# Auditing Evaluation Leakage in Formula 1 Race-Strategy Machine Learning: A Track-Identity Confound and a Corrected Protocol

Shubham Kumar Gupta
shubham.kumar59@s.amity.edu
ASET, Amity University Noida

**Abstract**

Machine learning models for Formula 1 race strategy are typically evaluated with random train/test splits over lap-level data, which lets adjacent laps from the same race leak between train and test. We quantify this directly on lap time and pit-stop prediction using 2023 FastF1 telemetry (22 races, 20,447 laps). Removing this leakage collapses a naively evaluated XGBoost lap-time model's $R^2$ from 0.994 to $-2.26$ on unseen circuits once track-identity features are also removed, showing that absolute lap time is dominated by track identity rather than the race-condition features (tyre life, stint, weather) typically used to justify these models. A track-aware evaluation (holding out drivers rather than circuits) restores an honest $R^2 = 0.991$ (MAE= 0.54s, vs. a persistence baseline's $R^2 = 0.949$, MAE= 0.61s). The same corrected pipeline yields a pit-stop-next-lap classifier (F1= 0.303, ROC-AUC= 0.888, vs. a logistic regression baseline's ROC-AUC= 0.669) that trails current published deep-learning approaches on a closely related task (Sasikumar et al. 2025: F1= 0.81), which we report as a limitation rather than a contribution. Applying the lap-time model without retraining to 2024 data shows a substantial generalization gap ($R^2$ 0.991 to 0.172; pit-stop ROC-AUC 0.888 to 0.721), suggesting season-over-season performance changes are not captured by the current feature set. We release code, corrected metrics, and a leakage decomposition to support more reproducible evaluation in this domain.

## 1 Introduction

Formula 1 teams make pit-stop and tyre-management decisions under significant time pressure, and a growing body of applied machine learning work uses lap-level telemetry (much of it sourced from the open FastF1 API) to model lap time degradation and pit-stop timing. A recurring methodological pattern in this literature, including in an earlier version of the pipeline this paper corrects, is to evaluate these models with a row-wise random train/test split over individual laps. Because laps from the same race and the same driver's stint are highly autocorrelated, this split lets information leak between train and test: a model can effectively interpolate within a race it has already partially seen, rather than generalize to conditions it has not.

This paper is a leakage audit, not primarily a new modeling technique. Its central contribution is a decomposition that isolates *exactly how much* of a plausible-looking result is attributable to (a) the row-wise split itself and (b) one-hot encoding driver and circuit identity as model features. We show that these two choices, independently or combined, are sufficient to reproduce a near-perfect lap-time $R^2$ that has essentially nothing to do with modeling tyre degradation. We then propose and validate an honest evaluation protocol: grouped cross-validation, with an explicit

distinction between generalizing to an *unseen circuit* (track-agnostic) versus an *unseen driver on a known circuit* (track-aware). We report what a corrected XGBoost pipeline actually achieves under each. We complement this with SHAP-based explainability, a set of non-trivial baselines (persistence, linear/logistic regression, majority class), and a cross-season generalization test that evaluates 2023-trained models on 2024 data without retraining.

We do not claim state-of-the-art predictive performance. On the pit-stop classification task specifically, our corrected classifier trails several recently published results (Section 2), and we report this plainly in Section 5 rather than around it. The contribution we do make is a methodology and an empirical demonstration, on a real dataset, of how large an evaluation-leakage confound can be in this application domain.

## 2 Related Work

**Machine learning for lap-time prediction.** Several recent studies apply machine learning directly to FastF1-sourced lap-time data. Zhao [8] trains a deep neural network to forecast qualifying lap times, categorizing historical data by driver and circuit to produce per-driver, per-track predictions; this specialization is functionally close to encoding driver and circuit identity as features, the same mechanism this paper identifies as the dominant confound in our own leakage decomposition (Section 3.3). Brusik [1] compares univariate and multivariate LSTM and LSTM-FCN architectures for multi-step lap-time forecasting on FastF1 data spanning the 2018–2023 seasons (136,122 laps), using a strict chronological train/test split and including driver and constructor identity among the multivariate model's features; the multivariate models outperform their univariate counterparts, and error analysis identifies track-status changes (yellow flags, safety cars) as the primary source of large forecasting errors across all tested architectures. Neither study's evaluation protocol is directly comparable to ours, since both address multi-step or per-category forecasting rather than the single-lap, race-grouped setting studied here, so we do not attempt a numeric comparison; both are nonetheless consistent with our finding that circuit and driver identity are highly informative for absolute lap time.

**Deep learning for pit-stop prediction.** Sasikumar, Leema, and Balakrishnan [5] benchmark five deep sequence and convolutional architectures (Bi-LSTM, TCN-GRU, GRU, InceptionTime, and CNN-BiLSTM) on FastF1-sourced pit-stop prediction, evaluated on a temporal holdout (the final eight races of the 2024 season), which avoids the within-race leakage this paper is concerned with. Their best model, Bi-LSTM, reaches precision= 0.77, recall= 0.86, F1= 0.81, and ROC-AUC= 0.988 on the pit-stop class. This is a strong result, and our own corrected pit-stop classifier (F1= 0.303, ROC-AUC= 0.888; Section 4) does not match it. We do not attempt to explain this gap away: a temporally-evaluated deep sequence model may simply capture pit-stop timing dynamics that a single-lap tabular gradient-boosted-tree model does not. Older comparators are also ahead of our result on the same broad task: Fatima and Johrendt's Deep-Racing embedded neural network [2] reports F1= 0.67, and García Tejada's SVM-based classifier [3] reports F1= 0.621 for pit-stop occurrence. We discuss this comparison further, and what we think it does and does not imply, in Section 5.

**Explainable F1 strategy models.** Two contemporaneous preprints from an Imperial College London / Mercedes-AMG PETRONAS collaboration take different approaches to interpretability in this domain. Todd et al. [7] apply explainable deep learning and XGBoost to tyre energy degradation prediction, using feature importance and counterfactual explanations. Thomas et al. [6] target the sequential strategy decision directly with a reinforcement learning policy (RSRL) for tyre compound and pit timing, interpreted via feature importance, surrogate decision trees,

and counterfactuals; RSRL achieved an average finishing position of P5.33 on the 2023 Bahrain Grand Prix, ahead of baseline strategies. Our use of SHAP (TreeExplainer) for both the lap-time and pit-stop models is in the same spirit (making model decisions auditable at both the global and individual-prediction level), applied to a simpler supervised tabular formulation rather than a time-series or RL one.

**Positioning.** Given the comparison above, we position this paper's contribution as methodological: a race- and driver-grouped cross-validation protocol, an explicit decomposition of how leaky evaluation inflates results in this specific domain, honest non-trivial baselines, and a cross-season generalization test, rather than a claim of predictive superiority over any cited comparator.

# 3 Methodology

## 3.1 Data

We use the FastF1 Python API [4] to fetch lap-by-lap timing data for all 22 races of the 2023 Formula 1 season (24,420 raw laps). Unlike a prior version of this pipeline, weather is joined properly: for each session, `session.weather_data` is merged onto `session.laps` with `pd.merge_asof` on timestamp (nearest match, 5-minute tolerance) before races are concatenated. For the cross-season generalization test (Section 3.6), we additionally fetch six 2024 races (Bahrain, Japan, Monaco, Great Britain, Singapore, Brazil) chosen to reuse 2023 circuit names, so that circuit-identity features are not evaluated on unseen categories purely as an artifact of race selection.

Laps are cleaned by requiring `IsAccurate`, dropping deleted laps, and dropping rows missing tyre life, stint, lap number, compound, or position, leaving 20,873 laps. Building the PitStopNextLap target (Section 3.2) additionally drops each driver's final lap of each race, since it has no following lap to compare against, leaving the 20,447 laps used throughout the rest of this paper.

## 3.2 Features and targets

Numeric features: tyre life, stint (FastF1's native `Stint` column, not recomputed), lap number, lap-time delta from the previous lap (same driver, same race), an approximate gap to the car ahead (derived from the difference in cumulative session time between adjacent-position cars on the same lap, since FastF1 does not provide this directly), track temperature, air temperature, rainfall, position, humidity, and wind speed. Categorical features: tyre compound and track status, one-hot encoded with the first level dropped.

The regression target is lap time in seconds. The classification target, PitStopNextLap, is defined as $(\text{NextTyreLife} < \text{TyreLife}) \wedge (\text{TyreLife} > 2)$, where NextTyreLife is each driver's tyre life on their following lap, shifted within *(Grand Prix, Driver)* groups rather than by driver alone. Shifting by driver alone incorrectly compares a driver's last lap of one race to their first lap of the next. Under this definition, 707 of 20,447 laps (3.5%) are positive.

## 3.3 Leakage decomposition

A prior version of this pipeline evaluated both models with a row-wise random 80/20 split and one-hot encoded driver and circuit (*GP*) identity as features. To isolate how much each choice contributes to inflated performance, we evaluate four scenarios for the lap-time regression task, using an identical feature pipeline throughout:

**A** Random row-wise split, driver and GP included as one-hot features (reproduces the original leaky setup).

**B** `GroupKFold` grouped by race (5-fold), driver and GP included: GP is then an unseen category for every held-out race, isolating what identity alone contributes once row-wise leakage is removed.

**C** `GroupKFold` grouped by race (5-fold), no driver or GP features at all ("track-agnostic"): tests generalization to a genuinely unseen circuit.

**D** `GroupKFold` grouped by *driver* (5-fold), GP kept as a feature but driver excluded ("track-aware"): tests generalization to an unseen driver on a circuit already represented in training. This is a legitimate deployment framing, since a race team always knows which circuit it is racing at.

Scenario C is our primary track-agnostic result; Scenario D is our primary track-aware result and the version used for SHAP analysis, the deployed model, and the 2024 generalization test. The same track-agnostic/track-aware distinction (Scenarios C and D) is also applied to the pit-stop classifier.

### 3.4 Baselines and model

For lap-time regression we compare XGBoost against a persistence baseline (predict the previous lap's time, within driver/race) and linear regression. For pit-stop classification we compare XGBoost against a majority-class baseline and logistic regression with balanced class weights. XGBoost hyperparameters follow the original paper's specification without re-tuning, since the purpose of this pass is evaluation correction, not hyperparameter search: for regression, 300 estimators, max depth 6, learning rate 0.05, subsample 0.85, column subsample 0.9, $\ell_2 = 2.0$, $\ell_1 = 1.0$; for classification, 250 estimators, max depth 5, learning rate 0.08, subsample 0.9, column subsample 0.85, with `scale_pos_weight` set from the training fold's class ratio.

### 3.5 Explainability

We compute SHAP values (`TreeExplainer`) for both final (track-aware) models on a random sample of up to 1,500 rows, producing global feature-importance summaries and individual-prediction waterfall plots for illustrative cases.

### 3.6 Cross-season generalization test

Both track-aware models, trained once on the full 2023 season, are evaluated without retraining on the 6-race 2024 holdout described above (6,133 laps after cleaning), to test whether performance transfers across a season boundary.

## 4 Results

### 4.1 Leakage decomposition

Table 1 confirms that the original evaluation's near-perfect $R^2$ is reproducible almost exactly once both the row-wise split and identity features are restored (Scenario A), and that identity features alone provide no benefit once the split is honest and the circuit is genuinely unseen (Scenario B, comparable to Scenario C). Scenario D shows that keeping circuit identity is not illegitimate in general (it is only a problem when combined with a leaky split or evaluated against unseen circuits), and yields a strong, honestly-validated result.

Table 1: Leakage decomposition, lap-time regression (Section 3.3).

| Scenario | Split | Driver/GP features | $R^2$ |
|---|---|---|---|
| A | Random row-wise | Both | 0.994 |
| B | Grouped by race | Both (unseen at test) | $-0.58 \pm 0.47$ |
| C (track-agnostic, primary) | Grouped by race | None | $-2.26 \pm 1.56$ |
| D (track-aware, primary) | Grouped by driver | GP only | $0.991 \pm 0.002$ |

### 4.2 Lap time regression

Table 2 reports the corrected metrics alongside the original notebook and paper figures, neither of which we were able to reproduce from the available pipeline. The track-agnostic model performs *worse than predicting the training-set mean* on unseen circuits ($R^2 < 0$), because absolute lap time is dominated by track length and layout, which no feature in this study represents. The track-aware model is a real improvement over a strong persistence baseline (MAE 0.54s vs. 0.61s), but the gap is modest, not the near-perfect result the original evaluation implied.

Table 2: Lap time regression: old vs. corrected metrics.

| Model | MAE (s) | RMSE (s) | $R^2$ |
|---|---|---|---|
| Notebook (as found, not reproducible) | N/A | 0.70 | 0.9973 |
| Paper (as claimed, not reproducible) | 0.174 | 0.769 | 0.995 |
| Track-agnostic: XGBoost | $13.40 \pm 2.14$ | $16.24 \pm 1.78$ | $-2.26 \pm 1.56$ |
| Track-agnostic: linear regression | $8.28 \pm 1.89$ | $10.63 \pm 2.53$ | $-0.29 \pm 0.57$ |
| Track-agnostic: persistence | $0.61 \pm 0.15$ | $2.20 \pm 0.95$ | $0.943 \pm 0.038$ |
| **Track-aware: XGBoost** | $\mathbf{0.54 \pm 0.01}$ | $\mathbf{1.03 \pm 0.12}$ | $\mathbf{0.991 \pm 0.002}$ |
| Track-aware: persistence | $0.61 \pm 0.05$ | $2.41 \pm 0.27$ | $0.949 \pm 0.012$ |

### 4.3 Pit-stop classification

Table 3 shows the same pattern: the track-agnostic classifier barely beats a majority-class baseline on F1 and is actually *worse* than logistic regression on ROC-AUC (0.645 vs. 0.669), while the track-aware classifier is a clear improvement over both non-trivial baselines. As discussed in Section 5, even the track-aware result trails recent published comparators on this task.

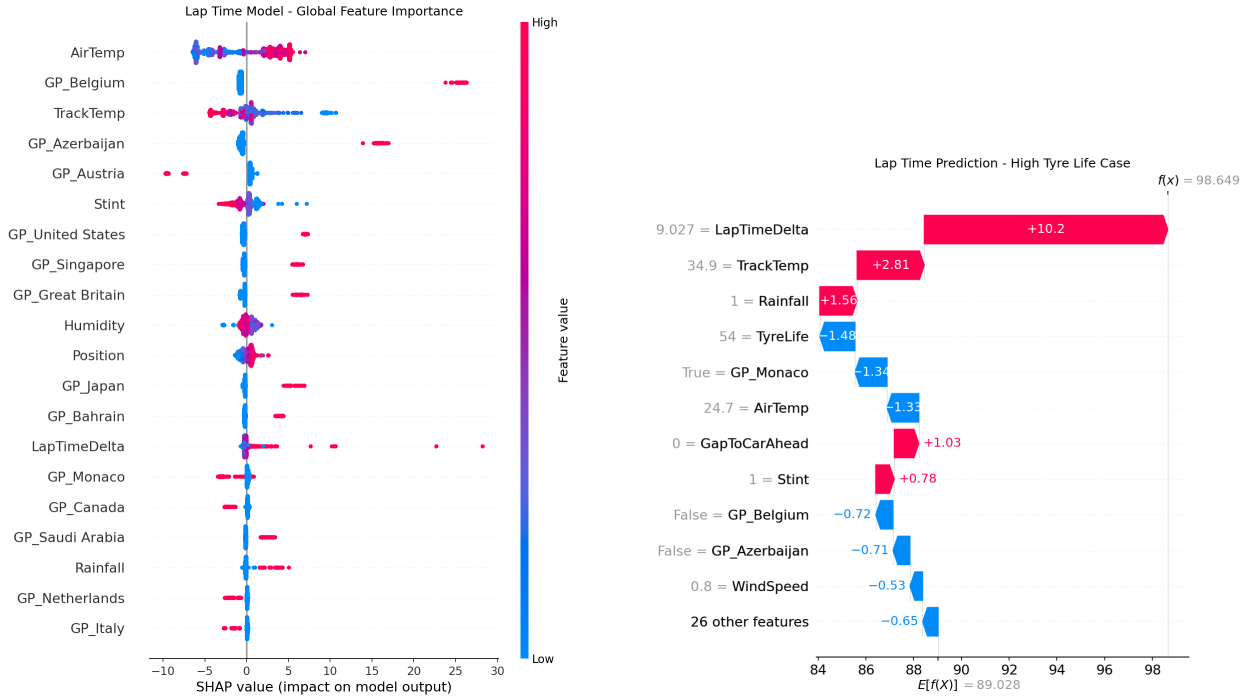
Table 3: Pit-stop-next-lap classification: old vs. corrected metrics.

| Model | Accuracy | Precision | Recall | F1 | ROC-AUC |
|---|---|---|---|---|---|
| Notebook (as found, not reproducible) | 98.62% | N/A | N/A | 0.72 | N/A |
| Paper (as claimed, not reproducible) | 89.53% | 0.71 | 0.81 | N/A | 0.94 |
| Track-agnostic: XGBoost | $87.8\% \pm 5.7$ | $0.087 \pm 0.031$ | $0.213 \pm 0.082$ | $0.114 \pm 0.043$ | $0.645 \pm 0.060$ |
| Track-agnostic: majority class | $96.6\% \pm 0.4$ | 0.0 | 0.0 | 0.0 | 0.500 |
| Track-agnostic: logistic regr. | $65.7\% \pm 6.0$ | $0.057 \pm 0.009$ | $0.565 \pm 0.059$ | $0.103 \pm 0.015$ | $0.669 \pm 0.041$ |
| **Track-aware: XGBoost** | $\mathbf{89.7\% \pm 1.5}$ | $\mathbf{0.200 \pm 0.027}$ | $\mathbf{0.643 \pm 0.039}$ | $\mathbf{0.303 \pm 0.030}$ | $\mathbf{0.888 \pm 0.005}$ |

Precision/recall/F1 in Tables 3 and 4 refer to the pit-stop (positive) class throughout.

### 4.4 Explainability

Figure 1 shows that the track-aware lap-time model's global feature importance is dominated by circuit identity (*GP*) and weather, with tyre-condition features ranking lower than the original paper's framing would suggest, consistent with the leakage-decomposition finding that absolute lap time is set primarily by track length and layout. Figure 2 shows the opposite, more domain-sensible pattern for the pit-stop model: tyre life and stint dominate, with circuit identity contributing comparatively little, matching the intuition that pit-stop timing is driven by tyre condition rather than which track is being raced.




(a) Global feature importance

(b) Individual case: high tyre life, wet lap

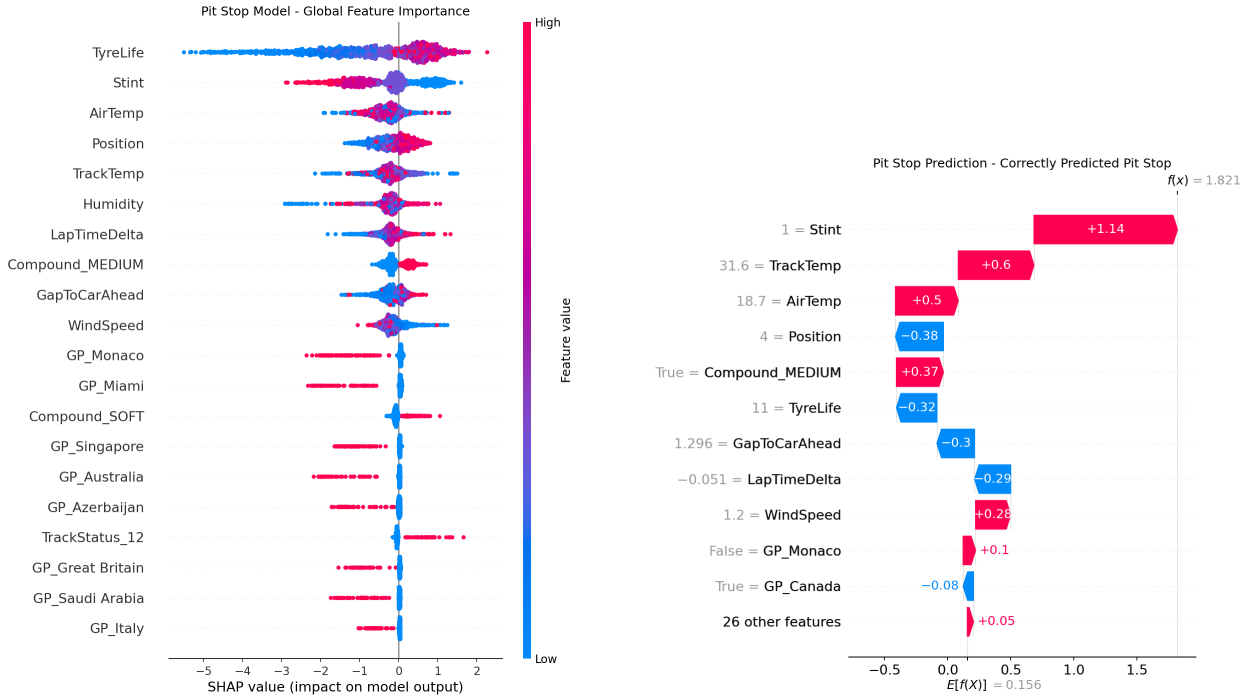
Figure 1: SHAP analysis, track-aware lap-time model.

### 4.5 Cross-season generalization

Table 4 reports both track-aware models, trained only on 2023 data, evaluated without retraining on the 2024 holdout. Both degrade substantially. The lap-time model's larger relative drop is consistent with its dependence on circuit identity as a pace-baseline proxy, which goes stale as cars are developed season over season; the pit-stop model's smaller relative drop is consistent with its dependence on tyre-life dynamics, which are more stable year over year, though still affected by Pirelli's per-round compound allocation changing independently of season.

## 5 Limitations

**The pit-stop classifier trails the current published field, and we do not soften this.** Our track-aware classifier reaches F1= 0.303 and ROC-AUC= 0.888. Sasikumar et al.'s Bi-LSTM [5] reaches F1= 0.81 and ROC-AUC= 0.988 on a temporally-evaluated holdout; Fatima and Johrendt's




(a) Global feature importance

(b) Individual case: correctly predicted pit stop

Figure 2: SHAP analysis, track-aware pit-stop model.

Table 4: Cross-season generalization: 2023-trained, evaluated on 2024 without retraining.

| Model | Metric | 2023 (in-season CV) | 2024 (held out) |
|---|---|---|---|
| Lap time | MAE (s) | 0.54 | 7.85 |
| Lap time | RMSE (s) | 1.03 | 10.01 |
| Lap time | $R^2$ | 0.991 | 0.172 |
| Pit stop | Accuracy | 89.7% | 87.7% |
| Pit stop | Precision | 0.200 | 0.080 |
| Pit stop | Recall | 0.643 | 0.329 |
| Pit stop | F1 | 0.303 | 0.129 |
| Pit stop | ROC-AUC | 0.888 | 0.721 |

Deep-Racing [2] reports F1= 0.67; García Tejada's SVM [3] reports F1= 0.621. All three are ahead of our result. We do not claim our evaluation methodology explains this gap: Sasikumar et al.'s temporal holdout is itself a leakage-conscious design, not a naive split, and a sequence model may genuinely capture pit-timing dynamics that a single-lap tabular classifier cannot. This paper's contribution is the leakage audit and corrected evaluation protocol applied to a tabular gradient-boosted-tree pipeline, not a claim that this pipeline is competitive with the current best published pit-stop predictors.

**The track-aware framing has a specific, narrower scope than "lap time prediction" as originally framed.** It generalizes to an unseen driver on a circuit already represented in training, not to an unseen circuit (Table 1, Scenario C). This is a reasonable deployment framing (a race team always knows its own circuit), but it does mean the model has not been shown to isolate a track-independent tyre-degradation signal, which the original task framing implicitly claimed.

**Cross-season generalization is weak without retraining** (Table 4). Neither model should be deployed across a season boundary without at least recalibration against current-season data.

**Safety cars, red flags, and team orders are not modeled.** Pit-stop timing is frequently driven by external strategic triggers not represented in the feature set; *TrackStatus* is a coarse proxy at best.

**Driver and constructor identity are deliberately excluded** from the pit-stop model and from the lap-time model's driver dimension, specifically to avoid the identity-memorization problem this paper documents. This means neither model represents genuine, non-random variation in team strategy philosophy or individual driver tyre management.

**Training data is single-season.** The 2024 data introduced in this paper is used only for evaluation, not training.

**The lap-time delta feature is quasi-derived from the regression target.** It is defined as the current lap's time minus the previous lap's time for the same driver, so it is linearly related to the target it helps predict. It is retained because the original paper's feature specification includes it; an ablation removing it changed the track-agnostic result by less than 0.03 $R^2$, indicating it is not the dominant confound relative to track identity, but its role should not be left implicit.

## 6 Conclusion

We audited an existing Formula 1 lap-time and pit-stop prediction pipeline and found that its previously reported near-perfect performance was substantially an artifact of a row-wise random train/test split combined with one-hot-encoded driver and circuit identity, not genuine modeling of race conditions. A grouped, leakage-audited evaluation protocol collapses the lap-time model's performance on genuinely unseen circuits below a mean-prediction baseline, revealing that absolute lap time in this feature set is dominated by track identity rather than tyre degradation. A track-aware reformulation (holding out drivers rather than circuits, and treating circuit identity as legitimate known context) restores strong, honestly-validated performance for lap time, and a more modest but still non-trivial improvement over baselines for pit-stop classification, which nonetheless trails the current published state of the art on the same task. A cross-season generalization test further shows that neither model transfers cleanly to a new season without retraining. We release the corrected pipeline, the leakage decomposition, and all metrics reported here to support more reproducible evaluation practice in this application area.

## References


[1] Fleur Brusik. Predicting lap times in a Formula 1 race using deep learning algorithms: A comparison of univariate and multivariate time series models. Master's thesis, Tilburg University, 2024.

[2] Syeda Sitara Wishal Fatima and Jennifer Johrendt. Deep-racing: An embedded deep neural network (EDNN) model to predict the winning strategy in Formula One racing. *International Journal of Machine Learning*, 13(3):97–103, 2023.

[3] L. García Tejada. Applying machine learning to forecast Formula 1 race outcomes. Master's thesis, Aalto University, 2023.

[4] Theo Sander. FastF1: Python api for accessing Formula 1 telemetry and timing data. GitHub repository, version 3.3, 2023.

[5] Abhijai Sasikumar, A. Anny Leema, and P. Balakrishnan. Data-driven pit stop decision support for Formula 1 using deep learning models. *Frontiers in Artificial Intelligence*, 8:1673148, 2025.

[6] Devin Thomas, Junqi Jiang, Avinash Kori, Aaron Russo, Steffen Winkler, Stuart Sale, Joseph McMillan, Francesco Belardinelli, and Antonio Rago. Explainable reinforcement learning for Formula One race strategy, 2025.

[7] Jamie Todd, Junqi Jiang, Aaron Russo, Steffen Winkler, Stuart Sale, Joseph McMillan, and Antonio Rago. Explainable time series prediction of tyre energy in Formula One race strategy, 2025.

[8] Zhixuan Zhao. Deep neural network-based lap time forecasting of Formula 1 racing. *Applied and Computational Engineering*, 47:61–66, 2024.