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

duBoef M [1,2], Romeas T [3,4,5], Charbonneau M [3], Svejstrup A [1]

[1] Jabbr, Denmark
[2] University of Massachusetts Amherst, United States
[3] Institut national du sport du Québec, Canada
[4] École de kinésiologie et des sciences de l'activité physique, Université de Montréal, Canada
[5] Department of Psychology, York University, Canada

BACKGROUND

In professional boxing 46.5% of matches are decided by judges' scorecards rather than knockout or stoppage [1]. Three judges score each round based on highly-subjective scoring criteria [2]. The sport is rife with scoring controversy [3] and intransparency as fans, athletes and analysts struggle to understand what it means to win a round of boxing [4].

Boxing lacks a large repository of quality statistics, mainly employing live clicker-based systems that are considered inaccurate and lacking in detail [5]. Existing research has instead relied on manual annotation, severely limiting sample size. Studies typically analyze fewer than 50 rounds, restricting researchers' ability to identify subtle judging patterns or develop robust predictive models [4, 6].

Using computer vision-generated statistics, we aim to (1) construct automated scoring models achieving professional-level accuracy and (2) identify which performance factors most strongly influence judges' decisions.

METHODS

We leverage DeepStrike, a new computer vision system never before used in academic work, to generate detailed round-by-round statistics on 7,323 rounds of professional boxing broadcast video. We map those statistics to judges' scores, constructing predictive models to automatically judge rounds (Figure 1). Since decisions are made using end-of-round statistics alone, this system is consistent and free from biasing factors like crowd noise, fighter nationality, and reputation. We train large 39-metric models for maximum accuracy. To gauge the accuracy of simpler scoring systems and reveal how features are valued relative to one another, we train minimal models based on just 5-6 metrics.

(1) We use two different methods to predict average score based on end-of-round statistics:

1. Points-based scoring (PB) optimized with gradient descent

$$R_{\text{points}} = aR_1 + bR_2 + cR_3 + \ldots$$

$$B_{\text{points}} = aB_1 + bB_2 + cB_3 + \ldots$$

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

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

$\{R_1, R_2, R_3, \ldots\}$ and $\{B_1, B_2, B_3, \ldots\}$ are sets of performance metrics for the red and blue corners.

$\{a, b, c, \ldots\}$ are the weights determining how the system values each metric.

Score (R_Θ) is calculated as a ratio of points using sharpness (S) and dampening (D) parameters.

2. Multi-layer perceptron (MLP)

This is a deep-learning approach. Neural networks are effective function approximators but remain opaque. Thus, predictions may be accurate, but the reasoning behind them is not interpretable.

Our MLP consists of an input layer, two hidden layers (32 and 16 neurons with ReLU activation), and a single output neuron with sigmoid activation. We train using the Adam optimizer.

(2) We use L1 logistic regression to identify important performance elements (*Figure 2*). To isolate the effect of punch impact we train a 6-parameter PB model, featuring only punch stats. By analyzing its optimized weights, we uncover how judges value punches based on their impact (*Table 2*).

RESULTS

(1) All models all ranked within accuracy range of a professional judge (*Table 1*) with PB classification (75.98%) achieving higher accuracy than MLP classification (75.52%). Our most detailed PB model ranks in the 22nd percentile among all judge with 20+ rounds scored (10th percentile with shrinkage). Among judges with 100+ rounds scored, it fell in the 7th percentile (all models ranked in the 4th percentile with shrinkage).

"Tiny PB", a points-based model based on only 5 performance metrics, achieved similar accuracy (75.54%). It ranked slightly higher than the 39-metric MLP model, demonstrating that very few metrics are needed to achieve professional-level accuracy

Table 1: *Predictive models and profession judges with 20+ rounds ranked by pairwise comparison accuracy. Each judge's accuracy reflects agreement rate with their two co-judges, Model's accuracy is based on agreement with all three judges in each round.*

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
…		…	…
177	Judge K	76.09%	23
178	Judge L	76.04%	48
—	*PB Model (Test Set)*	*75.98%*	*1450*
179	Judge M	75.86%	29
180	Judge N	75.77%	130
181	Judge O	75.61%	41
—	*Tiny PB Model (Test Set)*	*75.54%*	*1450*
—	*MLP Model (Test Set)*	*75.52%*	*1450*
182	Judge P	75.37%	67
183	Judge Q	75.00%	42
…		…	…
225	Judge R	60.87%	23
226	Judge S	60.71%	28
227	Judge T	54.55%	22
Avg	**All Judges**	**81.41%**	**7323**

Table 2: *Weights for a 6-metric PB model based on missed punches and punches landed by impact category (minimum to maximum). Values normalized around the weight of a min impact landed punch.*

Metric	Normalized Weight
missed	0.24
min impact	1.00
low impact	1.45
mid impact	2.54
high impact	4.40
max impact	10.50

(2) Higher impact punches are valued exponentially higher with a maximum impact punch assigned 10.5× the weight of a minimum impact punch

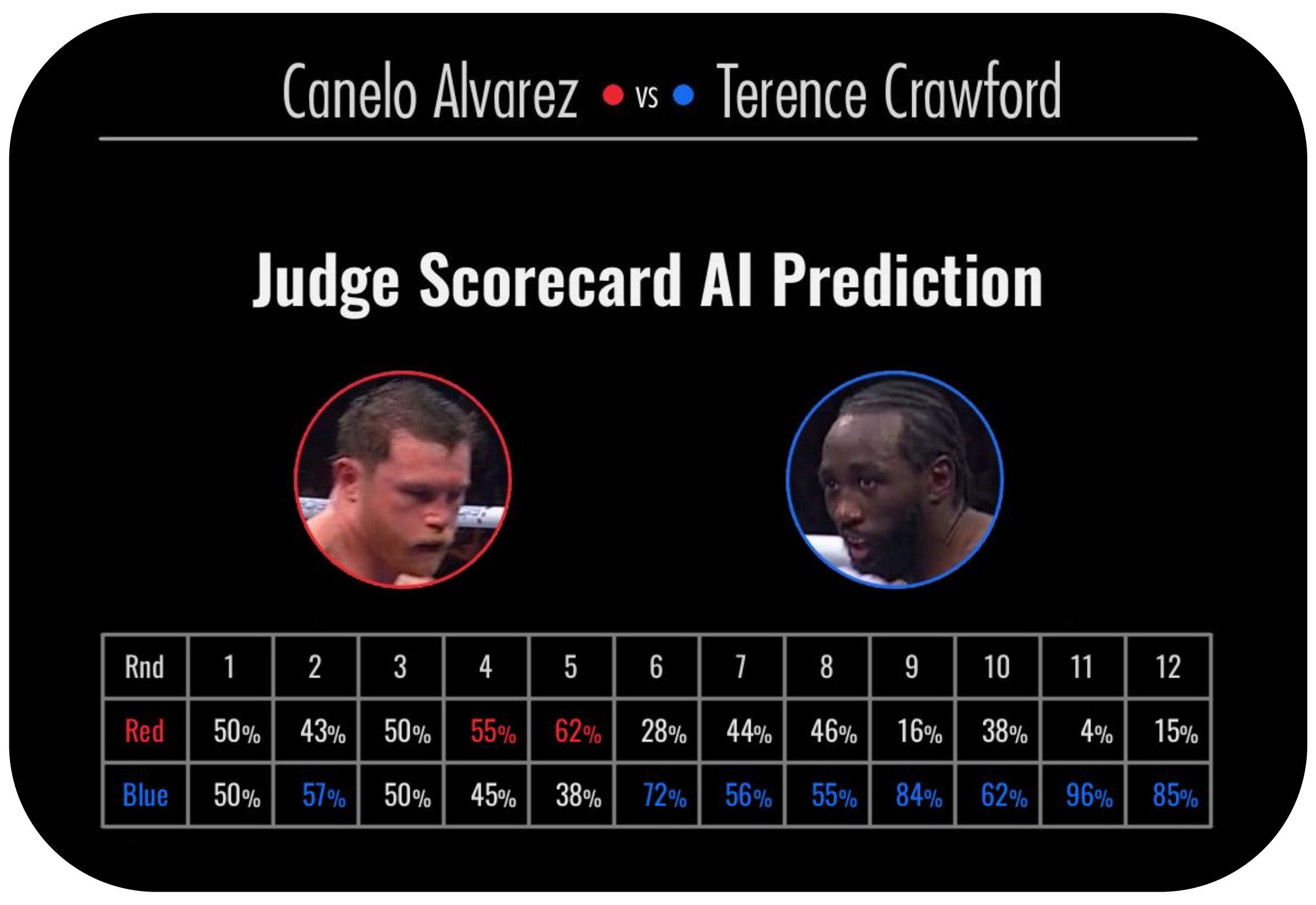
TAKEAWAYS

A simple points-based scoring system, using automated statistics achieves accuracy within the range of top-level professional judges, offering a scalable, consistent, transparent, and bias-free scoring standard.

Landing with impact and throwing with high power-commit are key factors driving judges' decisions, while commonly emphasized tactics like "walking down your opponent" show minimal predictive value.

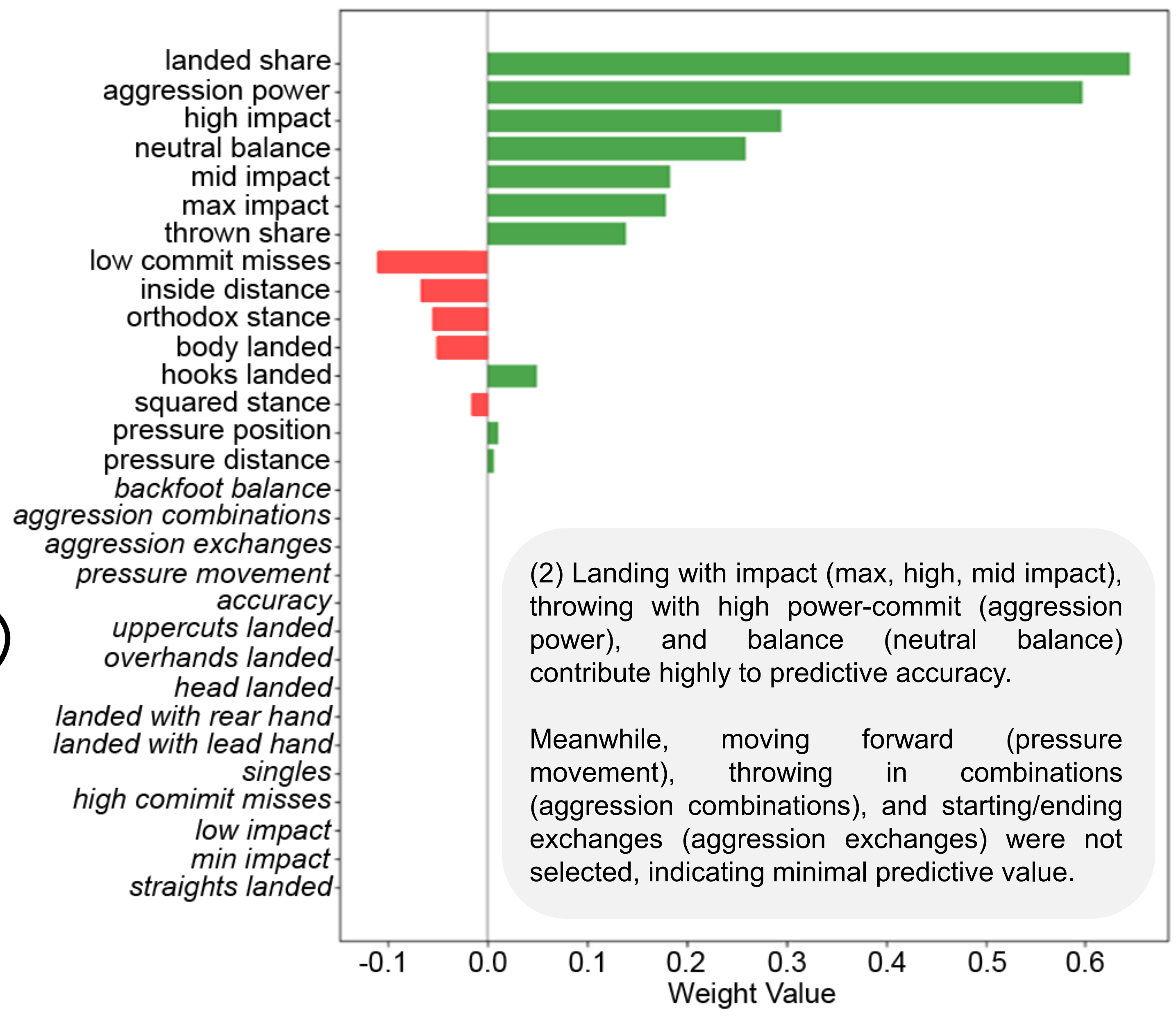
Punches are valued exponentially by impact. Max impact punches are worth over 10× more than minimum impact, highlighting the importance of differentiating punches by impact when tracking.

Figure 1: *Scorecard based on points-based automated scoring.*


Canelo Alvarez • vs • Terence Crawford
Judge Scorecard AI Prediction

Rnd	1	2	3	4	5	6	7	8	9	10	11	12
Red	50%	43%	50%	55%	62%	28%	44%	46%	16%	38%	4%	15%
Blue	50%	57%	50%	45%	38%	72%	56%	55%	84%	62%	96%	85%

Figure 2: *L1 logistic regression. 15 of 30 features selected. Unselected features italicized.*



(2) Landing with impact (max, high, mid impact), throwing with high power-commit (aggression power), and balance (neutral balance) contribute highly to predictive accuracy.

Meanwhile, moving forward (pressure movement), throwing in combinations (aggression combinations), and starting/ending exchanges (aggression exchanges) were not selected, indicating minimal predictive value.

REFERENCES

[1] Velasco et al. (2019). *Neurology*, 93:S11-S12.

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

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

[4] Kapo et al. (2021). *J Phys Educ Sport*, 21:2124-2130.

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

[6] Thomson et al. (2016). *Int J Perf Analysis Sport*, 16:203-215.

www.jabbr.ai ✉ mduboef@gmail.com in Mason duBoef www.insquebec.org/en