative modeling method. Data is partitioned into training (60%), validation (20%), and testing (20%) sets. The testing set is identical to the set used to evaluate PB. Features are standardized using StandardScaler. Training employs Adam optimization (learning rate: 0.001) with MSE loss.

2.4.3 Rankings to Contextualize Model Performance

We develop a benchmark where our models' accuracy could be directly compared to that of the most prolific judges in our dataset.

Pairwise Comparison Methodology: For each round, judges were compared against both co-judges (2 comparisons per judge per round), while models were compared against all three judges (3 comparisons per round).

$$\text{accuracy} = \frac{\text{total pairwise agreements}}{\text{total pairwise comparisons}} \tag{6}$$

To account for sample size bias, we apply Empirical Bayes shrinkage, which regularizes small-sample accuracy estimates toward the population mean (μ = 82.78%). The new accuracy estimate is formulated as follows:

$$\text{accuracy}_{\text{shrunk}} = \frac{n \cdot \text{accuracy}_{\text{observed}} + k \cdot \mu}{n + k} \tag{7}$$

where n is the number of rounds scored, k is the shrinkage parameter controlling the strength of regularization toward the mean. The optimal k value balances stabilizing noisy small-sample estimates while preserving variance attributable to genuine skill differences. We use Method of Moments estimation to derive an appropriate k value from the judging data itself.

Method of Moments decomposes observed variance into between-judge skill differences (τ^2) and within-judge measurement noise. For binomial data, the expected within-judge variance for a judge with n rounds is approximately $\frac{\mu(1-\mu)}{n}$ and the between-judge variance (τ^2) can then be estimated as:

$$\tau^2 = s^2 - \mathbb{E}[\text{within-judge variance}] \tag{8}$$

where s^2 is the sample variance of observed accuracies. The optimal shrinkage parameter is then:

$$k = \frac{\mu(1-\mu)}{\tau^2} \tag{9}$$

Shrinkage was applied only to judges (not models), as mean accuracy of human judges is not a meaningful prior for model performance. Moreover, models have sufficiently large sample sizes (4,350+ comparisons), rendering small-sample bias negligible.

2.5 Determining Feature Importance

Plus-minus performance metrics are formulated as differentials between opponents ($metric_{self}$ - $metric_{opponent}$), except for total volume and percentage-based ratio statistics, which are analyzed as raw values.

We employ three complementary approaches to assess feature importance:

2.5.1 Points-Based Approaches

One of the advantages of the PB model is that final weights grant insight into the relative importance of the corresponding feature. Large positive weights suggest judges favor that metric, large negative weight values suggest penalization, and near-zero weights suggest minimal influence. However, weights are context dependent. Their values change dramatically based on which other features are included in the model.

Furthermore, different combinations of features will result in different predictive accuracy. To determine the value of punch impact differentiation (Hypothesis 1), we compared two PB models: one using only the number of landed and missed punches, and another differentiating landed punches by five impact levels (min, low, mid, high, and max). By comparing the accuracy of these two models and analyzing their weights, we quantify the value placed on punches of different impact categories.

2.5.2 Correlation Analysis

Spearman rank correlations between plus-minus metrics and judge scores assessed relationship strength and direction, addressing Hypothesis 2 by indicating the relative importance of pressure and aggression indicators. Also, a small correlation matrix shows inter-metric relationships between pressure, aggression, and volumetric statistics. It serves to distinguish features measuring distinct aspects of performance from those that merely correlate with more important measures.

2.5.3 L1 Logistic Regression

L1 logistic regression was implemented to identify important features for Hypothesis 2. The L1 regularization automatically drives coefficients of unimportant variables to zero, revealing which factors contribute most to judging decisions. Multicollinearity among input features can lead to suboptimal feature selection. Thus, we preemptively reduce the feature space from 47 plus-minus metrics to a smaller set of 30, retaining conceptually distinct representatives from each performance category. This included compound statistics like "high impact" (high + max impact punches) not used in PB/MLP modeling. Regularization strength is optimized via 5-fold cross-validation across C values from 10^{-6} to 0.01.

3. Results

3.1 Accuracy of Predictive Models

The points-based model is more successful than MLP in scoring new rounds across all three measures of prediction accuracy. Furthermore, Tiny PB model, a points-based model with only five metrics (head, high impact, aggression power, inside and neutral) achieved accuracy very close to that of the larger 39 parameter models.

Table 2: *Comparing model performance on testing set (1450 rounds)*

Measure of Prediction Accuracy	PB Model	MLP Model	Tiny PB Model
Pairwise comparison accuracy	75.98%	75.52%	75.54%
Agreement with the majority of judges	77.59%	77.31%	77.24%
Mean square error	0.38327	0.39190	0.40323

Method of Moments estimation yielded k = 28.71, confirming genuine skill differences exist among judges (between-judge variance $\tau^2 = 0.005$) beyond random sampling variation. This k value provides moderate shrinkage: judges with fewer than 30 rounds are substantially regularized toward the population mean (82.78%), while well-sampled judges (100+ rounds) experience minimal adjustment.

When ranked against all judges with at least 20 rounds (n=227) without the use of shrinkage PB model ranks in the 22nd percentile while Tiny PB model and MLP model fall in the 20th percentile.

Table 3: *Professional judges and predictive models ranked by raw pairwise comparison accuracy. Only judges with 20+ rounds.*

Rank	Judge	Accuracy	Rounds
1	Judge A	98.33%	60
2	Judge B	97.83%	46
3	Judge C	96.51%	86
4	Judge D	95.45%	44
5	Judge E	94.44%	108
6	Judge F	93.75%	48
7	Judge G	93.75%	48
8	Judge H	93.55%	62
9	Judge I	93.55%	62
10	Judge J	93.48%	46
	...	...	...
173	Judge K	76.56%	64
174	Judge L	76.52%	164
-	*PB model (all rounds)*	*76.45%*	*7250*
-	*MLP model (all rounds)*	*76.33%*	*7250*
175	Judge M	76.19%	42
176	Judge N	76.09%	23
177	Judge O	76.09%	23
178	Judge P	76.04%	48
-	*PB model (test set only)*	*75.98%*	*1450*
179	Judge Q	75.86%	29
-	*Tiny PB model (all rounds)*	*75.78%*	*7250*
180	Judge R	75.77%	130
181	Judge S	75.61%	41
-	*Tiny PB model (test set only)*	*75.54%*	*1450*
-	*MLP model (test set only)*	*75.52%*	*1450*
182	Judge T	75.37%	67
183	Judge U	75.00%	42
	...	...	...
225	Judge V	60.87%	23
226	Judge W	60.71%	28
227	Judge X	54.55%	22
avg	**all judges**	**81.41%**	**7250**

Table 4: *Same methodology as Table 3 except with Empirical Bayes shrinkage (k=28.71) applied to judges only; models ranked by raw pairwise comparison accuracy.*

Rank	Judge	Accuracy	Rounds
1	Judge A	98.33%	60
2	Judge B	97.83%	46
3	Judge C	96.51%	86
4	Judge D	95.45%	44
5	Judge E	94.44%	108
6	Judge F	93.75%	48
7	Judge G	93.75%	48
8	Judge H	93.55%	62
9	Judge I	93.55%	62
10	Judge J	93.48%	46
	...	...	...
202	Judge K	76.56%	35
203	Judge L	76.52%	35
-	*PB model (all rounds)*	*76.45%*	*7250*
204	Judge M	76.40%	77
-	*MLP model (all rounds)*	*76.33%*	*7250*
-	*PB model (test set only)*	*75.98%*	*1450*
205	Judge N	76.56%	21
206	Judge O	75.93%	23
-	*Tiny PB model (all rounds)*	*75.78%*	*7250*
-	*Tiny PB model (test set only)*	*75.54%*	*1450*
-	*MLP model (test set only)*	*75.52%*	*1450*
207	Judge P	75.48%	20
208	Judge Q	75.48%	20
209	Judge R	75.46%	22
210	Judge R	74.98%	57
	...	...	...
225	Judge S	60.87%	28
226	Judge T	60.71%	56
227	Judge U	54.55%	44
avg	**all judges**	**81.41%**	**7250**

Once shrinkage is applied to the 227-judge sample, PB model is in the 10th percentile while Tiny PB model and MLP model fall in the 9th percentile.

When the models are ranked against all judges with 100+ rounds (n=45), all three models rank in the 4th percentile with shrinkage applied. Without shrinkage PB model ranks one place higher, 7th percentile (Appendix Table 8 and Table 9).

3.2 Value of Punch Impact Differentiation (Hypothesis 1)

We compare a minimal PB model using only the number of landed punches and the number of missed punches against a larger model that differentiates landed punches by impact level. The comparison demonstrates meaningful value in accounting for punch impact.

Table 5: *Comparison of the predictive accuracy of PB with and without punch impact differentiation, evaluated on testing set of 1450 rounds.*

Measure of Prediction Accuracy	No Impact Diff.	With Impact Diff.
Pairwise comparison accuracy	71.89%	73.15%
Agreement with the majority of judges	73.103%	74.62%
Mean square error	0.48156	0.44828

Table 6: *Weights assigned to parameters when training the points-based model without punch impact differentiation*

Parameter	Raw Weight
missed	0.193
landed	2.886
D (dampener)	127.040
S (sharpness)	8.776

Table 7: *Weights assigned to parameters when training the points-based model with punch impact differentiation.*

Parameter	Raw Weight	Normalized
missed	0.150	0.24
min	0.636	1.00
low	0.924	1.45
mid	1.612	2.54
high	2.796	4.40
max	6.676	10.50
D (dampener)	1.000	-
S (sharpness)	2.424	-

Normalized values use min impact as baseline (1.00). Maximum impact punches are valued over 10× higher than minimum impact punches

3.3 Importance of Pressure and Aggression Indicators (Hypothesis 2)

The mix of analytical approaches provides partial support for Hypothesis 2, revealing differences in importance among pressure and aggression indicators.

3.3.1 Feature Selection with Logistic Regression

L1 logistic regression identifies aggression power as the second most important metric (coefficient = 0.598), indicating judges highly value high-commitment punches regardless of landing success. Pressure position and pressure distance rank in the top 50% of features, suggesting moderate importance for controlling ring geography. However, pressure movement, aggression combinations, and aggression exchanges are not selected, indicating minimal importance.

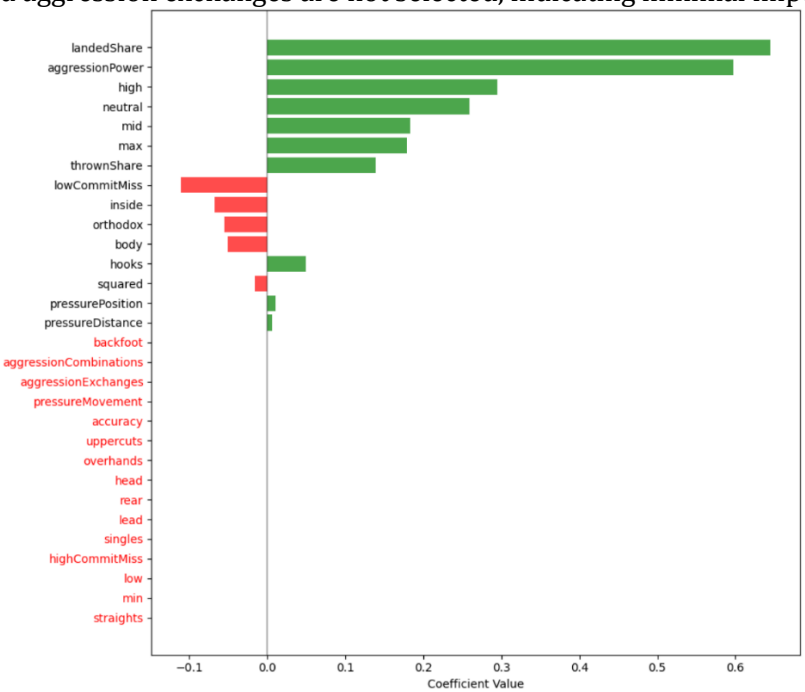

Figure 2: *L1 logistic regression. 15 of 30 features selected. Unselected features shown in red.*

3.3.2 Correlation to Round Score

Spearman correlations support the feature selection findings: aggression power shows the strongest correlation among pressure/aggression indicators ($r = 0.493$, $p < 0.001$), followed by aggression combinations ($r = 0.374$). Remaining indicators show weaker correlations ($r = 0.139$-0.168), all statistically significant but practically modest.

3.3.3 Multicollinearity Analysis

Aggression indicators and pressure distance show moderate correlations with offensive volume measures, while pressure movement (all |r| < 0.29) and pressure position (all |r| < 0.19) demonstrate weak inter-metric correlations, indicating distinctiveness. This explains pressure position's importance in feature selection despite weak correlation to judge scores.

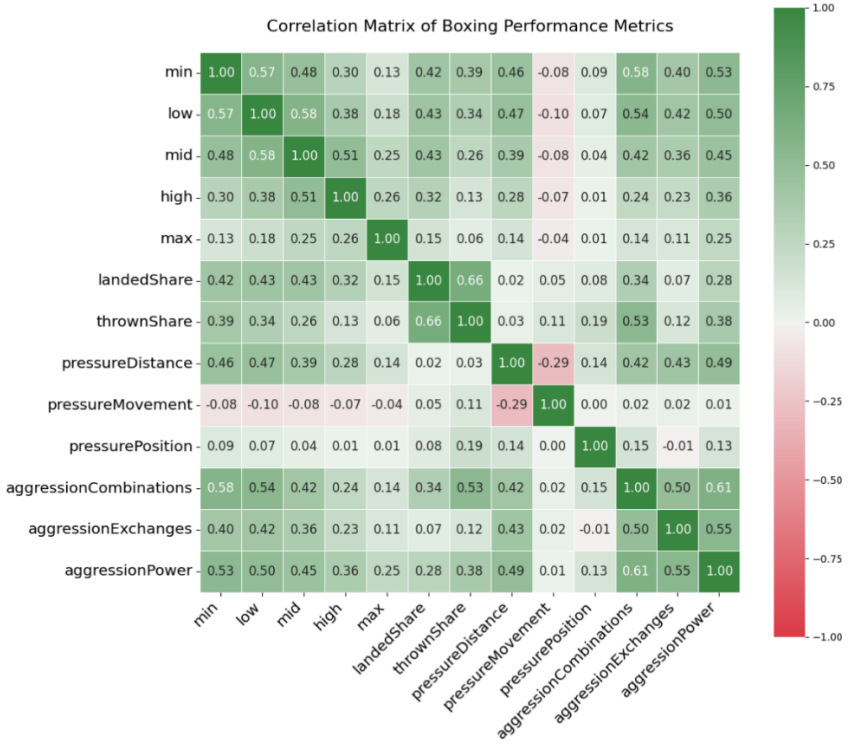

Figure 3: *Correlation matrix demonstrating relationships between aggression, pressure, and volumetric metrics.*

4. Discussion

This study represents the first large-scale automated analysis of professional boxing performance, leveraging AI-generated metrics from 1,003 bouts to understand what drives judging decisions. Our findings also demonstrate that fully automated systems can achieve accuracy comparable to a professional judge utilizing the 10-point must system, while maintaining transparency and interpretability.

4.1 Automated Judging Performance

Both models demonstrate professional-level judging capability. The point-based and MLP models respectively rank within the range of professional judges. The point-based model had 76% agreement, below the average pairwise agreement rate of 81% across all the professional judged rounds. This ranks in the 26th percentile of the 227 judges in the dataset with more than 20 rounds.

The points-based model produced slightly better predictions than the MLP model, a surprising result given that MLPs are typically regarded as more expressive. This result supports the use of interpretable points-based models for understanding judge preferences.

The Tiny PB model, using just 5 stats, produced slightly better predictions than the MLP model. This result supports the use of easily interpretable live-scoring points-based models for understanding judge preferences or even automated scoring, particularly considering the constraints and limitations outlined in section 4.5.

The competitive performance of point-based models is noteworthy given the inherent constraints on them that don't apply to professional judges.

1. The point-based model (and Tiny PB in particular) is a simple point-scale based on a few basic metrics being counted automatically by DeepStrike, thereby making its decisions transparent, reproducible, and interpretable

2. The PB models have no visibility into potentially biasing factors that may have influenced the human professional judges collectively, such as crowd noise, outcomes of previous rounds, fighter nationality, popularity, and reputation.

Judge bias is a known problem in boxing [30] and can even affect agreement-based rankings like those used in this paper. If biasing elements skew judges in the same direction, the judges' agreement rates can artificially inflate.

Including biasing factors in our models may increase prediction accuracy. However, their inclusion would fundamentally undermine the fairness of the judging system. The system's exclusive focus on the round-by-round performance metrics guarantees an objective and neutral basis for scoring.

4.2 Value of Punch Impact Differentiation

Punch impact differentiation represents a meaningful advancement over simple landed/missed punch counting. The exponential weighting pattern, with maximum impact punches valued over 10× more than minimum impact punches, quantifies what observers have long understood: not all landed punches are equal. This finding directly challenges binary classification systems and supports the use of sophisticated approaches to tracking that prioritize detailed punch classification.

4.3 Relative Importance of Pressure and Aggression Indicators

Our analysis reveals distinct patterns in how judges value pressure and aggression indicators, providing partial support for Hypothesis 2. While we expected all pressure and aggression indicators to demonstrate at least moderate importance, several failed to meaningfully contribute to predictive accuracy. Surprisingly, metrics we anticipated would carry limited value, such as balance and stance indicators, significantly outperformed most pressure and aggression measures.

Aggression power emerged as one of the most important performance metrics across all analytical approaches, demonstrating judges' appreciation for high power-commit punches independent of success landing. This likely reflects judges' difficulty distinguishing between fully blocked shots and

those that partially penetrate defenses, their appreciation for offensive commitment, and recognition that powerful punches can cause damage even when blocked. This finding aligns with established boxing wisdom that "occupying the guard" demonstrates dominance and control irrespective of clean punching.

Two tactical indicators showed moderate importance: pressure position (controlling ring geography by placing opponents on ropes/corners) and pressure distance (maintaining close distance to the opponent). Pressure position demonstrated predictive value despite weak individual correlation with judge scores, likely due to its distinctiveness as the only measure relative to ring geography.

On the other hand, aggression combinations (throwing punches in combinations) showed moderate correlation ($r = 0.374$) but provided little value in multivariate models, likely due to its overlap with impact metrics and aggression power.

Two commonly emphasized tactical elements showed surprisingly limited predictive value. Pressure movement had essentially no meaningful relationship to judge scores, challenging traditional emphasis on "walking down" opponents. Aggression exchanges also contributed negligibly to predictive accuracy, indicating that initiating or ending exchanges matters far less than the quality and power of punches within those exchanges.

4.4 Implications and Applications

In mapping DeepStrike metrics to 17,575 individual judge scores, we created a tool that brings transparency and objectivity to professional boxing while reflecting current judging practices. This research also quantifies which performance factors judges prioritize when interpreting subjective scoring criteria.

Three primary applications emerge:

Supplementing Human Judges: The system provides consistent, bias-free scoring at events lacking experienced judges, particularly in smaller venues or developing boxing markets where qualified judges are scarce. Unlike human judges, the model is effectively immune to bias based on fighter nationality, crowd reaction, or promotional pressure.

Judge Performance Evaluation: Comparing human judges against our automated scoring system is a valuable alternative to "deviation from the majority", a common agreement-based evaluation technique which not only ties a judge's accuracy estimate to the competency of their co-judges but also incentivizes collusion between judges [30]. The ability to score bouts based on DeepStrike stats enables objective evaluation against a consistent baseline that can be deployed at an unprecedented scale.

Training Strategy: Quantifying which metrics judges value allows fighters to focus their training on what judges in practice tend to favor. Our methodology and sample also make it possible to retrain the PB model on data from certain individual judges with hundreds of rounds scored and uncover stylistic differences between judges. Information like this could be immensely beneficial to fighters seeking to succeed on the judges' scorecards.

4.5 Limitations and Potential Improvements

Several limitations warrant consideration when interpreting these results.

The version of DeepStrike used for this study is a beta-model from 2024 provided by Jabbr. The statistics in this study were generated autonomously without any human validation. DeepStrike does occasionally make visible mistakes, with an estimated per-punch accuracy of around 95%. Furthermore, all stats were gathered on single-angle dirty-feed production footage, running at either 25 or 30fps. Single-camera footage like this suffers from occlusion issues where punches and their outcome are frequently blocked from view by referees, graphics, or fighters' bodies. DeepStrike supports multi-cam input, with confidence-weighted selection that significantly improves the accuracy of the automated statistics on which the scoring is mapped. We expect that repeating this same study with multi-cam data or a newer updated version of DeepStrike would provide more accurate statistics and lead to better predictive performance.

Furthermore, the dataset used in this study may contain geographical bias due to underrepresentation of UK bouts caused by scorecard unavailability. Furthermore, we expect there to be systematic underrepresentation of knockout artists who generate fewer usable rounds. For example, defensive specialist Floyd Mayweather Jr. averages 9.8 usable rounds per bout in our dataset while power-puncher Mike Tyson averages only 3.9 usable rounds, 60% fewer. This effect biases data toward defensive, decision-oriented fighting styles.

Several factors that likely influence judging remain untracked: visible signs of damage (cuts, swelling, unsteady movement), temporal context of defensive behavior following impact (e.g., clinching after absorbing a hard shot versus strategic clinching), and potential recency bias favoring late-round action. Finally, aggression and pressure indicators use arbitrarily set timing windows within DeepStrike (e.g. 3 seconds of aggression power for throwing with power commit level 4). The length of these timing windows could themselves be optimized to maximize predictive value.

5. Conclusion

This study applies machine learning-based analysis to statistics generated by DeepStrike, a cutting-edge computer vision system for combat sports. In doing so, we analyze judging patterns at unprecedented scale. Where previous research has relied on manual annotation of fewer than 50 bouts, our approach analyzed 1,003 professional bouts. The scale of data not only allowed us to investigate what factors drive judges' decisions but model the complex, opaque decision-making of professional judges. The results demonstrate that DeepStrike statistics can serve as the basis for objective, interpretable, and accurate scoring that accurately reflects the tendencies of professional judges.

The methodology offers practical solutions to problems that have plagued combat sports for decades. By creating transparent, bias-free scoring that reflects how rounds are actually judged at the professional level, this work provides a foundation for improving judge evaluation, informing training strategies, and even supplementing human judging. The framework established here demonstrates how automated performance analysis can bring greater objectivity and understanding to areas of sport limited by their subjectivity.

References

[1] Velasco, G., Neidecker, J., Muzzi, D., & Sethi, N. (2019). Retrospective analysis of professional boxing fight outcomes in the United States during a 6 month study period in 2017. *Neurology*, *93*(14_Supplement_1), S11–S12.

[2] Association of Boxing Commissions and Combative Sports. (2024). *Boxing judge manual*.

[3] Fédération Internationale de Gymnastique. (2024). *Men's artistic gymnastics code of points 2025-2028*.

[4] U.S. Figure Skating. (n.d.). *Understanding the international judging system: Scoring cheat sheet*.

[5] International Ski Federation. (2019). *FIS freestyle skiing judging handbook*.

[6] U.S. Senate. (2002). *A review of the professional boxing industry - Is further reform needed?*

[7] Thomson, E., & Lamb, K. (2016). The technical demands of amateur boxing: Effect of contest outcome, weight and ability. *International Journal of Performance Analysis in Sport*, *16*(1), 203–215.

[8] Kapo, S., El-Ashker, S., Kapo, A., Colakhodzic, E., & Kajmovic, H. (2021). Winning and losing performance in boxing competition: A comparative study. *Journal of Physical Education and Sport*, *21*(4), 2124–2130.

[9] Wylie, L. (2015). The inherent problems with CompuBox statistics. *The Sweet Science*.

[10] Ruebusch, C. (2013, May 29). The CompuBox lie: Why punch stats don't tell the whole story. *Bloody Elbow*.

[11] Starks, T. (2018). The problem with CompuBox. *The Ring Magazine*.

[12] McCarson, K. (2023). Why the UFC's fight stats need an overhaul. *The Rant 365*.

[13] A4 Fitness. (2023). *What is a significant strike in MMA? Understand the impact and rules*.

[14] James, N., Hughes, M., & James, N. (2016). Performance analysis in mixed martial arts: A systematic review. *International Journal of Performance Analysis in Sport*, *16*(2), 492–507.

[15] Fujitsu. (2020). *Judging support system for gymnastics*.

[16] Allen, E., Fenton, A., & Parry, K. (2021). Computerised gymnastics judging scoring system implementation – An exploration of stakeholders' perceptions. *Science of Gymnastics Journal*, *13*(3), 357–370.

[17] Spitz, J., Wagemans, J., Memmert, D., Williams, A., & Helsen, W. (2021). Video assistant referees (VAR): The impact of technology on decision making in association football referees. *Journal of Sports Sciences*, *39*(1), 1–7.

[18] Badiola-Bengoa, A., & Mendez-Zorrilla, A. (2021). A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. *Sensors*, *21*(18), Article 5996.

[19] Stefański, P., Kozak, J., & Jach, T. (2024). Boxing punch detection with single static camera. *Entropy*, *26*(8), Article 617.

[20] Feiz, H., Labbé, D., Romeas, T., Faubert, J., & Andrews, S. (2025). *Multi-person physics-based pose estimation for combat sports* (arXiv:2504.08175). arXiv.

[21] Fleisig, G. S., Slowik, J. S., Wassom, D., Yanagita, Y., Bishop, J., & Diffendaffer, A. (2024). Comparison of marker-less and marker-based motion capture for baseball pitching kinematics. *Sports Biomechanics*, *23*(12), 2950–2959.

[22] Lai, C., Mo, J., Xia, H., & Wang, Y. (2024). *FACTS: Fine-grained action classification for tactical sports* (arXiv:2412.16454). arXiv.

[23] Jabbr.ai, Nielsen, A. S., & Obeid, E. (2024). *Deepstrike: The world's first & best computer vision AI for combat sports* [Computer software].

[24] Chowdhury, S. R. (2024). First in history, Tyson Fury vs. Oleksandr Usyk will use AI to generate boxing's most debated stats. *Essentially Sports*.

[25] Jay, P. (2024). Jabbr AI stats launch shows Usyk clearly beat Fury. *World Boxing News*.

[26] McGoldrick, S. (2022). New AI technology being developed that could revolutionise boxing and help eliminate judging issues. *Irish Independent*.
[27] Professional boxing video archive. (2025). [Video sources included: Top Rank, PBC, DAZN, Golden Boy, Queensberry, Matchroom, Salita Promotions, ProBox, Sky Sports, TNT Sports Boxing, HBO, Showtime (YouTube); Boxing Fights Videos, Boxing Laboratory, Boxing Online (Dailymotion); and other boxing archive channels. Complete list available upon request].
[28] BoxRec. (2024). *BoxRec: Boxing's official record keeper*.
[29] Association of Professional Boxing Commissions. (2024). *Pocket guide - Judges*.
[30] McLaren, R. H. (2021). *McLaren independent AIBA investigation report* [Investigation report]. McLaren Global Sport Solutions Inc.
[31] github.com/mduboef/Jabbr-AI-Boxing-Judging

Appendix

Table 7: *Details on the 47 end-of-round statistics generated by DeepStrike.*

Metric	Description	Range
Thrown	The number of punches thrown in a given round.	≥ 0
Landed	The number of punches landed in a given round.	≥ 0
Missed	The number of missed punches thrown in a given round.	≥ 0
Accuracy	The rate at which a fighter is landing their thrown punches. Derived by $\frac{\text{landed}}{\text{thrown}}$.	[0, 100]
Min	The number of minimum impact punches landed in a given round.	≥ 0
Low	The number of low impact punches landed in a given round.	≥ 0
Mid	The number of medium impact punches landed in a given round.	≥ 0
High	The number of high impact punches landed in a given round.	≥ 0
Max	The number of maximum impact punches landed in a given round.	≥ 0
High Impact	The number of high and max impact punches landed in a given round.	≥ 0
Min Miss	The number of misses thrown with minimum power commit.	≥ 0
Low Miss	The number of misses thrown with low power commit.	≥ 0
Mid Miss	The number of misses thrown with medium power commit.	≥ 0
High Miss	The number of misses thrown with high power commit.	≥ 0
Max Miss	The number of misses thrown with maximum power commit.	≥ 0
Low Commit Miss	The number of min misses and low misses.	≥ 0
High Commit Miss	The number of mid, high and max misses.	≥ 0
Singles	What percentage of punches are thrown as a single.	[0, 100]
Doubles	What percentage of punches are thrown as part of a two-punch combination.	[0, 100]
Triples	What percentage of punches are thrown as part of a three-punch combination.	[0, 100]
Quads+	What percentage of punches are thrown as part of a combination of four or more punches.	[0, 100]
Lead	What percentage of punches are thrown with the lead hand.	[0, 100]
Rear	What percentage of punches are thrown with the rear hand.	[0, 100]
Straights	What percentage of punches are a straight or jab.	[0, 100]
Hooks	What percentage of punches are a hook.	[0, 100]
Overhands	What percentage of punches are an overhand.	[0, 100]
Uppercuts	What percentage of punches are an uppercut.	[0, 100]
Head	What percentage of punches target the opponent's head.	[0, 100]
Body	What percentage of punches target the opponent's body.	[0, 100]
Aggression Combinations	What percentage of the round is a fighter throwing punches in combinations.	[0, 100]
Aggression Exchanges	A measure of how often a fighter throws the first and last punch of an exchange with the opponent.	[0, 100]
Aggression Power	How often a fighter throws punches with high power commit.	[0, 100]
Aggression	How much of the round a fighter spends with at least one aggression metrics triggered.	[0, 100]
Pressure Distance	How much of the round a fighter spends in a close or mid-range distance to their opponent. If a boxer is on/near the ropes or the corners of the ring they will never trigger pressure distance.	[0, 100]
Pressure Movement	What percentage of the round a fighter is moving forward, towards their opponent, and making their opponent move backwards.	[0, 100]
Pressure Position	What percentage of the round a fighter has their opponent on/near the ropes or the corners of the ring.	[0, 100]
Pressure	How much of the round a fighter spends with at least one pressure metrics triggered.	[0, 100]
Orthodox	The percentage of the round a fighter spends in an orthodox stance, measured in 1 second intervals around punch events.	[0, 100]
Southpaw	The percentage of the round a fighter spends in a southpaw stance, measured in 1 second intervals around punch events.	[0, 100]
Squared	The percentage of the round a fighter spends in a squared stance, measured in 1 second intervals around punch events.	[0, 100]
Outside	The percentage of the round a fighter spends at an outside distance, measured in 1 second intervals around punch events.	[0, 100]
Mid-Range	The percentage of the round a fighter spends in a mid-range distance, measured in 1 second intervals around punch events.	[0, 100]
Inside	The percentage of the round a fighter spends in an inside distance, measured in 1 second intervals around punch events.	[0, 100]
Clinch	The percentage of the round a fighter spends in a clinch, measured in 1 second intervals around punch events.	[0, 100]
Back-Foot	The percentage of the round a fighter spends with the balance on their back foot, measured in 1 second intervals around punch events.	[0, 100]
Front-Foot	The percentage of the round a fighter spends with the balance on their back foot, measured in 1 second intervals around punch events.	[0, 100]
Neutral	The percentage of the round a fighter spends with the even balance, measured in 1 second intervals around punch events.	[0, 100]

Table 8: *Professional judges and predictive models ranked by raw pairwise comparison accuracy. Only judges with 100+ rounds shown.*

Rank	Judge	Accuracy	Rounds
1	Judge A	89.50%%	100
2	Judge B	88.36%	159
3	Judge C	86.11%	108
4	Judge D	86.02%	161
5	Judge E	85.82%	141
6	Judge F	84.96%	123
7	Judge G	84.43%	427
8	Judge H	84.03%	144
9	Judge I	83.93%	112
10	Judge J	83.66%	465
	. . .	. . .	. . .
39	Judge K	78.23%	272
40	Judge L	77.72%	184
41	Judge M	77.19%	171
42	Judge N	76.52%	164
-	*PB model (all rounds)*	*76.45%*	*7250*
-	*MLP model (all rounds)*	*76.33%*	*7250*
-	*PB model (test set only)*	*75.98%*	*1450*
43	Judge O	75.77%	130
-	*Tiny PB model (all rounds)*	*75.78%*	*7250*
-	*Tiny PB model (test set only)*	*75.54%*	*1450*
-	*MLP model (test set only)*	*75.52%*	*1450*
44	Judge P	73.03%	165
45	Judge Q	69.38%	129
avg	**all judges**	**81.41%**	**7250**

Table 9: *Same methodology as Table 8 except with Empirical Bayes shrinkage (k=28.71) applied to judges only; models ranked by raw pairwise comparison accuracy.*

Rank	Judge	Accuracy	Rounds
1	Judge A	88.00%	100
2	Judge B	87.51%	159
3	Judge C	85.53%	161
4	Judge D	85.41%	108
5	Judge E	85.30%	141
6	Judge F	84.55%	123
7	Judge G	84.32%	427
8	Judge H	83.82%	144
9	Judge I	83.69%	112
10	Judge J	83.60%	465
	. . .	. . .	. . .
39	Judge K	78.64%	294
40	Judge L	78.40%	184
41	Judge M	78.00%	171
42	Judge N	77.46%	164
43	Judge O	77.04%	130
-	*PB model (all rounds)*	*76.45%*	*7250*
-	*MLP model (all rounds)*	*76.33%*	*7250*
-	*PB model (test set only)*	*75.98%*	*1450*
-	*Tiny PB model (all rounds)*	*75.78%*	*7250*
-	*Tiny PB model (test set only)*	*75.54%*	*1450*
-	*MLP model (test set only)*	*75.52%*	*1450*
44	Judge P	74.48%	165
45	Judge Q	71.82%	129
avg	**all judges**	**81.41%**	**7250**