Scilight

Intelligence & Control
https://www.sciltp.com/journals/ic

Article

Check for updates

Behavior-Induced False Positives in Vehicle Telemetry Anomaly Detection: An Empirical Study

Tudor Hirtopanu, Zidong Wang, Alan Serrano and Weibo Liu *

Department of Computer Science, Brunel University of London, Uxbridge, Middlesex UB8 3PH, UK
* Correspondence: Weibo.Liu2@brunel.ac.uk

How To Cite: Hirtopanu, T.; Wang, Z.; Serrano, A.; et al. Behavior-Induced False Positives in Vehicle Telemetry Anomaly Detection: An Empirical Study. *Intelligence & Control* **2026**, *2*(2), 3. https://doi.org/10.53941/ic.2026.100006

Received: 24 March 2026
Revised: 19 April 2026
Accepted: 29 May 2026
Published: 26 June 2026

Abstract: Unsupervised anomaly detection is a promising approach for vehicle health monitoring, but its deployment in human-driven telemetry is often limited by nominal false positives. A key difficulty is that many telemetry channels are influenced not only by vehicle dynamics, but also by partially observed driver behavior and operating context, making some signals systematically less predictable than others. This paper presents a systematic empirical analysis of how such variability affects false-positive behavior in vehicle telemetry anomaly detection. Using multiple public vehicle telemetry datasets and representative detector families, the feature-level structure of nominal residual error and its contribution to false-positive detections are analysed. Experimental results show that residual error is not distributed uniformly across the telemetry space but is instead concentrated in a small subset of driver-controlled channels. Residual-error concentration persists across model classes, becoming more pronounced in falsely flagged nominal windows, and is not eliminated by increasing training-driver diversity. Feature-level analysis further shows that steering-, braking-, and accelerator-related variables dominate the residual hierarchy, whereas more tightly regulated vehicle-state channels remain comparatively stable. The findings identify a structured and recurrent false-positive failure mode in human-driven telemetry. It is further suggested that reliable deployment will require channel-aware or context-conditioned scoring strategies, rather than increased model complexity alone.

Keywords: anomaly detection; vehicle telemetry; false positives; driver behavior; unsupervised learning; multivariate time series

1. Introduction

Modern vehicles are complex cyber-physical systems in which mechanical components, embedded electronics, and software-driven control loops interact under diverse and continuously changing operating conditions [1,2]. The increasing integration of mechanical, electronic, and software components improves functionality and performance, but also complicates system monitoring and fault diagnosis [1,3]. Because many fault events are rare, heterogeneous, and difficult to label, unsupervised anomaly detection has become an attractive approach for vehicle health monitoring [4,5].

Deep sequence models are increasingly used to learn nominal patterns in multivariate vehicle telemetry, offering a scalable alternative to rule-based and model-based diagnostic methods [1,3]. By modeling temporal dependencies directly from historical sensor and control data, the deep sequence methods can capture rich correlations that are difficult to specify manually [4,5]. Deep sequence models may only generalize to the extent that nominal telemetry reflects learnable and stable system dynamics [4], an assumption that is often violated in human-driven settings [6]. Consequently, anomaly detectors may produce false positives even under nominal operation [4,7]. High false-positive behavior in human-driven telemetry is a deployment-critical problem, because the practical value of a health-monitoring system depends not only on detecting genuine faults, but also on maintaining stable and interpretable alarm behavior under nominal driving variability. Elevated false-positive rates are particularly problematic in vehicle health monitoring because they reduce trust in automated alerts, increase unnecessary diagnostic effort, and complicate threshold calibration across drivers and operating conditions [8,9].

The predictability of human-driven vehicle telemetry is fundamentally limited by partial observability. A substantial portion of signal variability is governed not only by deterministic physical laws, but also by unobserved factors such as driver intent, traffic conditions, and road geometry [10–13]. As a result, behavior-sensitive channels can exhibit heteroscedastic and context-dependent variability that is unrelated to vehicle degradation [12, 14, 15]. Sequence-based anomaly detectors generally assess abnormality through prediction or reconstruction error; consequently, behavior-sensitive channels with high contextual variability can yield elevated anomaly scores even under nominal operation [4,8]. As a result, benign behavioral variation may be mistaken for evidence of system abnormality [16,17]. Nominal error in human-driven telemetry therefore reflects a structured limitation of the monitoring setting itself, rather than merely insufficient model capacity.

False positives are widely recognized as a practical limitation of anomaly detection, but their feature-level causes in human-driven vehicle telemetry remain insufficiently understood [18,19]. The limited understanding of feature-level false-positive causes is particularly important because deployment reliability depends not only on detecting genuine faults, but also on avoiding persistent false alarms under nominal behavior [4,9]. When benign driver-induced variability repeatedly produces elevated anomaly scores, alerts become less interpretable, thresholds transfer less reliably across drivers and conditions, and genuine fault signatures become harder to distinguish from ordinary behavioral noise [11,20].

It remains unclear whether nominal false positives are broadly distributed across vehicle signals or concentrated in a smaller subset of behavior-sensitive channels. Existing research has focused primarily on improving model architectures and aggregate detection performance [4,5,21], with less attention to the structure of nominal prediction error across telemetry features [17,22]. It is further unclear whether this concentration in behavior-sensitive channels persists across datasets, model classes, and evaluation protocols.

Driver variability has been extensively studied in related contexts such as driver identification and behavior modeling. By contrast, its role as a source of false positives in anomaly detection has not been systematically analyzed. Understanding how driver variability contributes to false positives is particularly important in practical settings, where deployment reliability depends not only on detecting true faults but also on understanding and mitigating systematic sources of false alarms.

In this paper, we present a systematic empirical analysis of nominal false positives in human-driven vehicle telemetry anomaly detection. Unlike prior work, which has largely emphasized overall detection performance or the development of new detector architectures, this study focuses on the structure of false-positive behavior itself. Across multiple public datasets and representative anomaly detection baselines, we show that nominal residual error is not distributed uniformly across telemetry channels, but is instead disproportionately concentrated in a small subset of driver-controlled features. Residual error remains concentrated in driver-controlled features across detector families, datasets, and evaluation protocols, indicating a structural property of human-driven telemetry rather than a model-specific artifact. The findings identify a structured and practically important failure mode in vehicle telemetry anomaly detection, and suggest that reliable deployment in human-driven systems will require more than improved model capacity alone.

The specific contributions of this work are summarized as follows:

- We present a systematic empirical analysis of nominal false-positive behavior in human-driven vehicle telemetry.
- We show that nominal residual error is not uniformly distributed, but is concentrated in a small subset of driver-controlled (behavior-sensitive) features.
- We demonstrate that residual error remains consistently concentrated in behavior-sensitive features across multiple detector families, datasets, and evaluation protocols.
- We find that increasing training-driver diversity does not reduce residual error in behavior-sensitive features, highlighting a limitation of standard generalization strategies.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 formulates the problem and defines the feature partition, followed by Section 4, which presents the research hypotheses. The experimental methodology is described in Section 5, including datasets, models, evaluation protocols, and metrics. Empirical results are reported in Section 6, with their implications, limitations, and future directions discussed in Section 7. Finally, Section 8 concludes the paper.

2. Related Work

2.1. Deep Anomaly Detection for Vehicle Telemetry

Unsupervised deep anomaly detection has become a prominent approach in vehicle health monitoring, leveraging sequence models to learn temporal and cross-variable structure directly from nominal telemetry [4,8,23]. Unlike

traditional model-based diagnostics, data-driven frameworks—including reconstruction-based, forecasting-based, and probabilistic methods—do not require explicit fault hypotheses [3,4,24]. In contrast, data-driven frameworks define abnormality through deviation from learned patterns of nominal behavior [4,7].

A limitation of many anomaly detection frameworks is the lack of explicit accounting for large differences in predictability across telemetry channels [17,22,25]. In particular, existing methods are typically not designed to distinguish between exogenous driver control inputs and endogenous vehicle system responses when constructing anomaly scores [26,27]. Signals shaped by stochastic and partially observed human behavior are evaluated using the same anomaly-scoring framework as signals governed by deterministic physical dynamics [7,12].

From a diagnostic perspective, the distinction between exogenous driver inputs and endogenous vehicle responses is important. Behavior-sensitive channels can exhibit irreducible and context-dependent variability even under healthy operation. When anomaly scores are derived from prediction or reconstruction error, high-variability channels can disproportionately influence aggregate detection scores, increasing false positives under unseen drivers or operating contexts.

2.2. Driver-Induced Variability as a Source of Distributional Shift

A defining characteristic of human-driven vehicle telemetry is that many signals are shaped not only by vehicle dynamics, but also by variability in driver behavior [6, 11]. Driving style differs substantially across individuals [11,28] and can also vary within the same driver depending on context, including traffic conditions, road geometry, intent, fatigue, and environmental demands [10,13,29]. Nominal telemetry recorded under human operation is not generated by a single stationary process, but by a mixture of behaviorally and contextually dependent regimes [12,14,30].

Driver-induced variability is particularly pronounced in channels associated with steering, braking, acceleration, and their immediate physical consequences [10,11,31]. The combination of driver commands, controller responses, and vehicle dynamics makes steering-, braking-, and acceleration-related variables more sensitive to latent contextual factors than mechanically constrained system signals [14]. Since many factors–such as latent driver intent, perception errors, and memory effects–are only partially observed in onboard telemetry, their influence may appear as unpredictable but nominal variation [12,14].

From the perspective of anomaly detection, behavioral variability acts as a nuisance source of distributional shift. Nominal human behavior continuously evolves and exhibits strong temporal dependence [8,16], causing telemetry patterns that are nominal for one driver or driving context to deviate substantially from those observed during training, even when no fault is present [17,22,25]. Behavioral distributional shift is especially problematic under driver-disjoint or deployment-like evaluation, where detectors must generalize beyond the behavioral regimes represented in the training set [11]. Experimental results suggest that false positives in human-driven telemetry may arise not from uniformly poor modeling, but from structured shifts concentrated in behavior-sensitive channels.

2.3. Limitations of Contextual and Physics-Informed Diagnostics

Recent work has explored context-aware and physics-informed approaches to anomaly detection in vehicle systems, aiming to reduce spurious alarms by conditioning decisions on operating state, vehicle dynamics, or physical consistency constraints [27,32,33]. Context-aware and physics-informed methods reflect an important insight: abnormality should not be interpreted solely through raw deviation from nominal signal patterns, but relative to the conditions under which the vehicle is operating [17,22,27]. Incorporating observable context or known physical structure can improve robustness to changes in speed, maneuvering regime, or environmental conditions [32,33].

The aforementioned methods typically rely on the assumption that relevant context is either directly observable or can be sufficiently encoded through available telemetry and auxiliary variables. In human-driven vehicle systems, important sources of variability arise from latent and open-ended factors such as driver intent, behavioral style, anticipation, situational judgment, and interaction with traffic conditions [6, 12]. Latent and open-ended factors are typically only partially observed, and therefore contextual augmentation alone does not fully eliminate behavior-induced variability in telemetry [12].

The limitation of incomplete contextual observability is particularly important for channels associated with steering, braking, acceleration, and their immediate physical consequences [10, 11, 31]. Signals reflecting a mixture of driver commands, controller responses, and vehicle dynamics are sensitive to behavioral and contextual factors that are not exhaustively captured by onboard sensing [12,34]. As a result, even context-aware or physics-informed conditioning may leave substantial nominal variability unresolved in behavior-sensitive channels [12]. This study investigates whether context- and behavior-induced variability is uniformly distributed across telemetry

channels or concentrated in a small subset of high-entropy features, thereby imposing a structural limit to cross-driver predictability.

Contextual and physics-informed diagnostics can mitigate some forms of nuisance variation, but do not fully resolve the false-positive problem in human-in-the-loop monitoring. What remains insufficiently understood is how residual nominal variability is distributed across telemetry channels, and which classes of signals continue to dominate false-positive detections under realistic generalization settings. This motivates the need for a systematic empirical analysis of feature-level error structure in human-driven vehicle telemetry.

3. Problem Formulation

Consider a vehicle telemetry stream represented as a multivariate time series $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, where each observation $\mathbf{x}_t \in \mathbb{R}^d$ comprises d telemetry channels. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ denote the corresponding feature set. Given an observation $\mathbf{x}_t$, an anomaly detection model M, trained on nominal data, produces a model-dependent estimate $\tilde{\mathbf{x}}_t$ of expected nominal behavior. The anomaly score at time t is then defined as

$$S_t = \mathcal{L}(\mathbf{x}_t, \tilde{\mathbf{x}}_t),$$

where $\mathcal{L}(\cdot, \cdot)$ denotes a deviation measure, such as prediction error, reconstruction error, or a related score induced by the nominal model.

In this paper, the feature set $\mathcal{F}$ is divided into two disjoint subsets, $\mathcal{F}_d$ and $\mathcal{F}_v$, such that $\mathcal{F}_d \cup \mathcal{F}_v = \mathcal{F}$ and $\mathcal{F}_d \cap \mathcal{F}_v = \emptyset$.

(1) Driver-controlled features ($\mathcal{F}_d$): Channels directly manipulated by the driver, such as steering, braking, and accelerator inputs.

(2) Vehicle-state features ($\mathcal{F}_v$): The remaining channels, consisting predominantly of vehicle-response variables together with a small number of contextual variables not directly controlled by the driver.

In this work, the driver-controlled subset $\mathcal{F}_d$ is also referred to as the behavior-sensitive subset, as the channels are directly influenced by human driving behavior and its partially observed variability. The partition of the feature set $\mathcal{F}$ is analytical rather than strictly causal. Although driver actions propagate through the vehicle system and influence downstream variables, the subset $\mathcal{F}_d$ captures the primary points at which human control enters the observed telemetry. Our central hypothesis is that, in human-driven systems, the aggregate anomaly score S_t is disproportionately influenced by variability in $\mathcal{F}_d$ relative to $\mathcal{F}_v$. The stronger dependence of $\mathcal{F}_d$ on latent and only partially observed factors, such as driver intent and traffic context, can induce elevated nominal variability even in the absence of faults.

4. Research Hypotheses

To systematically investigate the structural failure modes of unsupervised anomaly detection in human-in-the-loop telemetry, we propose and test three primary hypotheses:

4.1. H1: The Error Concentration Hypothesis

Under nominal driving conditions, prediction and reconstruction errors are expected to be unevenly distributed across the feature space. We hypothesize that the mean residual associated with driver-controlled features exceeds that of vehicle-state features:

$$\mathbb{E}[\mathcal{L}(\mathbf{x}_{i \in \mathcal{F}_d})] > \mathbb{E}[\mathcal{L}(\mathbf{x}_{j \in \mathcal{F}_v})] \tag{1}$$

If H1 holds, aggregate anomaly scores will be disproportionately influenced by a small subset of driver-controlled channels, increasing the likelihood of nominal threshold crossings.

4.2. H2: Cross-Model Consistency Hypothesis

We hypothesize that the concentration of residual error in $\mathcal{F}_d$ is not specific to any single detector architecture, but persists across multiple model families. If H2 holds, the higher residual error in $\mathcal{F}_d$ relative to $\mathcal{F}_v$ will remain observable across linear, recurrent, and attention-based anomaly detection models.

4.3. H3: Limited Reduction Through Training Diversity Hypothesis

Increasing training diversity, such as enlarging the set of seen drivers, is expected to reduce overall false-positive rates but not eliminate the residual disparity between $\mathcal{F}_d$ and $\mathcal{F}_v$. If H3 holds, it would suggest that the

dominant variability in $\mathcal{F}_d$ arises from partially observed behavioral and contextual factors that are not fully resolved through additional nominal training diversity alone.

5. Experimental Methodology

To evaluate the hypotheses under diverse driving conditions and across multiple detector families, we adopt a multi-dataset, multi-model experimental framework. The design is intended to assess whether observed residual-error patterns reflect stable properties of vehicle telemetry rather than artifacts of a particular dataset, route, or model class.

5.1. Datasets and Split Protocols

We use three public CAN/OBD-II vehicle telemetry datasets that capture complementary sources of driver-dependent and contextual variability:

- HCRL Dataset [35]: Telemetry from ten drivers operating the same production vehicle over a fixed 46 km route. The HCRL dataset is used to examine inter-driver generalization while controlling for route geometry.
- Sonata Dataset [36]: Telemetry from four drivers collected over approximately 30 trips in heterogeneous urban environments. The Sonata dataset captures variation arising from driver behavior, route diversity, and traffic context.
- OBD Dataset [37]: A single-driver, multi-trip dataset used to examine intra-driver variability across differing road and traffic conditions.

For the multi-driver datasets, we employ a driver-disjoint evaluation protocol. Models are trained and validated using data from a subset of seen drivers, and are then evaluated on held-out unseen drivers excluded entirely from training. The driver-disjoint evaluation setting provides a direct test of cross-driver generalization and is particularly relevant to H3, which concerns the extent to which increased nominal training diversity reduces residual disparity in driver-controlled features.

All datasets are treated as nominal-only in this study, and no labeled fault events are used in training, validation, or test evaluation. Accordingly, any test window whose anomaly score exceeds the decision threshold is counted as a false positive.

5.2. Data Preprocessing and Window Construction

Raw telemetry files are ordered by their available time index, and non-informative metadata fields, such as timestamp, class-label, and path-order columns where present, are removed. All retained channels are treated as numeric telemetry variables and channels with near-zero variance in the training split are excluded.

To avoid information leakage, normalization is performed using statistics computed from the training split only. Each channel is standardized using the training-split mean and standard deviation, with the same transformation subsequently applied to the validation and test splits.

After preprocessing, each file is segmented into fixed-length sliding windows for model input. Windowing is performed independently within each file using a window length of $L = 30$ and a stride of 1, so that no window spans file boundaries. The sliding-window procedure produces standardized multivariate sequences while preserving within-file temporal structure.

5.3. Feature Categorization

Following the partition defined in Section 3, the retained telemetry channels in each dataset are assigned to either the driver-controlled subset $\mathcal{F}_d$ or the vehicle-state subset $\mathcal{F}_v$. Assignment is performed separately for each dataset using a strict actuation-based criterion: channels directly manipulated by the driver are assigned to $\mathcal{F}_d$, while all remaining retained channels are assigned to $\mathcal{F}_v$. The feature partition is intentionally conservative. Driver behavior may influence many downstream variables, only channels representing direct driver input are included in $\mathcal{F}_d$.

- Driver-controlled features ($\mathcal{F}_d$): Channels directly actuated by the driver, including variables associated with steering, braking, acceleration, and clutch input. Representative examples include Steering Wheel Angle, Steering Wheel Speed, Accelerator Pedal Value, Brake Switch Status, and Clutch Operation.
- Vehicle-state features ($\mathcal{F}_v$): All remaining retained telemetry channels, consisting predominantly of vehicle-state and vehicle-response variables, together with a small number of contextual channels not directly controlled by the driver. Representative examples include Engine Speed, Wheel Velocities, Fuel Pressure, Intake Air Pressure, and Throttle Angle.

5.4. Model Architectures

To evaluate H2, we consider a diverse set of anomaly detectors spanning linear, reconstruction-based, forecasting-based, and generative sequence-modeling paradigms. The evaluated models include PCA [38] as a linear baseline; a GRU-autoencoder [39] and an LSTM-autoencoder [40] as deterministic reconstruction models; an LSTM-forecaster [40] and a Transformer-forecaster [41] for one-step-ahead multivariate prediction; and probabilistic or generative approaches including LSTM-VAE [42] and USAD [20].

The selected anomaly detection models were chosen to cover detector families with distinct inductive biases and scoring mechanisms. Accordingly, any residual concentration that persists across the models is less likely to be attributable to a particular architectural choice and more likely to reflect a stable property of the underlying telemetry distribution.

For the forecasting-based models, prediction is restricted to a one-step-ahead horizon to reduce the influence of compounding forecast error and to provide a more controlled basis for comparing feature-wise residual structure. Restricting the horizon is particularly important in the present setting because the telemetry is sampled at 1 Hz, so successive observations may exhibit substantial changes, making longer-horizon forecasting errors accumulate rapidly and potentially confounding the analysis.

5.5. Training and Implementation Protocol

To ensure comparability across detector classes, all models are implemented in PyTorch and trained using the same optimization settings (Adam optimizer, a fixed learning rate of 10^{-3}, and early stopping based on validation performance) on an NVIDIA RTX 3090 GPU. Input data are segmented into sliding windows of length $L = 30$ with stride 1, and all models are trained using nominal data only. Early stopping is used to reduce overfitting and to ensure that comparisons reflect differences in model behavior rather than unequal training duration.

To improve comparability across architectures, anomaly scores are computed as the mean feature-wise error, thereby reducing sensitivity to feature dimensionality and window length. Threshold calibration is performed on validation data, as described in Section 5.

5.6. Dataset-Specific Evaluation Protocols

To evaluate false-positive behavior under both inter-driver and intra-driver variability, each dataset is partitioned according to its collection structure.

HCRL Dataset: The HCRL dataset contains telemetry from 10 drivers operating the same vehicle over a fixed route, enabling controlled evaluation of cross-driver variability. Driver-disjoint training/held-out partitions are constructed with ratios of 3/7, 5/5, and 7/3. For each partition, the training-driver subset is divided into training and validation portions using a 70/30 split, while the held-out-driver subset is reserved entirely for testing. To reduce sensitivity to any particular driver assignment, 10 random split realizations are generated for each ratio, and results are aggregated across realizations.

Sonata Dataset: The Sonata dataset contains telemetry from 4 drivers across heterogeneous urban trips. A leave-one-driver-out protocol is adopted, in which data from 3 drivers are used for model development and the remaining driver is reserved as a held-out test subject. Within the development subset, the data is divided into training and validation portions with a ratio of 70/30. The procedure is repeated four times so that each driver serves once as the held-out test driver.

To further examine the effect of training-driver diversity, an additional protocol is applied in which, for each held-out test driver, separate models are trained using all possible combinations of the remaining 1, 2, and 3 development drivers. The protocol allows controlled analysis of whether increasing driver diversity during training reduces residual error and false-positive concentration on the same held-out individual.

OBD Dataset: The OBD dataset contains 81 telemetry files collected from a single driver across multiple trips. Since inter-driver partitioning is not possible, this dataset is used to evaluate intra-driver variability across differing operating contexts. Files are partitioned into training, validation, and test subsets with target proportions of 60/20/20. The split is constructed to balance both the number of files and the total number of time steps assigned to each subset.

5.7. Evaluation Metrics

As the study is conducted on nominal-only telemetry, evaluation focuses on false-positive behavior rather than label-based discrimination metrics. The primary unit of analysis is the sliding window. For each test window, the detector produces an anomaly score S_t, and a false positive is recorded when $S_t > \tau$. The primary operational

metric is the window-level false positive rate,

$$\mathrm{FPR} = \frac{N_{\mathrm{FP}}}{N_{\mathrm{test}}},$$

where N_{FP} is the number of nominal test windows whose scores exceed the threshold, and N_{test} is the total number of nominal test windows.

Detection thresholds are calibrated separately for each model and split using the nominal validation set. Specifically, τ is defined as the empirical 0.99 quantile of the validation-score distribution and is then applied unchanged to the corresponding test set.

To characterize the structure of false positives, we compute per-feature error terms $e_{t,i}$ for each test window. For reconstruction-based detectors, $e_{t,i}$ denotes the mean squared reconstruction error of feature f_i over the window; for forecasting-based detectors, it denotes the squared prediction error of feature f_i at the forecast target step. The errors are then summarized separately over false-positive and non-false-positive nominal windows to obtain feature-wise mean error profiles.

Feature attribution is further quantified using the within-window contribution share

$$c_{t,i} = \frac{e_{t,i}}{\sum_{j=1}^{d} e_{t,j}},$$

which measures the relative contribution of feature f_i to the total error in window t. Averaging $c_{t,i}$ over false-positive and non-false-positive windows identifies the channels that contribute disproportionately when nominal windows are falsely flagged.

To test the central hypothesis, feature-level statistics are also aggregated over the driver-controlled and vehicle-state subsets introduced in Section 3. For each subset, we compute the mean error and total contribution share over false-positive and non-false-positive windows. In addition, we define the group-wise inflation ratio

$$\rho_g = \frac{\bar{e}_g^{(\mathrm{FP})}}{\bar{e}_g^{(\mathrm{nonFP})}},$$

which quantifies the relative increase in mean error for group g when a nominal window is falsely flagged. At the individual-feature level, false-positive inflation is quantified by

$$\Delta_i = \bar{e}_i^{(\mathrm{FP})} - \bar{e}_i^{(\mathrm{nonFP})},$$

which allows features to be ranked according to their association with nominal false alarms.

All reported metrics are aggregated across experimental splits using the mean and standard deviation.

6. Results

This section is organized around the hypotheses introduced in Section 4. Sections 6.1 and 6.2 evaluate H1 and H2 by examining the extent to which residual error is concentrated in driver-controlled features and persists across different anomaly detection models. Section 6.3 evaluates H3 by testing whether increasing the number of training drivers reduces the disparity between driver-controlled and vehicle-state feature groups. Finally, Section 6.4 provides feature-level attribution to identify the specific telemetry channels that contribute most strongly to the observed false-positive behavior, thereby offering signal-level support for the group level findings reported in Sections 6.1 and 6.2.

6.1. Concentration of Nominal Error in Driver-Controlled Features

We first evaluate H1 and H2 by examining whether nominal residual error is distributed uniformly across the telemetry feature space or concentrated in the subset of features directly controlled by the driver. Across all three datasets, driver-controlled features exhibit substantially higher mean per-feature error than vehicle-state features for the majority of evaluated models. Higher residual error in driver-controlled features is consistently observed across linear, recurrent, attention-based, and generative detectors, indicating that nominal modeling error is not evenly distributed across the telemetry space, but is concentrated in channels most directly associated with driver input.

Figure 1 summarizes the difference in residual error between driver-controlled and vehicle-state features for the HCRL, Sonata, and OBD datasets. In both HCRL and Sonata, the separation between the two feature groups is large and consistent across all evaluated detectors. In OBD, the overall error scale is lower, but the same pattern

remains clearly visible: driver-controlled features continue to exhibit markedly higher residuals than vehicle-state features. Taken together, the results support H1 and indicate that elevated nominal residuals arise primarily in driver-controlled features rather than being distributed uniformly across the full telemetry set.

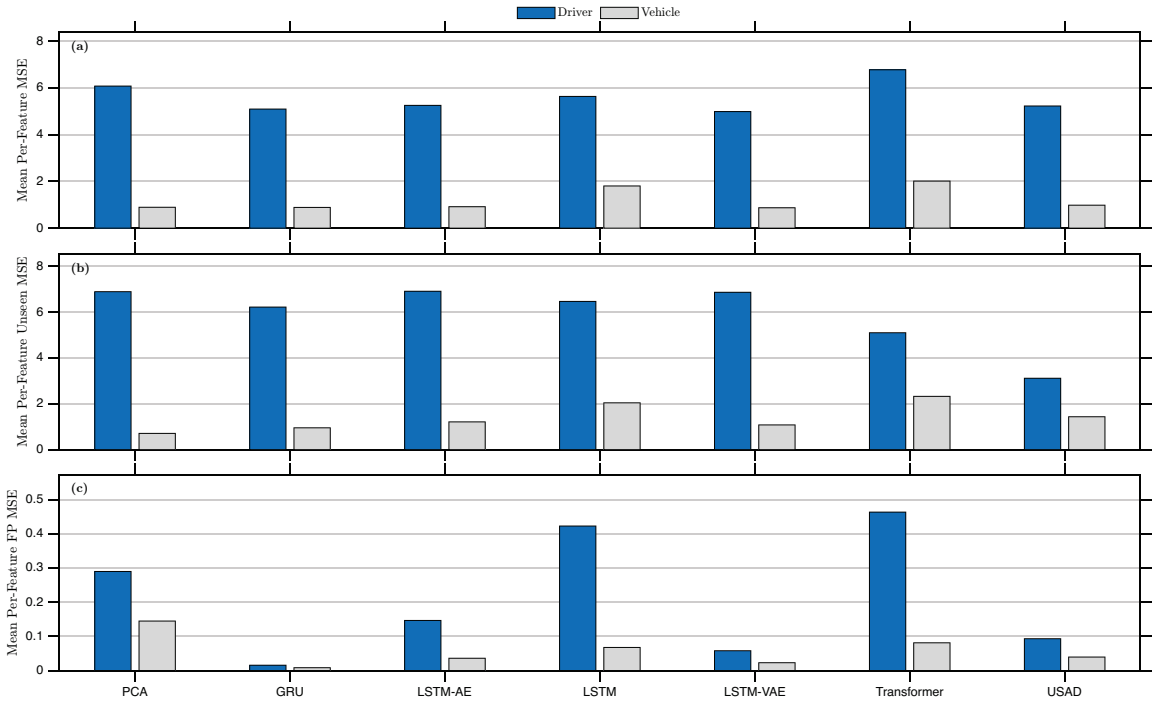

Figure 1. Mean per-feature mean-squared error (MSE) for driver-controlled and vehicle-state feature groups across the (**a**) HCRL, (**b**) Sonata, and (**c**) OBD datasets.

The results in Figure 1 therefore provide empirical evidence that the observed nominal false positives are predominantly associated with behavior-sensitive telemetry channels.

6.2. Amplification of Driver-Controlled Error in False-Positive Windows

To further evaluate H1 and H2, we next examine whether nominal false positives arise from uniform error growth across the telemetry space or from disproportionate amplification in a smaller subset of channels. Specifically, we compare the inflation ratio

$$\rho_g = \frac{\bar{e}_g^{(\mathrm{FP})}}{\bar{e}_g^{(\mathrm{nonFP})}}$$

for the driver-controlled and vehicle-state feature groups. A larger value of ρ_g indicates a stronger increase in residual error for that group when nominal windows are falsely flagged relative to nominal windows that remain below threshold.

Table 1 shows that, across the HCRL and Sonata datasets, the inflation ratio is consistently larger for the driver-controlled group than for the vehicle-state group. In HCRL, the inflation ratio remains higher for driver-controlled features across all evaluated models, with values ranging from 14.15 to 35.89 compared with 3.11 to 13.64 for vehicle-state features. A similarly clear pattern is observed in Sonata, where driver-controlled features again exhibit higher values across all detector families.

Higher error and greater variability in driver-controlled features is also observed in OBD for the majority of models. Although PCA is an exception, and the gap is less uniform than in HCRL and Sonata, most learned detectors continue to exhibit substantially larger inflation in the driver-controlled group. Forecasting models, in particular, exhibit especially large differences in inflation ratios between driver-controlled and vehicle-state features, indicating that false positives in OBD arise from concentrated amplification in driver-controlled channels rather than uniform error growth across the feature space.

The inflation-ratio comparison indicates that nominal false positives are not produced by uniform degradation across the telemetry space, but by a concentrated escalation of residual error in driver-controlled features. Residual error remains consistently concentrated in behavior-sensitive channels across datasets and detector families, indicating that they dominate the error structure of falsely flagged nominal windows.

Table 1. Group-wise inflation ratios ρ_g on held-out test data. Values are reported as mean ± standard deviation across splits.

Dataset	Model	Driver-Controlled	Vehicle-State
HCRL	PCA	15.61 ± 5.95	3.11 ± 0.41
	GRU	21.53 ± 8.31	4.88 ± 1.01
	LSTM	21.89 ± 8.84	4.81 ± 0.97
	LSTM-Forecast	27.84 ± 11.14	12.07 ± 1.79
	LSTM-VAE	21.54 ± 8.59	4.80 ± 0.99
	Transformer-Forecast	35.89 ± 14.10	13.64 ± 2.86
	USAD	14.15 ± 5.20	3.16 ± 0.43
Sonata	PCA	36.29 ± 19.70	4.04 ± 0.69
	GRU	14.29 ± 6.86	2.88 ± 2.21
	LSTM	25.12 ± 13.91	4.04 ± 0.64
	LSTM-Forecast	27.44 ± 10.28	14.11 ± 1.86
	LSTM-VAE	23.60 ± 11.77	4.90 ± 1.23
	Transformer-Forecast	20.86 ± 9.68	11.15 ± 2.85
	USAD	10.83 ± 4.96	2.60 ± 0.44
OBD	PCA	18.79 ± 1.10	20.14 ± 1.18
	GRU	32.37 ± 8.85	24.16 ± 4.50
	LSTM	42.59 ± 10.88	25.43 ± 5.65
	LSTM-Forecast	262.67 ± 16.01	85.16 ± 2.97
	LSTM-VAE	37.47 ± 14.73	17.95 ± 2.48
	Transformer-Forecast	258.47 ± 32.19	86.34 ± 6.88
	USAD	20.04 ± 2.51	18.13 ± 1.85

Note: HCRL results are aggregated over three driver-disjoint splits (3/7, 5/5, 7/3).

In summary, results show that nominal false positives are driven primarily by disproportionate amplification in driver-controlled features rather than by a uniform degradation of model accuracy across all channels. The dominant mechanism underlying nominal false alarms is therefore not merely elevated baseline residual error, but a structured increase concentrated in the channels most directly influenced by driver input. The findings strengthen the evidence for H1 and further support the interpretation that the observed failure mode persists across multiple detector families.

Together with the feature-group disparity reported in Section 6.1, the inflation-ratio results provide further empirical support that the observed nominal false positives are predominantly behavior-related.

6.3. Persistence of Driver-Controlled Error Across Driver-Diversity Regimes

To assess H3, namely whether increasing training-driver diversity reduces the disparity between driver-controlled and vehicle-state channels, we first analyze the HCRL dataset across 10 randomized driver assignments under three development/held-out split configurations (3/7, 5/5, and 7/3). For clarity, Figure 2 reports results for the LSTM model as a representative deterministic sequence detector. The same pattern of higher error and greater variability in driver-controlled features relative to vehicle-state features is observed more broadly across the model families considered in Section 6.

As shown in Figure 2, the driver-controlled group consistently exhibits higher median mean-squared error and substantially greater dispersion than the vehicle-state group across all three split regimes. The vehicle-state channels remain tightly distributed with comparatively low error, whereas the driver-controlled channels retain both elevated central tendency and broader spread. The persistence of the group-level gap indicates that increasing the number of development drivers does not eliminate the residual disparity between the two groups. Although some reduction in variability is visible in selected split realizations, the driver-controlled group remains substantially above the vehicle-state baseline in all three configurations. The persistence of high-error outliers further suggests that threshold exceedances are driven not by diffuse instability across the full telemetry space, but by concentrated variability in a small subset of driver-controlled channels.

A complementary fixed-test-driver experiment on the Sonata dataset yields the same conclusion that increasing training-driver diversity does not eliminate the concentration of residual error in driver-controlled features (Table 2). For each held-out test driver, separate LSTM-Forecaster models were trained using all possible combinations of one, two, and three of the remaining development drivers. As training-driver diversity increased, the overall held-out false-positive rate decreased modestly, but the driver-controlled group remained the dominant contributor

to false-positive-window error. Across the aggregated one-, two-, and three-driver training conditions, held-out driver-controlled false-positive MSE remained extremely large (approximately 60–63), whereas the corresponding vehicle-state false-positive MSE remained below 1.2 and decreased as additional training drivers were included. At the same time, the driver-controlled contribution share in held-out false-positive windows increased from approximately 0.90 to 0.94. Experimental results indicate that increasing driver diversity may reduce overall false positives, but does not remove the dominant concentration of residual error in driver-controlled channels.

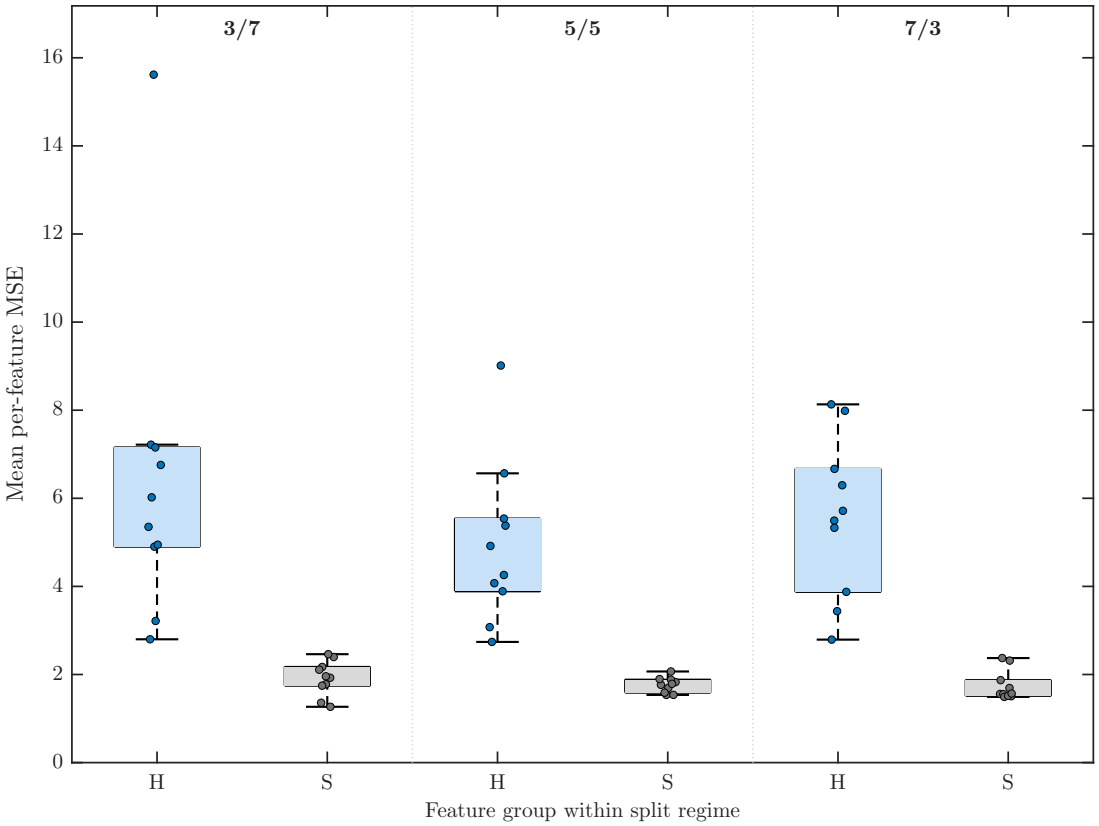

Figure 2. Comparison of mean per-feature MSE for driver-controlled (H, Human; blue) and vehicle-state (S, System; gray) feature groups across three HCRL driver-disjoint split regimes (3/7, 5/5, and 7/3). Each point denotes one randomized split realization. Driver-controlled features exhibit consistently higher median error and broader dispersion across all regimes.

The Sonata fixed-test-driver results indicate that increasing training-driver diversity reduces the overall false-positive rate primarily through improvements in vehicle-state prediction, rather than through any reduction in driver-controlled error. The persistent magnitude of driver-controlled false-positive MSE, together with the increasing contribution share, demonstrates that behavior-sensitive features remain the dominant source of false positives even as nominal training diversity is expanded.

Overall, the HCRL and Sonata experiments indicate that increasing nominal driver coverage does not eliminate the elevated residual error associated with driver-controlled channels. The results support H3 and suggest that the dominant source of error in driver-controlled channels is not fully resolved by additional training-driver diversity alone.

Table 2. Sonata fixed-test-driver LSTM-Forecaster results under increasing training-driver diversity. Values are averaged across held-out test drivers and runs within each training-driver-count condition.

Training Drivers	FPR	Driver FP MSE	Vehicle FP MSE	Driver FP Share
1	0.00165	59.98	1.10	0.899
2	0.00139	60.23	0.82	0.921
3	0.00118	63.02	0.54	0.944

The HCRL and Sonata diversity results further support the interpretation that the observed nominal false positives are not eliminated by increased training diversity and remain primarily associated with behavior-sensitive channels.

6.4. Feature-Level Attribution of Nominal False Positives

To provide signal level support and interpretation for the group-level findings associated with H1 and H2, we examine the feature-wise distribution of residual error in a representative held-out-driver setting. Figure 3 ranks telemetry channels by median mean-squared error for the HCRL LSTM-Forecaster under the 3/7 driver-disjoint split. The resulting hierarchy is strongly non-uniform: the largest residuals are concentrated in a small subset of driver-controlled and behavior-linked channels, rather than being distributed broadly across the telemetry space.

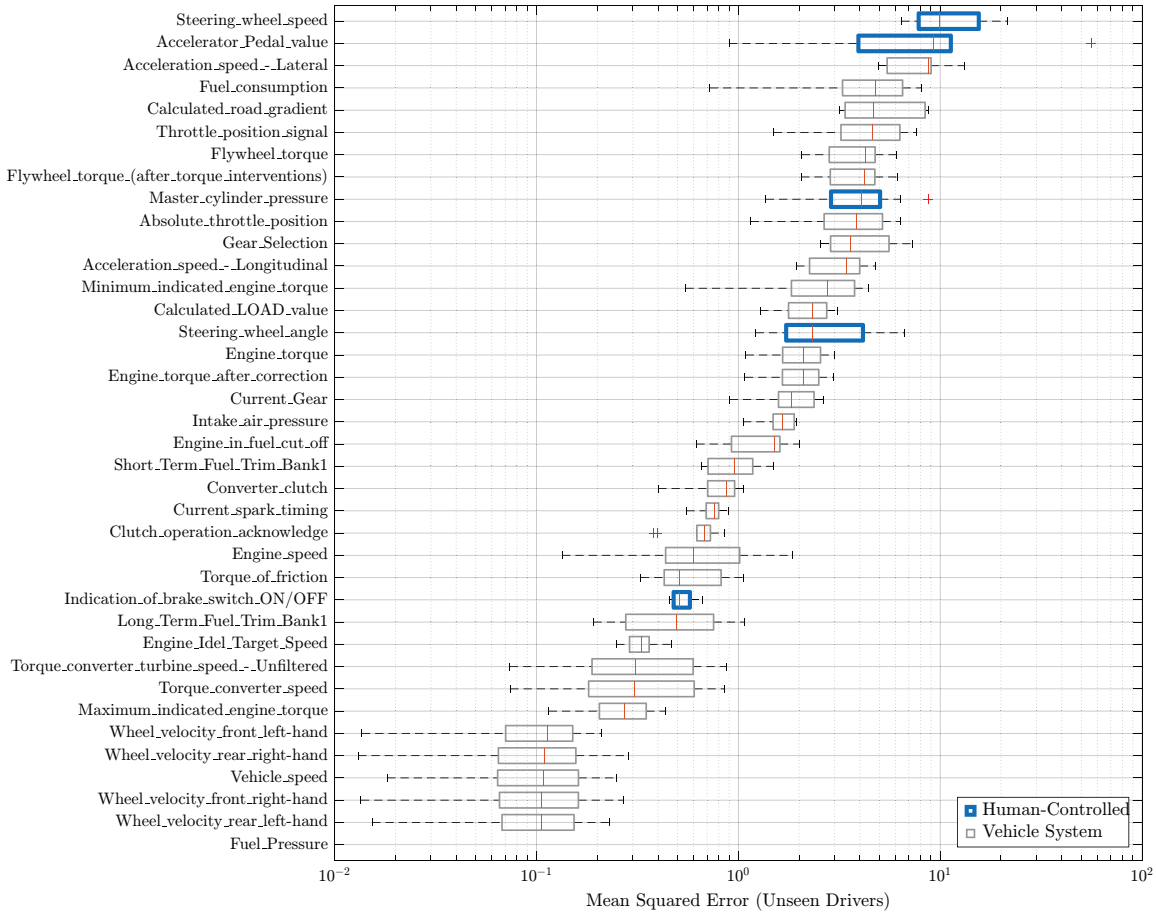

Figure 3. Feature-level distribution of median MSE for the HCRL LSTM-Forecaster under the 3/7 driver-disjoint split, evaluated on held-out drivers. Driver-controlled features (blue) and several context-coupled vehicle-state variables dominate the upper error hierarchy. Fuel Pressure exhibits near-deterministic behavior ($\approx 10^{-11}$ MSE) and falls below the lower bound of the log-scale axis.

The highest-error features include Steering Wheel Speed, Accelerator Pedal Value, and Master Cylinder Pressure, all of which are closely associated with direct driver input or its immediate physical consequences. In contrast, more tightly regulated vehicle-state variables exhibit substantially smaller residual magnitudes. For example, driver-linked channels such as Accelerator Pedal Value and Steering Wheel Speed attain median MSE values on the order of 10^{1}, whereas variables such as Vehicle Speed are closer to 10^{-1}, and highly regulated internal states such as Fuel Pressure approach near-deterministic error levels.

The ranking also indicates that residual variability propagates beyond the direct driver-input interface. Several downstream or context-coupled variables, including Fuel Consumption, Acceleration Speed—Lateral, and Calculated Road Gradient, appear among the higher-error channels despite not belonging to the driver-controlled subset itself. The ranking pattern suggests that stochasticity associated with driver behavior and unobserved operating context can propagate into physically coupled system responses, even though the strongest concentration of error remains in channels most directly influenced by driver input.

Overall, the feature-level analysis confirms that nominal false-positive behavior is driven by a small subset of high-variance channels rather than by diffuse residual growth across the full telemetry set. The feature-level ranking provides a signal-level explanation for why uniformly aggregated residual-based detectors are particularly vulnerable to false alarms under novel but nominal driving behavior.

The feature-level hierarchy is not limited to the representative example shown in Figure 3. Across repeated

runs, the same small subset of driver-controlled channels consistently reappears near the top of the ranking. In HCRL, for example, Master Cylinder Pressure appears in the top 5 features in 78.8% of runs and in the top 10 in 94.6% of runs, while Accelerator Pedal Value has a mean rank of 1.70 and median rank of 1. Similar stability is observed in Sonata for steering-related channels and in OBD, where Accelerator Pedal Position D appears in the top 5 in 99.0% of runs and in the top 10 in 100% of runs. The repeated appearance of the same channels indicates that the feature-level concentration pattern is stable across split realizations rather than an artifact of a single held-out evaluation.

7. Discussion

The empirical results indicate that nominal false positives in vehicle telemetry anomaly detection arise from a structured concentration of residual error rather than from diffuse degradation across the full feature space. Across all three datasets, driver-controlled features exhibit consistently larger residuals than vehicle-state features in nominal windows that are not flagged as anomalies, and the difference becomes substantially more pronounced in falsely flagged nominal windows. The same relationship is observed across linear, reconstruction-based, forecasting-based, recurrent, attention-based, and generative detectors, and persists even as training-driver diversity is increased. The findings suggest that the dominant false-positive mechanism is not architecture-specific, but reflects a recurrent property of human-driven telemetry.

A plausible explanation for the concentration of residual error in behavior-sensitive channels is the partial observability of human driving. Channels associated with steering, braking, and accelerator input are influenced not only by vehicle dynamics, but also by latent driver intent, traffic interactions, route geometry, and other contextual factors that are only incompletely represented in onboard telemetry. Consequently, the channels are systematically more variable in the observed nominal data than tightly regulated internal vehicle states. Residual-based detectors, however, treat departures from learned nominal patterns as evidence of abnormality irrespective of whether that variability arises from genuine system malfunction or from partially observed but still nominal behavior. Accordingly, the principal failure mode is not global modeling breakdown, but persistent over-penalization of the behavior-sensitive portion of the telemetry stream.

Nominal false positives arise primarily from variability in (i) driver-controlled signals such as steering, braking, and acceleration, (ii) their immediate physical consequences within the vehicle system, and (iii) a small number of context-coupled variables influenced by partially observed human driving behavior and operating conditions.

The observed concentration of error in driver-controlled features has several implications for the design of vehicle telemetry anomaly detectors. First, uniformly aggregating residuals across channels can produce unstable anomaly scores, because a small subset of driver-controlled features can dominate the total error. Second, threshold transfer across drivers is likely to remain fragile when the features are scored in the same manner as tightly regulated vehicle-state channels. Third, the results suggest that increasing model complexity alone is unlikely to resolve the problem, since the same disparity persists across multiple detector families and under increased training-driver diversity. More broadly, the present results indicate that the central challenge is not solely representational capacity, but also the mismatch between conventional residual aggregation and the mixed behavioral and physical uncertainty structure of human-driven telemetry. More reliable deployment may therefore require channel-aware scoring, differential treatment of driver-controlled variables, or context-conditioned detection strategies rather than reliance on increasingly expressive sequence models alone.

Several limitations should be acknowledged. First, the partition of features into driver-controlled and vehicle-state subsets is manually defined, although the grouping criterion is intentionally strict and transparent. Second, the datasets considered here remain limited in scale and do not span the full diversity of vehicles, drivers, and operating environments encountered in real deployments. Third, the analysis is restricted to nominal false-positive behavior and therefore does not, by itself, characterize the full trade-off between false-positive reduction and sensitivity to true system faults. Finally, although the observed error concentration in driver-controlled features is consistent with variability induced by partially observed behavioral and contextual factors, the present study does not formally separate aleatoric uncertainty from omitted context, model misspecification, or other sources of residual error.

The aforementioned limitations motivate several directions for future work.

A related direction is the use of multi-step forecasting with weighting to model the temporal evolution of control signals over a longer prediction horizon. Multi-step forecasting with weighting may help capture the dynamics of driver-controlled actions once initiated, but its effectiveness depends strongly on the temporal resolution of the data. In the present setting, where telemetry is sampled at 1 Hz, key control inputs such as steering adjustments and pedal actions occur at sub-second timescales, so successive observations may not capture intermediate transition dynamics. As a result, extending the prediction horizon is likely to introduce compounding forecast error rather

than reduce the observed concentration of false positives. More fundamentally, the primary challenge relates to the unpredictability of the onset of driver-initiated actions, which depends on unobserved contextual and behavioral factors and is therefore not fully addressed by modeling longer-horizon dynamics alone. Exploring multi-step forecasting in higher-frequency telemetry, where transient dynamics are more finely resolved, remains an interesting direction for future work.

The input window length is another relevant design factor. Longer input sequences provide more historical information, but their usefulness depends on how well the data capture the relevant transition dynamics. In the present 1 Hz setting, successive observations may not fully resolve the onset of driver-initiated actions, so extending the window may add temporally distant information that is only weakly informative for predicting current behavior. The limited informativeness of distant history suggests that increasing input length alone may not be sufficient to fully address the concentration of false positives in driver-controlled features.

A natural next step is to develop channel-aware and context-aware scoring methods that reduce the disproportionate influence of driver-controlled channels while preserving sensitivity to genuinely informative fault signatures. Driver-conditioned or behavior-aware anomaly detectors may also improve calibration under cross-driver deployment. At a broader level, the present analysis should be extended to larger and more heterogeneous fleets in order to determine the extent to which the same feature-level concentration pattern persists across vehicle platforms and operating conditions. An equally important question is whether mitigation strategies that reduce nominal false positives can do so without eroding detection sensitivity to true system abnormalities.

8. Conclusions

This paper has presented an empirical study of nominal false-positive behavior in unsupervised vehicle telemetry anomaly detection under human driving. The results show that residual error is not distributed uniformly across the telemetry space, but is instead concentrated in a small subset of driver-controlled channels. Across the datasets considered, the channels consistently exhibit larger nominal residuals than vehicle-state channels and contribute disproportionately to falsely flagged nominal windows. Our findings support H1 and indicate that nominal false positives are driven primarily by structured error concentration rather than diffuse degradation across all features.

The study has shown that the concentration of residual error in driver-controlled features is not specific to any single detector family. Linear, reconstruction-based, forecasting-based, recurrent, attention-based, and generative models exhibit the same pattern of higher residual error in driver-controlled features relative to vehicle-state features, despite differing absolute false-positive rates. The consistency of residual error concentration in driver-controlled features across models provides support for H2 and suggests that the observed failure mode reflects a recurring property of human-driven telemetry rather than an artifact of a particular model architecture.

Additional experiments on HCRL and Sonata have further shown that increasing training-driver diversity does not eliminate the residual disparity between driver-controlled and vehicle-state channels. Although broader driver coverage reduces overall false-positive rates, the available evidence suggests that this improvement is driven primarily by reduced error in vehicle-state variables, while driver-controlled false-positive error remains persistently high and increasingly dominant. The continued dominance of driver-controlled error under increased driver diversity supports H3 and indicates that the principal source of false-positive behavior is not resolved by additional nominal diversity alone.

At the feature level, the dominant contributors to nominal false positives are a small number of steering-, braking-, and accelerator-related channels, together with several closely coupled context-sensitive variables. The feature-level attribution provides a signal-level explanation for why residual-based detectors can behave unstably under nominal human driving: a small subset of high-variance channels can dominate aggregate anomaly scores even when the overall vehicle remains healthy.

Overall, the findings of this paper suggest that reliable deployment of vehicle telemetry anomaly detection in human-driven settings will require more than improved model capacity alone. More robust approaches are likely to require channel-aware scoring, context-conditioned modeling, or alternative aggregation strategies that explicitly account for the distinct statistical role of driver-controlled channels in nominal telemetry.

Author Contributions

T.H.: methodology, writing—original draft, visualization, software; Z.W.: methodology, writing—reviewing and editing, funding acquisition; A.S.: conceptualization, investigation; W.L.: supervision, writing—reviewing and editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the Royal Society of the UK and the Alexander von Humboldt Foundation of Germany.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets used in this study are publicly available from the corresponding original sources. The HCRL dataset is available from [35], the Sonata dataset is available from IEEE DataPort [36], and the Automotive OBD-II dataset is available from RADAR4KIT [37].

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Al-Zeyadi, M.; Andreu-Perez, J.; Hagras, H.; et al. Deep learning towards intelligent vehicle fault diagnosis. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
2. Åström, K.J.; Murray, R.M. *Feedback Systems: An Introduction for Scientists and Engineers*; Princeton University Press: Princeton, NJ, USA, 2008.
3. Abbaspour, A.; Mokhtari, S.; Sargolzaei, A.; et al. A survey on active fault-tolerant control systems. *Electronics* **2020**, *9*, 1513.
4. Zamanzadeh Darban, Z.; Webb, G.I.; Pan, S.; et al. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.* **2024**, *56*, 263.
5. Pang, G.; Shen, C.; Cao, L.; et al. Deep learning for anomaly detection: A survey. *ACM Comput. Surv.* **2021**, *54*, 15.
6. Negash, N.M.; Yang, J. Driver behavior modeling toward autonomous vehicles: Comprehensive review. *IEEE Access* **2023**, *11*, 22788–22821.
7. Tuli, S.; Casale, G.; Jennings, N.R. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow.* **2022**, *15*, 1201–1214.
8. Su, Y.; Zhao, Y.; Niu, C.; et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2828–2837.
9. Lavin, A.; Ahmad, S. Evaluating real-time anomaly detection algorithms: The Numenta Anomaly Benchmark. In Proceedings of the 14th International Conference on Machine Learning and Applications (IEEE ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 38–44.
10. Chu, H.; Zhuang, H.; Wang, W.; et al. A review of driving style recognition methods from short-term and long-term perspectives. *IEEE Trans. Intell. Veh.* **2023**, *8*, 4599–4612.
11. Wei, C.; Qin, Z.; Li, S.; et al. PDB: Not all drivers are the same—A personalized dataset for understanding driving behavior. *arXiv* **2025**, arXiv:2503.06477.
12. Zhang, C.; He, Z.; Wu, C.; et al. When context is not enough: Modeling unexplained variability in car-following behavior. *arXiv* **2026**, arXiv:2507.07012.
13. Mannering, F.L.; Shankar, V.; Bhat, C.R. Unobserved heterogeneity and the statistical analysis of highway accident data. *Anal. Methods Accid. Res.* **2016**, *11*, 1–16.
14. Driggs-Campbell, K.; Shia, V.; Bajcsy, R. Improved driver modeling for human-in-the-loop vehicular control. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1654–1661.
15. Nunes, P.; Santos, J.; Rocha, E. Challenges in predictive maintenance—A review. *CIRP J. Manuf. Sci. Technol.* **2023**, *40*, 53–67.
16. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 15.
17. Song, X.; Wu, M.; Jermaine, C.; et al. Conditional anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 631–645.

18. Wickramasinghe, C.S.; Marino, D.L.; Mavikumbure, H.S.; et al. RX-ADS: Interpretable anomaly detection using adversarial ML for electric vehicle CAN data. *IEEE Open J. Ind. Electron. Soc.* **2022**, *3*, 549–561.
19. Nazat, S.; Abdallah, M. XAI-based feature ensemble for enhanced anomaly detection in autonomous driving systems. *arXiv* **2024**, arXiv:2410.15405.
20. Audibert, J.; Michiardi, P.; Guyard, F.; et al. USAD: UnSupervised anomaly detection on multivariate time series. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 3395–3404.
21. Caetano, F.; Carvalho, P.; Cardoso, J.S. Deep anomaly detection for in-vehicle monitoring—An application-oriented review. *Appl. Sci.* **2022**, *12*, 10011.
22. Li, Z.; van Leeuwen, M. Explainable contextual anomaly detection using quantile regression forests. *Data Min. Knowl. Discov.* **2023**, *37*, 2432–2469.
23. Cherdo, Y.; Miramond, B.; Pégatoquet, A.; et al. Unsupervised anomaly detection for cars CAN sensors time series using small recurrent and convolutional neural networks. *Sensors* **2023**, *23*, 5013.
24. Wang, F.; Jiang, Y.; Zhang, R.; et al. A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors* **2025**, *25*, 190.
25. King, S.; Zhang, Z.; Yu, R.; et al. Contextual learning for anomaly detection in tabular data. *Trans. Mach. Learn. Res.* **2026**. Available online: https://openreview.net/forum?id=PmqZslRENW (accessed on 2 November 2025).
26. Han, X.; Absar, S.; Zhang, L.; et al. Root cause analysis of anomalies in multivariate time series through Granger causal discovery. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025. Available online: Available: https://openreview.net/forum?id=k38Th3x4d9 (accessed on 17 November 2025).
27. Chen, C.Y.; Shin, K.G.; Dadras, S. Context-aware anomaly detection using vehicle dynamics. In Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID), Padua, Italy, 30 September–2 October 2024; pp. 531–545.
28. Yao, X.; Calvert, S.C.; Hoogendoorn, S.P. Driving heterogeneity identification using machine learning: A review and framework for analysis. *Transp. Res. Interdiscip. Perspect.* **2025**, *32*, 101511.
29. Makridis, M.A.; Anesiadou, A.; Mattas, K.; et al. Characterising driver heterogeneity within stochastic traffic simulation. *Transp. B Transp. Dyn.* **2023**, *11*, 725–743.
30. Shi, Y.; Wu, T.; Guo, T.; et al. Traffic simulation optimization considering driving styles. *Commun. Transp. Res.* **2025**, *5*, 100181.
31. Yarlagadda, J.; Pawar, D.S. Heterogeneity in the driver behavior: An exploratory study using real-time driving data. *J. Adv. Transp.* **2022**, *2022*, 4509071.
32. Ghosh, S.; Zaboli, A.; Hong, J.; et al. A physics-informed context-aware approach for anomaly detection in tele-driving operations under false data injection attacks. *IEEE Access* **2025**, *13*, 1.
33. Wasicek, A.R.; Pesé, M.D.; Weimerskirch, A.; et al. Context-aware intrusion detection in automotive control systems. In Proceedings of the 5th ESCAR USA Conference, Ypsilanti, MI, USA, 21–22 June 2017; pp. 1–14. Available online: https://www.weimerskirch.org/files/WasicekEtAl_CAIDS.pdf (accessed on 9 November 2025).
34. Campbell, S.; O'Mahony, N.; Krpalcova, L.; et al. Sensor technology in autonomous vehicles: A review. In Proceedings of the 29th Irish Signals and Systems Conference (ISSC), Belfast, UK, 21–22 June 2018; pp. 1–6.
35. Kwak, B.I.; Woo, J.; Kim, H.K. Know your master: Driver profiling-based anti-theft method. In Proceedings of the 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, New Zealand, 12–14 December 2016; pp. 211–218.
36. Park, K.H.; Kwak, B.I.; Kim, H.K. This car is mine!: Driver pattern dataset extracted from can-bus. 2020. Available online: https://dx.doi.org/10.21227/qar8-sd42 (accessed on 20 October 2025).
37. Weber, M. Automotive OBD-II Dataset. 2023. Available online: https://radar.kit.edu/radar/en/dataset/bCtGxdTklQlfQcAq (accessed on 16 November 2025).
38. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202.
39. Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
40. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. Available online: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 23 September 2025).
42. Park, D.; Hoshi, Y.; Kemp, C.C. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1544–1551.