



Automated fault attribution for Formula 1 collisions: evidence for the limits of telemetry-based classification

Panchal D

Submitted: March 29, 2026, Revised: version 1, June 8, 2026, version 2, July 1, 2026

Accepted: July 1, 2026

Abstract

Fault attribution in Formula 1 collisions remains a manual process in which FIA stewards integrate video, telemetry, team radio, and regulatory interpretation. We present the first machine-learning study of whether telemetry alone can predict steward fault decisions. Using 164 steward-adjudicated collisions from the 2019–2025 seasons, we first extracted 225 statistical telemetry features. These failed to generalize across seasons, performing below a majority-class baseline under leave-one-season-out (LOSO) evaluation. We then developed 26 FIA-aligned interaction features representing spatial overlap, braking dynamics, temporal proximity, avoidance behavior, and car control. These reduced the within-distribution to cross-season generalization gap from 30.0 to 22.9 percentage points and achieved 72.9% balanced accuracy under stratified 10-fold cross-validation but only 52.3% under LOSO (AUROC = 0.577), only marginally above chance. Permutation importance and SHAP analyses showed that the limited predictive signal was concentrated in a small subset of braking-dynamics and spatial-overlap features, while additional telemetry features did not improve cross-season performance. These findings indicate that telemetry alone is insufficient for reliable fault attribution because critical evidence—including video, driver intent, race context, and evolving steward interpretation—is unavailable from telemetry. Rather than presenting a deployable classifier, this study establishes the first quantitative baseline for automated Formula 1 fault attribution and motivates future multimodal decision-support systems.

Keywords

Formula 1 telemetry, Fault attribution, Machine Learning, Sports analytics, Explainable artificial intelligence, Time series classification, Distribution shift, Human-AI decision support, Feature engineering, Automated officiating

Dhruv Panchal, Princeton Day School, 650 Great Road, Princeton, NJ 08540, USA. odiouschipmunk@gmail.com

1. Introduction

Formula 1 is governed by the Fédération Internationale de l'Automobile (FIA), which appoints a panel of stewards at each Grand Prix to adjudicate on-track incidents. When two cars make contact during a race, these stewards must decide whether the contact constitutes a racing incident, requiring no penalty, or whether one driver was predominantly at fault and should be penalized. To reach this determination, stewards review multiple information streams: onboard camera footage from both cars, live telemetry overlays, team radio communications, GPS position data, and their interpretation of the FIA Sporting Regulations (1). The outcome carries direct sporting consequences. Time penalties alter race results, grid penalties affect qualifying positions at future events, and accumulating penalty points across a season can result in race bans.

Despite the high stakes of these decisions, the steward process has attracted persistent criticism for perceived inconsistency. Commentators and teams have noted that the same category of incident may be penalized in one race and ruled a racing incident in another, even when the steward panel includes overlapping members (2). This perception of inconsistency is fueled in part by the subjective nature of the rules themselves, which require stewards to assess whether a driver was "predominantly at fault" and whether a collision was "avoidable," terms that leave substantial room for interpretation (1).

The FIA has invested in technology to support its officiating. The RaceWatch system uses computer vision to track all cars on circuit and

automatically flag potential track-limit violations and unsafe driving behavior (3). RaceWatch provides stewards with objective tracking data for certain rule infringements, such as exceeding track limits at specific corners. Yet no existing tool assists with the consequential question of fault attribution in car-to-car contact, where judgment depends on each driver's speed, trajectory, braking point, and positioning relative to the apex.

This paper asks whether machine learning classifiers, trained on telemetry from past incidents with known steward outcomes, can learn to predict those outcomes. We state upfront that under evaluation conditions approximating deployment, the answer with the original feature set is largely negative, yet we also identify why and introduce a more principled feature engineering framework that addresses the core limitation. The value of this work lies not in a deployable system but in three contributions: (a) the first formalization of the fault-attribution task as a supervised classification problem; (b) a reproducible collision-centered feature engineering pipeline for F1 telemetry, built on the open-source FastF1 API (4); and (c) the first FIA-aligned interaction feature set for this task, grounded in the specific right-of-way, overlap, and avoidability rules that stewards actually apply.

We address the high feature-to-sample ratio (225 features for 164 samples) through PCA (fitted exclusively within training folds), mutual-information feature selection, and regularized linear models. Modern time-series classification methods ROCKET and MiniROCKET serve as temporal baselines

alongside the traditional classifiers. Distribution shift is examined through LOSO, venue-grouped, and event-grouped cross-validation, with prior-shift correction methods (BBSE, threshold recalibration) evaluated and found ineffective. Calibration is assessed via ECE, MCE, Brier score, and reliability bins.

2. Related work

Automated officiating in sport has seen growing interest, particularly through computer-vision methods. Rabee et al. (5) recently demonstrated an automated foul detection system for football using machine learning on video data, achieving promising accuracy in identifying fouls from match footage. Similar approaches have been explored for basketball and cricket. In motorsport, published machine learning research has concentrated on adjacent problems: pit-stop strategy optimization and tire-selection modeling (6), race-outcome prediction from qualifying and practice data, and driver-performance analysis from telemetry and lap-time patterns (7). We are not aware of any prior study that applies machine learning to the fault-attribution question in racing collisions, making this work the first formalization of the problem.

Time-series classification (TSC) has advanced rapidly in recent years. ROCKET (8) applies a large bank of random convolutional kernels to extract features from time series, followed by a linear classifier. MiniROCKET (9) achieves comparable accuracy with a deterministic kernel set, reducing computational cost. Both methods have shown strong performance on the UCR/UEA archive benchmarks (10). They are applicable in our setting because the resampling step produces fixed-length multivariate

sequences. We include them as temporal baselines alongside our statistical feature approach, which trades temporal resolution for interpretability.

Feature engineering in sports analytics has increasingly moved toward physics-grounded and rule-aligned representations. In soccer, advanced metrics such as expected possession value and pressure ratings encode game concepts rather than raw positional summaries (11). Similarly, our FIA-aligned features translate the specific overlap and right-of-way rules in the FIA Sporting Regulations into computable telemetry quantities. This approach reflects the broader principle that domain-specific feature engineering outperforms generic statistical summaries when the decision process follows structured rules (12).

The non-stationarity challenge in our data connects to the broader literature on learning under distribution shift. Moreno-Torres et al. (13) provide a unifying taxonomy of dataset shift, distinguishing among covariate shift, prior probability shift, and concept shift. Label shift, where the marginal distribution $P(Y)$ changes between training and deployment while class-conditional distributions $P(X|Y)$ remain fixed, is one candidate explanation in our setting, given that per-season fault rates vary from 29.6% to 87.5%. BBSE (14) and earlier expectation-maximization methods (15) address label shift under this fixed-conditional assumption. Our experiments test the assumption and find it does not hold: correcting for label shift does not close the generalization gap, suggesting that steward interpretation criteria and driving patterns themselves change across seasons, consistent

with a mixture of prior probability shift and concept shift.

3. Data collection and preprocessing

3.1 Dataset overview

We collected telemetry for 164 collision incidents across the 2019-2025 F1 seasons using the FastF1 Python API (4). Each incident involved two cars and is labeled according to the FIA steward decision: at-fault if a penalty was

issued ($n = 104$, 63.4%), no-fault if the stewards ruled it a racing incident or took no further action ($n = 60$, 36.6%). Labels were extracted by parsing FIA steward decision documents, which are publicly available on the FIA website. Any penalty (time penalty, grid penalty, or penalty points) maps to the at-fault label; all other outcomes map to no-fault. The per-season distribution of incidents and fault rates is shown in Figure 1.

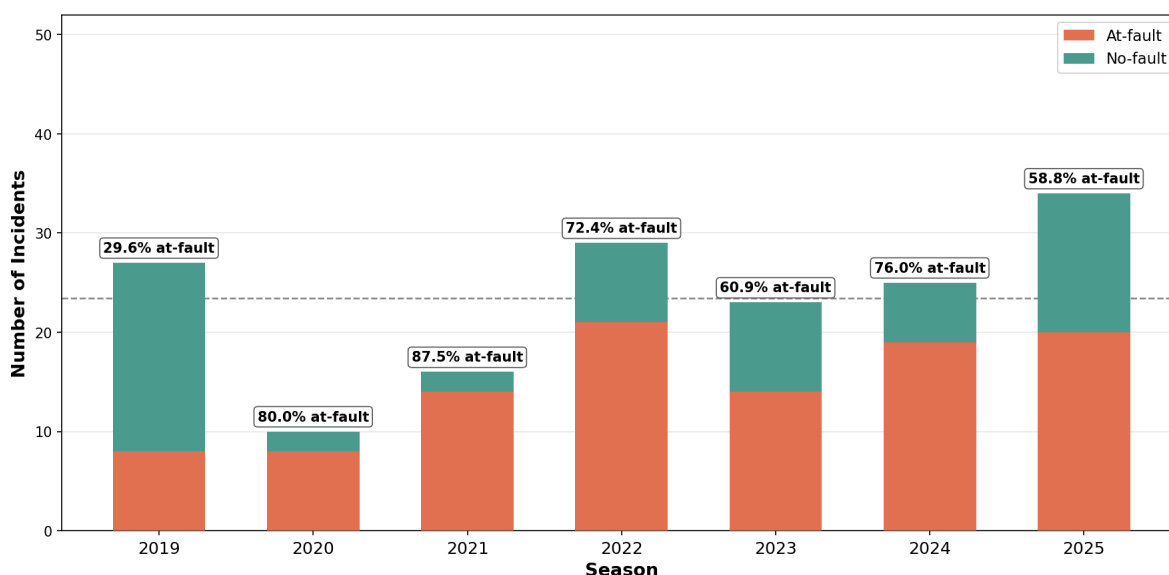


Figure 1. Per-season incident distribution ($n=164$), Percentages show at-fault rate

3.2 Raw telemetry characteristics

The FastF1 API returns telemetry at the native timing resolution of the F1 data feed, which varies across sessions and seasons. Sampling rates range from 7.5 Hz to 1000 Hz, with a median ~ 500 Hz, reflecting infrastructure changes over the seven-season span. Position data (X, Y coordinates) is available for 150 of 164 incidents (91.5%), which is what enables collision-point detection for those cases.

3.3 Collision-centered telemetry processing

Raw telemetry of variable length and variable sampling rate cannot be fed directly to standard classifiers. We developed a four-stage pipeline to produce consistent feature representations (Figure 2).

Stage 1: Collision detection. For each incident with position data (150 of 164), we computed the frame-by-frame Euclidean distance between the two cars and identified the collision point as

the frame with minimum inter-car distance. For the 14 incidents lacking position data, we used the temporal midpoint of the telemetry as a fallback.

Stage 2: Window extraction. We extracted a symmetric six-second window (± 3 seconds) centered on the estimated collision point from each telemetry channel for both cars.

Stage 3: Resampling. Each channel within the window was resampled to a uniform 600 timesteps via linear interpolation, removing the effect of sampling-rate variability.

Stage 4: Feature computation. From each of the 18 resampled channels (nine per car), we computed 11 statistical features: mean, standard deviation, minimum, maximum, range, 25th and 75th percentiles, interquartile range, median, variance, and linear slope. We also computed three inter-car difference features (mean difference, standard deviation difference, maximum absolute difference) for each of the nine matched channel pairs. The total is $(9 \times 2 \times 11) + (9 \times 3) = 225$ features.

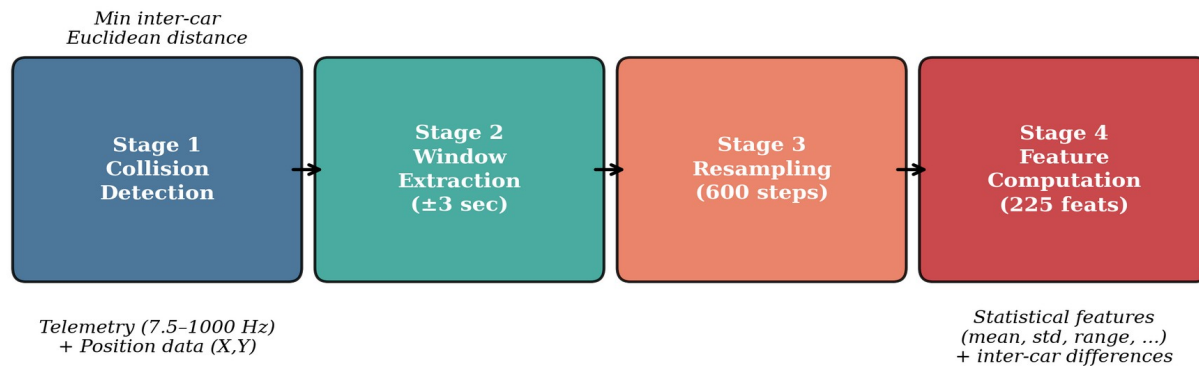


Figure 2. Collision-centered feature engineering pipeline

3.4 Normalization scope and leakage prevention

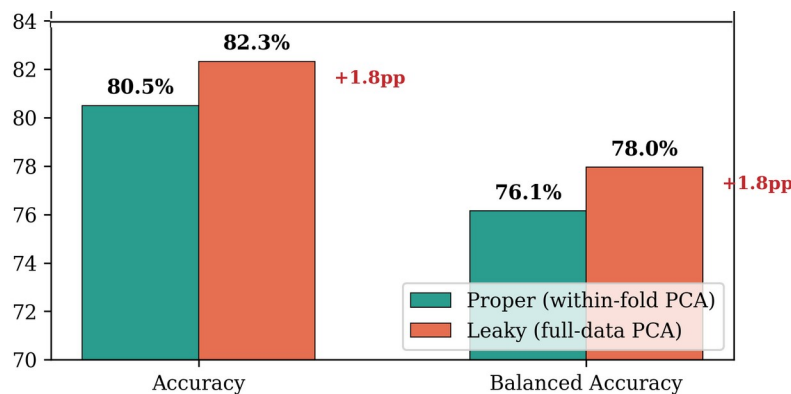


Figure 3. PCA(50) + RF: within-fold versus leaky PCA (1.8 pp gap confirms minimum leakage)

Four of the nine telemetry channels (speed, throttle, brake, and RPM) arrived from the FastF1 API already min-max normalized at the session level, using the full session range. All data-dependent transformations were fitted exclusively within training folds. StandardScaler was fitted on training-fold features and applied to the test fold. For PCA

experiments, PCA was likewise fit only on training-fold data within each CV split. A deliberate leakage test confirmed only 1.8 percentage points (pp) inflation from improper fitting (Figure 3).

3.5 Telemetry channels

Table 1. Telemetry channel definitions and derivations.

Channel	Description / Derivation
speed_normalized	Min-max normalized speed within session; re-normalized within collision window.
throttle_normalized	Throttle application (0-100 mapped to 0-1), from the F1 telemetry feed.
brake_normalized	Brake application (0/100 in most sessions, binary). Normalized to 0-1.
rpm_normalized	Engine RPM, min-max normalized within session.
speed_change	First-order difference of speed: $\text{speed}[t] - \text{speed}[t-1]$. Positive = acceleration.
distance_change	First-order difference of cumulative distance. Proportional to instantaneous speed.
gear	Current gear (integer 0-8, 0 = neutral). Taken directly from the telemetry feed.
drs	DRS activation state (integer). 0-1 mapped from activation codes.
cornering	Binary indicator: $ \text{speed_change} < 1.0$ AND $\text{speed} < \text{session median speed}$.

3.6 Dimensionality reduction strategy

The 225-to-164 feature-to-sample ratio presented a real overfitting concern. We addressed it through PCA dimensionality reduction (10-100 components), mutual-information feature selection (20-100 features), and L1/L2/elastic-net regularized linear models. PCA(50)+RF matched or slightly exceeded full-feature RF, and mutual-information selection down to 20 features retained 78.1% accuracy, so the discriminative signal was unlikely to be a noise-memorization artifact.

3.7 FIA-aligned interaction features

The core limitation of the original 225-feature set is that it summarizes each car independently through univariate telemetry statistics (e.g.,

means, variances, and slopes) rather than representing the spatial and temporal interactions between the two cars. FIA steward decisions are governed by relational, rule-based criteria: significant overlap, entitlement to space, avoidability, right-of-way. These can be approximated with telemetry-derived interaction features. We defined 26 FIA-aligned features (Table 2) by mapping FIA decision principles onto computable telemetry quantities. A review identified 25 candidate interaction features (referred to here as #1-#25); we implemented 23 of them, approximating steering change rate (#21) via position-derived lateral jerk (F25, $r = +0.685$), and added 8 features selected for direct correspondence to FIA fault-attribution criteria, plus lateral gap

closing rate (F26) from the FIA "leave space" steering magnitude (#16, a different feature principle, yielding 26 features in total. Note that F-numbered labels (F01-F26) denote implemented F16) required a steering-angle channel absent from the FastF1 standard implementation IDs in the adopted feature set output. The 26 features cover five categories: and are distinct from the expert-review spatial overlap geometry (F01-F06), braking candidate numbers (#1-#25). Two of the 25 dynamics (F07-F09), temporal proximity (F10-candidates were excluded: track position F12), avoidance behavior (F13-F15), and car classification (#7) required a circuit geometry control (F16-F26). map not available in FastF1, and late evasive

Table 2. The 26 FIA-aligned interaction features mapped to FIA decision principles.

ID	Feature name	Computation	FIA principle	Category
F01	Lateral offset at collision	$ \Delta d[\text{col}] / \text{car_width}$	Significant overlap	Spatial overlap
F02	Overlap ratio at turn-in	$(L - \Delta s[\text{ti}]) / L, \in [0,1]$	Right to contest corner	Spatial overlap
F03	Heading angle difference	$\arccos(\mathbf{v}_a \cdot \mathbf{v}_b / (\mathbf{v}_a \mathbf{v}_b))$ at turn-in	Trajectory alignment	Spatial overlap
F04	Distance at first overlap	$\text{dist}[\text{first } \Delta s < L] / 20\text{m}$	Right-of-way onset	Spatial overlap
F05	Time side-by-side fraction	$\text{mean}(\Delta s < \text{car_length})$	Sustained overlap	Spatial overlap
F06	Apex lead position	$\Delta s[\text{apex}] / \text{car_length}$	Corner ownership	Spatial overlap
F07	Relative speed at brake onset	$\text{speed_A}[\text{t}^{\text{br}}] - \text{speed_B}[\text{t}^{\text{br}}]$	Late braking	Braking dynamics
F08	Delta braking onset	$(\text{t}^{\text{br}}_{\text{A}} - \text{t}^{\text{br}}_{\text{B}}) / T$	Divebomb / late braking	Braking dynamics
F09	Peak deceleration diff	$\min(\text{sc_A}) - \min(\text{sc_B})$	Commitment to braking	Braking dynamics
F10	TTC at turn-in	$\text{dist}[\text{ti}] / \text{closing speed}, \in [0,1]$	Feasibility of manoeuvre	Temporal proximity
F11	Speed differential at turn-in	$\text{speed_A}[\text{ti}] - \text{speed_B}[\text{ti}]$	Speed advantage	Temporal proximity
F12	Rate of closing speed	$\nabla(\text{closing speed})[\text{ti}]$	Aggression of approach	Temporal proximity
F13	Reaction delay B to A	$(\text{response_B} - \text{t}^{\text{br}}_{\text{A}}) / T$	Avoidable contact	Avoidance behavior
F14	Throttle lift timing diff	$(\text{t_lift_A} - \text{t_lift_B}) / T$	Avoidance effort	Avoidance behavior
F15	B brake active at A's turn-in	$\text{brake_B}[\text{ti_A}]$	Opponent awareness	Avoidance behavior
F16	Lateral clearance at apex	$ \Delta d[\text{apex}] / \text{car_width}$	Leave-space obligation	Car control
F17	Mean curvature of A's trajectory	$\text{mean}(\kappa_{\text{A}})$ across window	Path stability	Car control
F18	Trajectory deviation from line	$\text{mean deviation from B's path} / 5\text{m}$	Maintaining position	Car control
F19	Corner entry speed vs mean	$\text{speed_A}[\text{ti}] - \text{mean}(\text{speed_A})$	Speed control	Car control
F20	Throttle-brake overlap	$\text{mean}((\text{thr_A} > 0.2) \& (\text{brk_A} > 0.1))$	Car control / panic	Car control
F21	Throttle persistence before braking	$\text{mean}(\text{thr_A} > 0.5)$ up to t^{br}	Late commitment	Car control
F22	Initiator flag	$\text{sign}(\text{first lateral velocity change})$	Sequential causality	Car control
F23	Closing angle convergence	$\cos\theta = (\mathbf{v}_a - \mathbf{v}_b) \cdot \mathbf{r} / (\mathbf{v}_a - \mathbf{v}_b \mathbf{r})$ at turn-in	Directional aggression	Car control
F24	Control input variance	$\text{var}(\text{thr_A}) + \text{var}(\text{brk_A})$ over window	Erratic vs smooth driving	Car control
F25	Lateral jerk at turn-in	$\text{mean}(d^2(\Delta d)/dt^2)$ near turn-in; pos. proxy for #21	Predictability / sudden shift	Car control
F26	Lateral gap closing rate	$\text{mean}(d \Delta d /dt)$ from turn-in to apex; $-0 = \text{squeezing}$	Leave-space obligation	Car control

A track-aligned decomposition was applied to the relative position vector between the two cars, yielding longitudinal (Δs , along-track) and lateral (Δd , across-track) components. The local track direction was approximated as the average velocity of both cars at each time: $\hat{t} = (v_a + v_b) / \|v_a + v_b\|$, with the lateral direction defined as perpendicular to $\hat{t}$.

The turn-in point was identified as the first timestep where the speed derivative $sc[t] < -0.005$ and the brake pressure exceeded 0.1. If no such point existed within the window, the first-third of the window was used as a fallback. The apex was identified as the time of maximum trajectory curvature $\kappa = |x'y'' - y'x''| / (x'^2 + y'^2)^{3/2}$. Braking onset for each car was the first frame where brake pressure exceeded 0.1; throttle lift was the first frame where throttle dropped below 0.5.

The FIA-26 feature set is a theoretically motivated, regulation-derived approximation of steward reasoning: each feature was constructed by mapping a specific clause or principle in the FIA Sporting Regulations (1) onto a computable telemetry quantity (Table 2). We did not independently validate the correspondence between FIA-26 feature values and the actual reasoning stewards record in their written decisions. The steward decision PDFs underlying our labels (Section 3.1, Section 8) were used only to extract the binary at-fault outcome, not to systematically code language such as “significant overlap,” “entitlement to racing room,” “avoidable contact,” “late braking,” “failure to leave space,” or “loss of control” for comparison against feature values. We did not pursue this analysis because steward

reports vary widely in length and specificity and rarely quantify the concepts they invoke, which would make a systematic content analysis at this sample size (164 incidents) unreliable rather than informative. We therefore present FIA-26 as a conceptual approximation of steward criteria rather than as an empirically validated representation of steward decision-making, and we calibrate the conclusions we draw from it accordingly (Section 6.6, Limitations). Coding steward-report language against FIA-26 feature values in a larger, multi-season corpus is an important direction for future work.

The full computation recipe for all 26 features, including pseudocode and formulas for each step, is provided with the submission code repository (Data availability section). All features are computable from the same FastF1 telemetry channels already used in the original pipeline; no additional data sources are required. Implementation uses only numpy and pandas.

4. Experimental setup

4.1 Classification models

We evaluated four classifiers from scikit-learn (16): Random Forest ($n_estimators = 200$), Gradient Boosting (50 trees, depth 3, $lr = 0.1$), Logistic Regression (liblinear, L2, $C = 2.01$), and SVM (RBF kernel with Platt scaling). We deliberately avoided extensive hyperparameter tuning: with only 164 samples and 225 features, any substantial tuning procedure risked overfitting the validation set.

ROCKET and MiniROCKET apply large banks of random convolutional kernels to the 18-channel, 600-timestep collision-centered

windows. The kernel-transformed features were standardized and classified with RidgeClassifierCV.

For the FIA-aligned feature experiments, we used L2-regularized logistic regression ($C = 2.0$) implemented via gradient descent in numpy, applied to three feature sets: (i) the original 225 statistical features, (ii) the 26 FIA-aligned features only, and (iii) the combined 251-feature set.

4.2 Baseline methods

Two non-informative baselines anchored our results. The majority-class baseline always predicted the at-fault class, achieving 63.4% accuracy but 50.0% balanced accuracy by construction. The stratified random baseline samples predicted in proportion to the training-set class frequencies. We also tested four parameter-free domain heuristics grounded in racing intuition: brake differential, speed differential, throttle differential, and a combined majority-vote heuristic. All four performed below the majority-class baseline, confirming that fault attribution could not be reduced to any single telemetry differential.

4.3 Metrics

Balanced accuracy (unweighted mean of per-class recall) was adopted as the primary metric given the class imbalance. Accuracy measures overall correctness but is misleading here, since always predicting at-fault achieved 63.4%. Matthews correlation coefficient (MCC) provides a single-number summary of classifier quality that is insensitive to class imbalance. AUROC measures the probability that the

model ranks a randomly chosen at-fault incident above a randomly chosen no-fault one; AUPRC accounts for the class imbalance more directly. Brier score measures probabilistic forecast quality. Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) quantify how closely predicted probabilities match observed outcome frequencies. McNemar exact binomial tests (17) were used for pairwise model comparisons under stratified 10-fold CV. Wilson score intervals (18) provided 95% confidence intervals for per-season LOSO accuracy.

4.4 Distribution-shift correction

We tested five correction strategies: default threshold ($t = 0.5$), training-prior threshold, oracle season-prior threshold, BBSE correction, and grid-search threshold tuning. All were evaluated under LOSO. The oracle threshold uses the test season's true class distribution as an impractical upper bound, showing the maximum possible benefit of perfect prior knowledge.

5. Results

5.1 Domain heuristic baselines

Every domain heuristic falls below the majority-class baseline, with accuracy (Acc) ranging from 43.3% to 47.0% and balanced accuracy (BAcc) from 41.2% to 45.5%. In other words, no single telemetry differential captures what stewards actually weigh when assigning fault. The driver who brakes harder, arrives faster, or holds the throttle longer is not systematically the one who is penalized, which is consistent with the multifactorial nature of steward decisions.

Table 3. Domain heuristic baselines (parameter-free rules).

Heuristic	Acc	BAcc	MCC	vs Maj	Definition
Brake differential	0.470	0.444	-0.164	-0.164	Higher brake = fault
Speed differential	0.470	0.455	-0.164	-0.164	Faster car = fault
Throttle differential	0.445	0.432	-0.189	-0.189	More throttle = fault
Combined vote (3)	0.433	0.412	-0.201	-0.201	Majority of above
Majority class	0.634	0.500	0.000	ref	Always at-fault

5.2 Leave-One-Season-Out evaluation (original features)

Table 4. LOSO results (original 225 features). Asterisked (*) values unchanged from default model.

Model / Method	Acc	BAcc	AUROC	AUPRC	Brier	MCC
RF default (t=0.5)	0.704	0.558	0.448	0.655	0.242	0.153
RF BBSE-estimated	0.689	0.546	0.448*	0.655*	0.239	+0.136
RF grid-search	0.643	0.577	0.448*	0.655*	0.242*	+0.062
GB default (t=0.5)	0.666	0.605	0.528	0.702	0.262	0.229
GB grid-search	0.665	0.620	0.528*	0.702*	0.262*	+0.249
ROCKET	0.628	0.537	0.390	0.554	0.342	+0.101
MiniROCKET	0.575	0.493	0.508	0.658	0.384	-0.016

**Figure 4.** Per-season LOSO results for Random Forest

LOSO provides the most operationally relevant ones. RF achieved 70.4% accuracy but only assessment, since a deployed system would be 55.8% balanced accuracy under LOSO, trained on historical seasons and applied to new meaning it defaults to the majority class on most

predictions (per-season detail in Figure 4). Gradient Boosting demonstrated somewhat better balanced accuracy (60.5%) despite lower raw accuracy (66.6%). Prior-shift correction methods were uniformly ineffective. Neither BBSE nor threshold recalibration consistently improved balanced accuracy, and even the oracle threshold, which assumed perfect knowledge of the test-season class distribution, produced only marginal gains. This pattern indicates that the distribution shift across seasons is not merely a label-shift problem where $P(Y)$ changes but $P(X|Y)$ stays fixed. The feature distributions themselves shifted as driving styles and steward interpretations evolved.

5.3 Per-season LOSO breakdown

Table 5. Per-season leave-one-season-out (LOSO) results for Random Forest (RF), Gradient Boosting (GB), and ROCKET (RKT). n = number of incidents; **Prev** = at-fault prevalence; **MajAcc** = majority-class accuracy; **Acc** = accuracy; **BAcc** = balanced accuracy. Calculated on original features.

Year	n	Prev	MajAcc	RF Acc	RF BAcc	GB Acc	GB BAcc	RKT Acc	RKT BAcc
2019	27	0.30	0.704	0.444	0.605	0.370	0.553	0.444	0.605
2020	10	0.80	0.800	0.800	0.500	0.800	0.688	0.800	0.688
2021	16	0.88	0.875	0.875	0.500	0.875	0.714	0.750	0.429
2022	29	0.72	0.724	0.724	0.500	0.690	0.554	0.552	0.381
2023	23	0.61	0.609	0.609	0.540	0.565	0.524	0.565	0.464
2024	25	0.76	0.760	0.800	0.640	0.800	0.697	0.800	0.697
2025	34	0.59	0.588	0.676	0.618	0.559	0.507	0.647	0.593

5.4 Venue-grouped and event-grouped cross-validation

Table 6. Venue-grouped and event-grouped 5-fold CV (original features).

Model	Venue Acc	Venue BAcc	Event Acc	Event BAcc	Δ BAcc
Majority	0.634	0.500	0.634	0.500	ref
RF	0.616	0.522	0.616	0.522	+0.022
GB	0.610	0.519	0.610	0.519	+0.019
SVM	0.659	0.578	0.659	0.578	+0.078
LogReg	0.505	0.473	0.505	0.473	-0.027

Under both grouping strategies, no model reliably exceeded the majority-class baseline. SVM reached the highest balanced accuracy at 57.8%, while RF, the strongest model under stratified CV, dropped to 52.2% balanced accuracy, which is essentially random.

5.5 Stratified 10-fold cross-validation (within-distribution, original features)

Under stratified 10-fold cross-validation (Table 7), which mixes incidents from all seasons and therefore represents an optimistic within-distribution evaluation, Random Forest achieved the strongest overall performance with 79.3% accuracy and 75.9% balanced accuracy, followed by ROCKET (72.9%) and Gradient Boosting (70.1%). Logistic Regression and SVM performed less well, suggesting that nonlinear models better capture the relationships present in the statistical telemetry

features. Calibration quality varied across classifiers: Gradient Boosting had the lowest Expected Calibration Error (ECE = 0.150), whereas Random Forest exhibited the best discrimination (Brier score = 0.184) despite poor maximum calibration (MCE = 0.960). Because this evaluation mixes seasons and venues across training and test folds, these results represent an upper bound on achievable performance and are contrasted with the more deployment-relevant cross-season results in Section 5.2.

Table 7. Stratified 10-fold CV results (original 225 features). Upper bound; mixes seasons and venues.

Model	Acc	BAcc	AUROC	AUPRC	F1(W)	MCC	Brier	ECE	MCE
Majority	0.634	0.500	0.500	0.634	0.492	0.000	0.366	0.366	0.366
LogReg	0.640	0.635	0.639	0.724	0.638	0.263	0.288	0.262	0.572
RF	0.793	0.759	0.703	0.691	0.785	0.550	0.184	0.225	0.960
GB	0.727	0.701	0.701	0.720	0.723	0.417	0.207	0.150	0.411
SVM	0.727	0.649	0.633	0.697	0.702	0.362	0.205	0.121	0.537
ROCKET	0.750	0.729	0.717	0.801	0.756	0.414	0.229	0.200	0.727
MiniROCKET	0.702	0.695	0.748	0.837	0.729	0.381	0.255	0.207	0.708

5.6 Regularization and feature selection (original features)

Regularization and dimensionality-reduction analyses indicate that the strong within-distribution performance of the original feature set is not solely a consequence of overfitting the high-dimensional feature space (Table 8). PCA reduced the feature set from 225 to as few as 50 components while maintaining or slightly improving Random Forest performance, with PCA(50) and PCA(100) achieving balanced accuracies of 76.1% and 75.2%, respectively. Similarly, mutual-information feature selection

retained competitive performance even with only 20 selected features (74.6% balanced accuracy), indicating that much of the predictive signal is concentrated in a relatively small subset of features. In contrast, L1- and L2-regularized logistic regression performed substantially worse than tree-based methods, suggesting that the decision boundary is not well approximated by a linear model. Overall, these results indicate that the original feature representation contains genuine predictive information under within-distribution evaluation, and that its poor cross-season performance arises primarily from

distribution shift rather than excessive feature dimensionality or memorization.

Table 8. Regularization and feature selection analysis (stratified 10-fold CV).

Configuration	Acc	BAcc	AUROC	vs RF(full)	Features
RF (225 feats)	0.793	0.759	0.699	ref	225
PCA(50)+RF	0.805	0.761	0.733	+0.012	50
PCA(100)+RF	0.810	0.752	0.763	+0.017	100
MI(50)+RF	0.781	0.742	0.696	-0.012	50
MI(20)+RF	0.781	0.746	0.672	-0.012	20
L1 (C=10)	0.621	0.624	0.634	-0.172	77
L2 (C=10)	0.647	0.647	0.683	-0.146	225

5.7 Collision-centering sensitivity

Shifting the collision-point estimate by up to one second in either direction changed accuracy by at most 1.2 percentage points and balanced accuracy by at most 1.0 percentage points (Figure 5). The six-second window was wide enough that moderate mis-centering did not materially affect the feature distributions, hence the collision-centering pipeline was robust to timing imprecision in the minimum-distance proxy.

5.8 Class-weighted training

Inverse-frequency class weighting had only a marginal effect on model performance. RF balanced accuracy decreased slightly from 75.9% to 74.2%; Logistic Regression saw the largest gain at 1.6 percentage points. The 63.4%/36.6% class split was evidently not severe enough to warrant cost-sensitive training for tree ensembles, which already handle moderate imbalance through their splitting criteria.

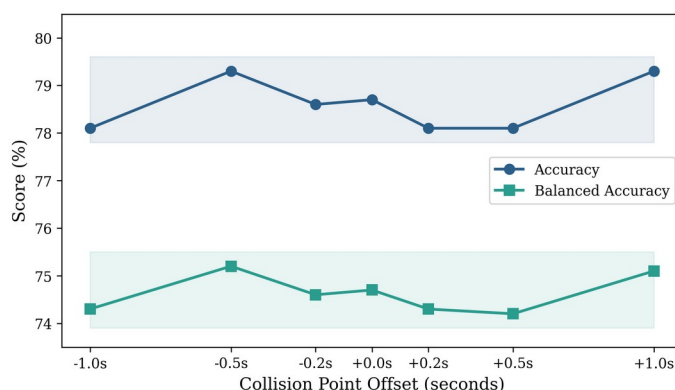


Figure 5. Collision-centering sensitivity (RF, stratified 10-fold), Maximum variation: 1.2pp accuracy, 1.0pp balanced accuracy

5.9 Calibration analysis

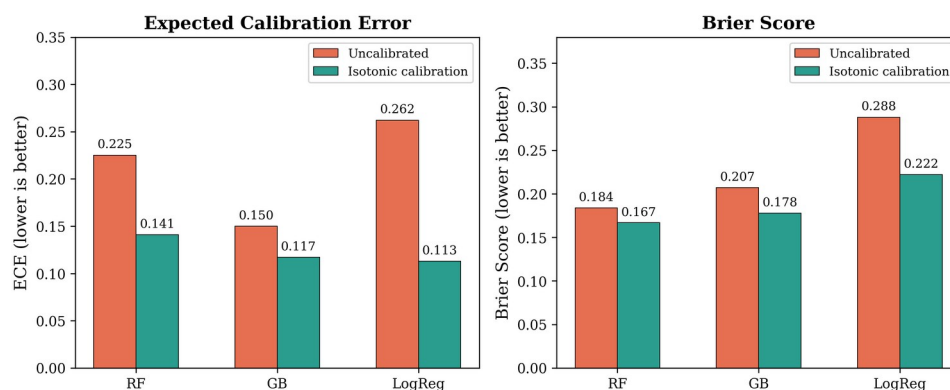


Figure 6. Calibration metrics before and after isotonic calibration

Uncalibrated GB showed the lowest ECE (0.150) and MCE (0.411). RF, despite strong discrimination (Brier 0.184), exhibited an MCE of 0.960, indicating that in at least one probability bin the predicted confidence was nearly 100% off from the true proportion. Isotonic calibration reduced ECE across all models (RF: 0.225 $\rightarrow$ 0.141; GB: 0.150 $\rightarrow$ 0.117; LogReg: 0.262 $\rightarrow$ 0.113), with a similar pattern for observed for the Brier score (Figure 6).

5.10 Feature set comparison (session-start vs collision-centered)

Under stratified CV, collision-centered and session-start features achieved nearly identical accuracy (RF: 79.3% for both). Under LOSO, collision-centered features yielded +9.8 percentage points of accuracy and +11.9 percentage points of balanced accuracy, which suggests that centering the window on the collision captured behaviorally relevant signal

that carried over to unseen seasons.

5.11 FIA-26 feature Set: real-telemetry results

Table 9 reports LOSO and stratified 10-fold cross-validation results for all 26 FIA-aligned interaction features applied to the full real F1 telemetry dataset of 164 incidents (2019–2025). Features were computed from real FastF1 position and telemetry data using the pipeline in Section 3.7; the 14 incidents without position data received median-imputed values. Four classifiers were evaluated under within-fold standardization: majority-class baseline, logistic regression (L2, $C = 2.0$, gradient descent), linear discriminant analysis (LDA), and random forest (150 trees, max depth = 5, $\sqrt{d}$ features per split). Note that the RF here uses a depth-constrained configuration to ensure a fair regularized comparison across feature sets; it differs from the unconstrained scikit-learn RF in Section 5.2, which achieves 55.8% LOSO balanced accuracy on the original 225 features.

Table 9. FIA-26 real-telemetry results vs. original 225-feature set (Orig-225). BAcc=balanced accuracy; AUROC=area under ROC curve. Results averaged over 7 LOSO folds or 10 stratified folds ($\pm$ std). Bold rows show strongest LOSO result. Note: Combined-251 (Orig-225 + FIA-26, 251 features) achieves LOSO RF BAcc=49.6% and Strat-10CV RF BAcc=75.1%; these do not improve on FIA-26 alone, confirming parsimony of the FIA-aligned feature set. See Section 5.11 discussion.

Protocol / Model	Feature Set	BAcc	AUROC	Δ BAcc vs Orig	Orig BAcc (depth-constrained RF)
LOSO / Majority	FIA-26	0.500 $\pm$ 0.000	0.500	0.000	0.500
LOSO / LogReg	FIA-26	0.523 $\pm$ 0.039	0.577	-0.005	0.528
LOSO / RF	FIA-26	0.500 $\pm$ 0.045	0.575	+0.041 $\uparrow$	0.459
LOSO / LDA	FIA-26	0.500 $\pm$ 0.000	0.584	N/A	N/A
Strat-10CV / Majority	FIA-26	0.500 $\pm$ 0.000	0.500	0.000	0.500
Strat-10CV / LogReg	FIA-26	0.592 $\pm$ 0.058	0.683	-0.057	0.649
Strat-10CV / RF	FIA-26	0.729 $\pm$ 0.116	0.797	-0.030	0.759

The key results across all three feature sets are shown in Table 9. FIA-26 achieved 72.9% stratified 10-fold balanced accuracy (RF) and 52.3% LOSO balanced accuracy (Logistic Regression, AUROC = 0.577). The combined 251-feature set reaches 75.1% stratified balanced accuracy (RF) but only 49.6% LOSO balanced accuracy, slightly below FIA-26 alone at 50.0%. The original 225 features performed the worst under LOSO at 45.9% balanced accuracy (RF). Three results confirm that the original features failed to capture the car-to-car relational quantities that generalize across seasons. With FIA-26, RF exceeded chance in two of seven seasons: 2019 (BAcc = 55.3%, AUROC = 0.681) and 2025 (BAcc = 57.1%, AUROC = 0.643). Logistic Regression achieved the best cross-season result overall at 52.3% LOSO balanced accuracy (AUROC = 0.577). The remaining seasons collapsed to the majority-class baseline, consistent with small per-season test sizes ($n = 10$ -34) and class

imbalance up to 88% at-fault in 2021. The combined 251-feature set did not improve over FIA-26 under LOSO (49.6% vs. 50.0%), showing that the additional 225 ad-hoc features add noise that impedes cross-season generalization, supporting the use of the FIA-aligned set. The Strat-10CV to LOSO balanced-accuracy drop for RF fell from 30.0 percentage points with the original features (75.9% to 45.9%) to 22.9 points with FIA-26 (72.9% to 50.0%), indicating better cross-season generalization. AUROC for RF under LOSO was 0.575, which is consistent with a real, if modest, discriminative signal given the constraints of small per-season test sets and subjective labels. Absolute performance remained below operational utility; as discussed in Section 6.1, the residual gap reflected concept shift from evolving steward standards and information that telemetry simply cannot provide. Per-season RF LOSO detail is shown in Table 10.

Table 10. Per-season LOSO detail, Random Forest, FIA-26 features. Prev=at-fault prevalence; Orig BAcc and Δ BAcc unavailable at per-season granularity from existing results (aggregate Orig LOSO RF BAcc=0.459).

Season	n	Prev	Acc	BAcc	AUROC	Orig BAcc	Δ BAcc
2019	27	0.30	0.370	0.553	0.681	N/A	N/A
2020	10	0.80	0.800	0.500	0.312	N/A	N/A
2021	16	0.88	0.875	0.500	0.786	N/A	N/A
2022	29	0.72	0.621	0.467	0.560	N/A	N/A
2023	23	0.61	0.478	0.433	0.532	N/A	N/A
2024	25	0.76	0.720	0.474	0.509	N/A	N/A
2025	34	0.59	0.647	0.571	0.643	N/A	N/A

5.12 Feature importance and interpretability

We computed permutation importance and SHAP (SHapley Additive exPlanations) values for the FIA-26 feature set using the same Random Forest configuration as in Table 9 (150 trees, max depth = 5) to determine whether the predictive signal reported in Section 5.11 originated from a few features or was spread evenly across the set. Permutation importance is the mean drop in held-out balanced accuracy across 30 repeats per fold, averaged over the 10 stratified folds; SHAP values use TreeExplainer on a model fit to the full standardized dataset and are reported as the mean absolute SHAP value for the at-fault class. Table 11 reports the top 10 features by each method (Figure 7 plots the same results).

Both methods identified F07 (relative closing speed at brake onset, a braking-dynamics feature) as the strongest contributor by both permutation importance (mean balanced-accuracy drop = 0.062) and mean |SHAP value| (0.054), ~ double the next-ranked feature. F18 (trajectory deviation) and F05 (time spent side-by-side) ranked in the top three under both methods, alongside F14 (throttle-lift timing

difference) and F23 (closing angle). The top five features by permutation importance accounted for ~ two thirds of the total importance mass across all 26 features, indicating that the predictive signal documented in Section 5.11 is concentrated in a small subset of braking-dynamics and spatial-overlap features rather than diffused evenly across the FIA-26 set.

This importance profile contrasts with the univariate point-biserial correlations reported for the eight-feature subset in Section 6.2. Features such as control input variance and lateral jerk exhibited relatively strong individual correlations with the fault label but ranked well below the top five under multivariate permutation and SHAP analyses of the full FIA-26 feature set. Conversely, F07 (relative speed at brake onset), the strongest multivariate predictor, showed little prominence in the univariate analysis. This divergence illustrates that features associated with fault attribution in isolation do not necessarily provide independent predictive information once correlated variables are considered jointly. In particular, braking-dynamics and spatial-overlap features account for much of the predictive signal, rendering

several car-control features largely redundant in predictive performance is driven primarily by a the multivariate model. The concentration of few FIA-aligned interaction variables rather importance within a small subset of relational than by broadly distributed contributions across features therefore suggests that cross-season all 26 features.

Table 11. Top 10 FIA-26 features by permutation importance and mean |SHAP value| (Random Forest, 150 trees, max depth = 5, stratified 10-fold CV).

Rank	Feature (permutation importance)	Perm. importance	Feature (mean SHAP)
1	F07_rel_speed_brake_onset	0.0620	F07_rel_speed_brake_onset (0.0542)
2	F05_time_sidebyside	0.0369	F18_traj_deviation (0.0361)
3	F01_lateral_offset_coll	0.0318	F05_time_sidebyside (0.0247)
4	F18_traj_deviation	0.0293	F11_speed_diff_turnin (0.0211)
5	F14_throttle_lift_diff	0.0272	F23_closing_angle (0.0176)
6	F23_closing_angle	0.0198	F12_rate_closing (0.0159)
7	F02_overlap_ratio_turnin	0.0192	F14_throttle_lift_diff (0.0139)
8	F04_dist_first_overlap	0.0187	F13_reaction_delay (0.0138)
9	F13_reaction_delay	0.0184	F10_ttc_turnin (0.0122)
10	F24_control_variance	0.0145	F03_heading_angle_diff (0.0118)

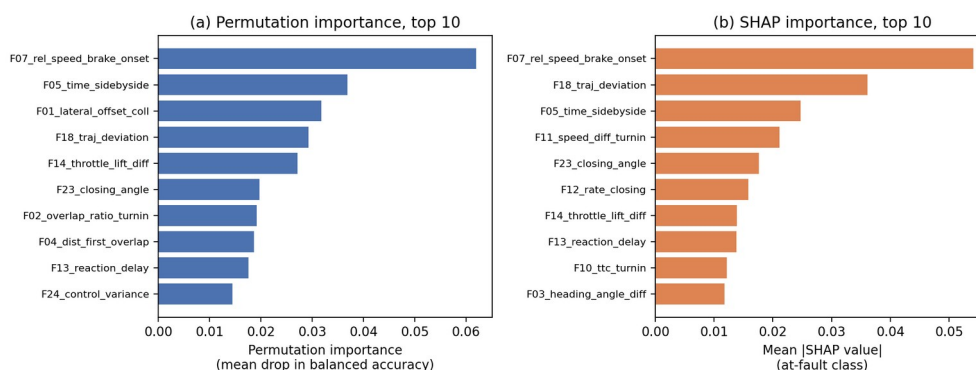


Figure 7. Permutation importance (a) and SHAP importance (b) for the top 10 FIA-26 features

6. Discussion

6.1 Cross-season generalization

The gap between within-distribution and out-of-distribution performance is the central finding (Figure 8). RF achieved 79.3% accuracy under

stratified CV but only 70.4% under LOSO and 61.6% under venue-grouped CV. Balanced accuracy dropped from 75.9% to 55.8% (LOSO) and 52.2% (venue-grouped). PCA experiments confirm this gap was not caused by overfitting in 225 dimensions. The gap instead reflects

distributional shift across seasons and circuits from three sources, each discussed below.

Regulatory evolution: FIA steward interpretation of the Sporting Regulations has shifted over the study period, particularly regarding defensive driving, the obligation to leave space, and the threshold for issuing time versus grid penalties. These interpretive shifts mean that a feature pattern predictive of penalties in 2019 may not be penalized under 2024 or 2025 standards.

Small per-season samples: With 10 to 34 incidents per season, individual season results carry wide confidence intervals, and the

aggregate LOSO metrics depend heavily on performance in the 2019 season, the one year with an inverted class balance. The combination of small sample sizes and shifting class distributions makes standard adaptation methods ineffective.

Non-stationary class distributions: Fault rates vary from 29.6% (2019) to 87.5% (2021) across seasons. Prior-shift methods do not close the gap, indicating that $P(X|Y)$ is not constant across seasons. The way features connect to labels shifts as steward interpretation standards evolve alongside changes in driving behavior and the regulatory environment.

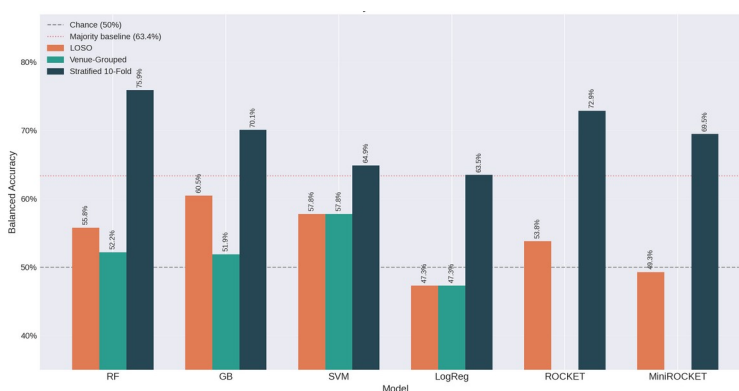


Figure 8. Balanced accuracy across evaluation protocols

6.2 Importance of rule-aligned features

The real-telemetry results in Section 5.11 point to a more fundamental problem than the generalization gap: the original 225 features are not aligned with the decision logic stewards actually use. The core mismatch is between what the model tries to learn ("what telemetry patterns correlate with penalties?") and what FIA stewards actually do ("apply structured rules about overlap, space, and control"). The model is trying to learn a rule-based judgment

process without access to the rule-relevant variables.

The 26 FIA-aligned features introduced in Section 3.7 address this mismatch directly. They encode relational quantities across five categories: spatial overlap geometry (F01-F06), braking dynamics (F07-F09), temporal proximity (F10-F12), avoidance behavior (F13-F15), and car control (F16-F26). The real-telemetry validation in Section 5.11, where FIA-26 outperforms the combined 251-feature

set under LOSO despite using only about 10% as many features, demonstrates that these interpretable variables carry signal the original 225-feature set does not. Individual features such as control input variance and lateral jerk show the strongest univariate correlations with the fault label and directly encode the FIA's "car control" and "predictability" principles. However, Section 5.12 shows that once the full FIA-26 set is available to the model, the multivariate predictive signal is instead concentrated in braking-dynamics and spatial-overlap features (F07, F18, F05), with car-control features contributing comparatively little in addition to that signal. The categories align with FIA principles conceptually, but multivariate feature importance rather than univariate correlation is the more reliable basis for claims about which principles the data actually support, and should guide future feature engineering.

The FIA-aligned features also degrade less under LOSO than the original feature set, narrowing the within-distribution to cross-season gap from 30.0 to 22.9 percentage points and suggesting they capture relational physics that generalizes across steward standards rather than season-specific artifacts. Absolute LOSO performance remains modest, indicating that residual concept shift and the limits of what telemetry alone can capture leave room for the multimodal extensions discussed in Section 7.

6.3 Limits of telemetry-only models

The results in Sections 5.11-5.12 and 6.1-6.2 point toward a limitation in the available information rather than in the modeling approach. FIA stewards do not rely solely on

speed, braking, and positioning channels. Written decisions and the steward process more broadly draw on onboard and broadcast video evidence, broader spatial context beyond the two cars directly involved, judgments about driver intent, expectations about a car's entitlement to a given racing line, situational awareness of blue flags, safety-car status, and track position, precedent from prior rulings at the same circuit or under comparable circumstances, and subjective interpretation of the Sporting Regulations that can vary between steward panels and across seasons (Section 6.1). Few of these factors are recoverable from the FastF1 telemetry channels used in this study (Table 1). Even the FIA-26 features, which are explicitly designed around the regulatory language stewards use (Section 3.7), reach only 52.3% LOSO balanced accuracy, barely above the 50% chance level, and the feature-importance results in Section 5.12 show that this small amount of signal is concentrated in only a handful of features. This pattern is consistent with an information ceiling: the residual gap between within-distribution and LOSO performance, and between LOSO performance and a practically useful classifier, may reflect what telemetry alone cannot encode rather than a shortcoming of feature engineering or model selection. Future systems for this task will likely require richer multimodal inputs beyond telemetry alone: video, track geometry, steward reports, and contextual race information. The FIA-26 results in Section 5.11 provide a quantitative baseline against which such systems can be measured.

6.3.1 *Adaptive feature weighting under concept drift*

Although telemetry cannot capture driver intent, video evidence, race context, or steward deliberation, some degradation in cross-season performance may arise from evolving steward interpretation rather than missing modalities alone. The present study found that correcting only the class prior through BBSE and threshold recalibration produced little improvement, suggesting that concept drift affects the decision function rather than the classification threshold. A potential extension is to estimate season-specific feature importance and identify features whose predictive contribution varies systematically with season-level at-fault prevalence. As steward decisions accumulate during a season, the evolving class prior could serve as an online proxy for the prevailing enforcement environment, enabling adaptive reweighting of policy-sensitive features while leaving rule-invariant features unchanged. Such an approach would test whether changes in steward interpretation can be accommodated through dynamic feature weighting before introducing additional data modalities. Determining whether season-level class prevalence reliably predicts changes in feature importance will require substantially larger multi-season datasets.

6.4 Distribution shift and adaptation

We tested five approaches to handling the LOSO distribution shift: default threshold, training-prior threshold, oracle season-prior threshold, BBSE correction, and grid-search threshold tuning. None yielded consistent improvement, and several actively worsened balanced accuracy. Under the taxonomy of Moreno-Torres et al. (13), BBSE assumes that $P(X|Y)$ is constant across domains and that only

the class prior $P(Y)$ changes. Our data violate this assumption: the way drivers interact and the criteria by which stewards evaluate that interaction both evolve over time. More sophisticated domain-adaptation methods require larger sample sizes than our 164-incident dataset allows.

6.5 Time-series versus engineered features

ROCKET and MiniROCKET achieved within-distribution balanced accuracy comparable to the traditional classifiers (72.9% and 69.5% versus RF's 75.9%), despite using the raw time-series signal rather than engineered features. Both methods degraded sharply under LOSO, falling to balanced accuracies of 53.7% and 49.3% respectively, performing no better than (and in the MiniROCKET case slightly below) the majority-class baseline. The six-second window may be too short for convolutional kernels to capture sequential patterns substantially different from what summary statistics already encode. The FIA-aligned features outperform both by encoding specific events (turn-in, apex, brake onset) rather than treating the window as a uniform time series.

6.6 Limitations

Several limitations bear emphasis. The dataset contained 164 incidents from seven seasons, which constrained both model complexity and evaluation precision. Ground-truth labels reflected subjective steward decisions, not an objective physical standard. The telemetry modality was inherently incomplete: stewards use video footage, driver radio, track-specific precedents, and contextual knowledge that our features could capture (Section 6.3). The FIA-26 feature set itself was not validated against the

text of actual steward reports (Section 3.7); it should be read as a conceptual, regulation-derived approximation of steward reasoning rather than an empirically confirmed representation of the decision process, and the strength of our conclusions is calibrated accordingly. Session-level normalization of speed and RPM channels entangled session identity with feature values, and while within-fold standardization prevented direct leakage, normalization artifacts may still inflate stratified-CV results. The FIA feature computation required position data, which was unavailable for 14 of 164 incidents (8.5%), and absolute LOSO performance with the FIA-26 set remained modest, reflecting concept shift and information limits (Section 6.3) beyond what feature engineering alone could address.

6.7 Ethical considerations

Automated fault attribution in professional sport raises concerns that extend beyond classification accuracy. Any system used in an advisory capacity must be transparent: feature-importance rankings and probability calibration are necessary but not sufficient for explaining individual decisions. Drivers and teams must retain the ability to contest model-informed outcomes. The training labels encode historical steward decisions, which may reflect systematic biases in how incidents involving particular teams or driver profiles have been adjudicated; deploying a model without auditing for such patterns risks perpetuating inequitable treatment. Finally, the LOSO results demonstrate that models trained on past seasons do not reliably generalize to new ones. Operational use would require continuous

monitoring and retraining as regulatory norms evolve.

7. Conclusion

This study reports a feasibility study of machine learning for fault attribution in Formula 1 racing collisions, formalizing the task as binary classification on telemetry data. The central contribution is a negative result with diagnostic value: even after introducing feature engineering explicitly aligned with FIA decision principles, fault attribution remains close to chance under the season-held-out evaluation that mirrors real deployment. Telemetry-only approaches appear to face a fundamental information ceiling rather than a solvable feature-engineering or modeling gap (Section 6.3). Several findings support this conclusion. First, telemetry-only models using generic statistical features do not achieve operationally useful performance under realistic evaluation conditions: under LOSO, RF reaches 70.4% accuracy but only 55.8% balanced accuracy, and under venue-grouped CV, no model reliably exceeds the majority-class baseline of 63.4%. Second, FIA-aligned interaction features, which better reflect the relational quantities stewards reason about, narrow but do not close this gap: LOSO balanced accuracy improves from 45.9% to 50.0% (Random Forest) while remaining barely above chance. Third, feature-importance analysis (Section 5.12) shows that even this modest signal is concentrated in a handful of braking-dynamics and spatial-overlap features, which argues against the view that more exhaustive feature engineering within the telemetry modality alone would meaningfully close the gap. Similarly, correcting only the class prior through BBSE and threshold

recalibration produced little improvement, indicating that cross-season degradation reflects changes in the decision function rather than label prevalence alone.

We introduced a set of 26 FIA-aligned interaction features representing spatial overlap, braking dynamics, temporal proximity, avoidance behavior, and car control. Although these features substantially improved cross-season robustness relative to generic statistical telemetry features, their predictive value remained concentrated in a small subset of braking-dynamics and spatial-overlap variables, reinforcing the conclusion that telemetry alone contains insufficient information for reliable fault attribution.

Within-distribution performance was considerably stronger, with Random Forest achieving 79.3% accuracy under stratified cross-validation. The collision-centered processing pipeline consistently improved generalization relative to session-start windows and remained robust to moderate timing offsets, indicating that behavior immediately preceding impact contains the most informative telemetry signal.

Section 6.3 points to several research directions. Multimodal integration of video footage, broadcast or onboard, track geometry, steward report text, and contextual race information (safety-car status, blue flags, championship stakes) is, in our view, more likely to close the residual gap than further feature engineering within the telemetry modality alone. Coding steward-report language against FIA-26 feature values (not attempted here, Section 3.7) would

test whether the feature set captures actual steward reasoning rather than serving as a conceptual approximation. Dataset expansion as future seasons add incidents is a simpler complementary path; the FIA-aligned feature pipeline and full code are provided for replication as the dataset grows. Temporal adaptation methods designed for small, nonstationary datasets, such as online learning with a sliding window, may be complemented by adaptive feature-weighting schemes that use the evolving season-level class prior as a proxy for changes in steward interpretation (Section 6.3.1). Both approaches warrant investigation given the concept shift documented in Section 6.1. Interpretability beyond the permutation importance and SHAP results in Section 5.12, such as case-based reasoning over similar past incidents, could make a future system more suitable for decision support, provided the information gap identified here is first closed. By establishing a quantitative baseline and explicitly identifying the information limits of telemetry-only models, this work provides a foundation for future research into explainable, multimodal decision-support systems for motorsport officiating.

Acknowledgments

The author thanks the developers of the FastF1 Python library (Oehrly, 2020) for maintaining open access to Formula 1 timing and telemetry data. FIA steward decision documents used for label extraction are publicly available on the FIA website.

The author acknowledges the use of AI language model tools (including ChatGPT, developed by OpenAI) during the preparation and revision of

this manuscript. Specifically, AI assistance was used to (a) explore feature engineering approaches and identify gaps in the original feature set relative to FIA decision criteria, (b) generate pseudocode for the FIA-aligned interaction features, and (c) assist with editorial review of the manuscript text. All technical content, experimental design, results, and scientific conclusions are the sole responsibility of the author. This disclosure follows the Journal's transparency requirements.

Data availability

Telemetry data was extracted using the FastF1 Python API (4). FIA steward decisions are publicly available on the FIA website. Labels were derived by parsing penalty fields from steward decision PDFs: any penalty (time, grid, or penalty points) maps to `at_fault = 1`; racing incident, no further action, or no penalty maps to `at_fault = 0`. All classifiers were implemented in scikit-learn with fixed random seeds (`seed = 42`). The FIA-aligned feature computation code, implemented in numpy and pandas, is provided with the submission and is fully reproducible from any FastF1 telemetry extraction for incidents with available position data. The permutation importance and SHAP analysis in Section 5.12 uses scikit-learn's `permutation_importance` and the shap package's `TreeExplainer` (19), with the same Random Forest configuration and random seed as the main FIA-26 results.

8. References

1. Fédération Internationale de l'Automobile. (2025). *2025 FIA Formula One sporting regulations*. FIA. <https://www.fia.com/regulation/category/110>
2. Murugan, A. A., & Panchalingam, L. (2025). *The final lap of justice: Reforming FIA governance in Formula 1*. *Entertainment and Sports Law Journal*, 23(1), 5. <https://doi.org/10.16997/eslj.1784>
3. Fédération Internationale de l'Automobile. (2023, November 23). *FIA Insights: Computer vision*. FIA. <https://www.fia.com/news/fia-insights-computer-vision>
4. Oehrly. (2020). *Fast-F1: A Python package for accessing and analyzing Formula 1 timing and telemetry data* [Computer software]. GitHub. <https://github.com/theOehrly/Fast-F1>
5. Rabee, A., Anwar, Z., AbdelMoety, A., Abdelsallam, A., & Ali, M. (2025). Comparative analysis of automated foul detection in football using deep learning architectures. *Scientific Reports*, 15(1), 14236. <https://doi.org/10.1038/s41598-025-96945-0>

6. Sasikumar, A., Leema, A. A., & Balakrishnan, P. (2025). Data-driven pit stop decision support for Formula 1 using deep learning models. *Frontiers in Artificial Intelligence*, 8, Article 1673148. <https://doi.org/10.3389/frai.2025.1673148>
7. Hojaji, F., Toth, A. J., Joyce, J. M., & Campbell, M. J. (2024). An AI approach for analyzing driving behaviour in simulated racing using telemetry data. In P. Dondio, M. Rocha, A. Brennan, A. Schönbohm, F. de Rosa, A. Koskinen, & F. Bellotti (Eds.), *Games and Learning Alliance: 12th International Conference, GALA 2023, Dublin, Ireland, November 29–December 1, 2023, Proceedings* (Lecture Notes in Computer Science, Vol. 14475, pp. 194–203). Springer. https://doi.org/10.1007/978-3-031-49065-1_19
8. Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5), 1454–1495. <https://doi.org/10.1007/s10618-020-00701-z>
9. Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). MiniROCKET: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 248–257). Association for Computing Machinery. <https://doi.org/10.1145/3447548.3467231>
10. Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3), 606–660. <https://doi.org/10.1007/s10618-016-0483-9>
11. Fernández, J., Bornn, L., & Cervone, D. (2019, March). *Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer* [Conference paper]. 13th MIT Sloan Sports Analytics Conference, Boston, MA, United States. <https://www.sloansportsconference.com/research-papers/decomposing-the-immeasurable-sport-a-deep-learning-expected-possession-value-framework-for-soccer>
12. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>
13. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1), 521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>

14. Lipton, Z. C., Wang, Y.-X., & Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 80, pp. 3122–3130). PMLR. <https://proceedings.mlr.press/v80/lipton18a.html>

15. Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new *a priori* probabilities: A simple procedure. *Neural Computation*, 14(1), 21–41. <https://doi.org/10.1162/089976602753284446>

16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>

17. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>

18. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>

19. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>