Scilight

*Journal of Machine Learning and Information Security*
https://www.sciltp.com/journals/jmlis

*Article*

Check for updates

# Uncertainty-Gated Mixture Modeling for Anomaly Detection in Human-in-the-Loop Vehicle Systems

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

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

**How To Cite**: Hirtopanu, T.; Wang, Z.; Serrano, A.; et al. Uncertainty-Gated Mixture Modeling for Anomaly Detection in Human-in-the-Loop Vehicle Systems. *Journal of Machine Learning and Information Securit* **2026**, *2*(2), 10. https://doi.org/10.53941/jmlis.2026.100010

Received: 24 March 2026
Revised: 20 May 2026
Accepted: 25 May 2026
Published: 28 May 2026

**Abstract:** Anomaly detection in human-driven vehicle telemetry is complicated by mixed uncertainty: nominal deviations may arise either from stochastic driver behavior or from genuine departures from learned vehicle dynamics. Conventional forecasting-based detectors typically treat both as predictive error, which can produce heavy-tailed anomaly-score distributions and elevated false-positive rates under unseen driver behavior. To address this limitation, we propose the Uncertainty-Gated Mixture Model (U-GMM), a feature-wise anomaly-scoring framework that combines conditional probabilistic forecasting with marginal plausibility estimation through an uncertainty-aware gating mechanism. The conditional component captures temporal consistency with recent history, while the marginal component evaluates whether an observation remains plausible under the broader nominal feature distribution. The learned gate then uses predictive uncertainty to adaptively balance these two sources of anomaly evidence, reducing undue score inflation in nominally stochastic channels while preserving sensitivity to dynamically inconsistent or globally implausible deviations. Experiments on real-world vehicle telemetry datasets show that the proposed framework improves threshold transfer under unseen-driver evaluation, achieving up to a $2.5\times$ reduction in extreme false-positive rate while maintaining competitive fault detection performance under injected anomalies. These results indicate that reliable anomaly detection in human-in-the-loop systems depends not only on predictive model capacity, but also on uncertainty-aware score construction that distinguishes difficult-to-predict nominal behavior from genuinely abnormal system dynamics.

**Keywords:** anomaly detection; driver behavior; deep learning; false positives; multivariate time-series forecasting

## 1. Introduction

Unsupervised anomaly detection has become an important paradigm for vehicle health monitoring because fault events are rare, heterogeneous, and difficult to label comprehensively [1,2]. By training models exclusively on nominal telemetry, unsupervised anomaly detection methods can identify deviations from learned normality without requiring exhaustive fault annotations [3,4]. The ability to learn directly from nominal operating data makes unsupervised anomaly detection particularly well suited to modern vehicle monitoring, where high-dimensional multivariate telemetry is readily available but representative fault datasets are generally scarce [5].

The need for unsupervised monitoring approaches is amplified by the increasing complexity of modern vehicle mechatronic systems, in which tightly coupled sensing, control, and embedded software can obscure the observable effects of incipient degradation [6, 7]. In closed-loop operation, compensation and fault-tolerant control may attenuate the external manifestation of internal drifts, so that early-stage faults appear only as subtle deviations in internal telemetry rather than as clear threshold violations in performance variables [7]. The weak external manifestation of early-stage faults makes nominal-data-driven monitoring particularly attractive for early fault detection in high-dimensional vehicle systems [6].

In human-driven vehicle systems, early fault detection from telemetry is further complicated by the fact that telemetry is shaped not only by deterministic vehicle dynamics but also by partially observed driver behavior and operating context [8,9]. Signals associated with steering, braking, and acceleration may vary substantially under nominal conditions due to driver intent, traffic interactions, and road geometry, even when the vehicle remains mechanically healthy [9–11]. As a result, deviations in observed telemetry do not uniformly indicate abnormality [12,13]: large residuals in some channels may reflect benign behavioral variability, whereas comparable deviations in more tightly regulated vehicle-state variables may be more indicative of genuine system faults.

Most existing anomaly detection frameworks for multivariate vehicle telemetry model nominal behavior through reconstruction, forecasting, or likelihood estimation, and then derive anomaly scores from residual magnitude or predictive deviation [2,3]. While effective in many settings, the aforementioned approaches typically do not distinguish between channels that are intrinsically difficult to predict under nominal operation and channels whose deviations are more indicative of abnormal system behavior [12,14,15]. Treating all prediction residuals as equally informative can lead to a formulation mismatch in human-driven telemetry: behavior-sensitive variables with high nominal uncertainty may dominate aggregate anomaly scores, inflating false-positive rates even when the underlying system remains healthy [9,13]. The resulting sensitivity to nominal behavioral variability suggests that the central challenge lies not solely in representational capacity, but in anomaly scoring under mixed uncertainty.

To address this limitation, we propose the Uncertainty-Gated Mixture Model (U-GMM), a feature-wise anomaly scoring framework that combines conditional probabilistic forecasting with marginal density-based plausibility estimation through an uncertainty-aware gating mechanism. The conditional forecasting component quantifies the consistency of each observation with its recent temporal context, whereas the marginal plausibility component evaluates whether the same observation remains likely under the broader nominal feature distribution. A learned gate then uses predictive uncertainty to adaptively fuse the complementary anomaly signals. In channels for which nominal behavior is inherently difficult to predict, large conditional residuals need not imply abnormality; in such cases, the proposed framework reduces reliance on predictability-based evidence and instead assesses whether the observation remains globally plausible under nominal operation. Conversely, when temporal predictions are confident, deviations from the conditional forecast are treated as stronger indicators of abnormal behavior. In effect, the proposed framework learns to distinguish between telemetry channels and operating regimes in which predictive deviation is informative of abnormality and those in which nominal behavior is inherently difficult to predict. As a result, deviations in conditionally predictable signals are penalized more strongly, whereas deviations in more aleatoric channels are evaluated primarily through global plausibility when conditional uncertainty is high.

Unlike conventional reconstruction- or forecasting-based detectors, the proposed framework does not assume that predictive deviations are uniformly informative across telemetry channels and operating contexts. Instead, U-GMM explicitly models heterogeneous feature-wise predictability by combining conditional temporal consistency with global marginal plausibility through uncertainty-aware gating. The central contribution therefore lies not merely in the use of a more expressive forecasting architecture, but in a different anomaly-scoring formulation for human-driven telemetry, where nominal stochastic variability and fault-relevant deviations must be treated differently. By adaptively modulating anomaly evidence according to predictive uncertainty, the proposed framework reduces false-positive inflation arising from behavior-sensitive telemetry channels while preserving sensitivity to genuinely abnormal vehicle dynamics.

The main contributions of this paper are as follows:

- We formulate anomaly scoring in human-driven vehicle telemetry as a mixed-uncertainty problem, highlighting the need to distinguish behavior-induced aleatoric variability from fault-relevant deviations in more deterministic vehicle dynamics.
- We propose the Uncertainty-Gated Mixture Model (U-GMM), a unified feature-wise anomaly scoring framework that combines conditional probabilistic forecasting with marginal plausibility modeling.
- We develop a dynamic uncertainty-aware gating mechanism that uses predicted conditional uncertainty to adaptively weight complementary anomaly signals at each feature and timestep.
- We show through ablation and interpretability analysis that the resulting gating behavior is not only effective, but also consistent with the heterogeneous predictability structure of vehicle telemetry.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 formulates the anomaly detection problem under mixed uncertainty. Section 4 presents the proposed U-GMM framework. Section 5 describes the experimental setup, and Section 6 reports the empirical results. Section 7 discusses the findings, Section 8 outlines limitations and future work, and Section 9 concludes the paper.

## 2. Related Work

### *2.1. Deep Anomaly Detection for Multivariate Time Series*

Traditional statistical and threshold-based anomaly detection methods are effective under low-dimensional settings and strong modeling assumptions, but do not scale well to modern high-dimensional multivariate time series with nonlinear temporal dependencies and cross-variable interactions [1,3,16]. Classical machine learning approaches, including distance-based, clustering, and one-class methods, face similar limitations as dimensionality and system complexity increase [17, 18]. The limitations of classical methods in modeling high-dimensional multivariate temporal structure have motivated deep learning-based approaches that learn temporal representations directly from multivariate telemetry [1,2].

Existing deep anomaly detection methods for time series can be broadly categorized into reconstruction-based methods [19], forecasting-based methods [20], probabilistic likelihood-based models [3], and adversarial or hybrid frameworks [21,22]. Models such as OmniAnomaly [3] and USAD [21] learn nominal temporal structure and detect anomalies through deviations in reconstruction error, prediction error, or likelihood. More recent work further incorporates attention mechanisms and graph structures to capture complex inter-variable dependencies [23,24].

Despite differences in anomaly detection architecture, most existing frameworks do not explicitly account for heterogeneous feature-wise predictability. In particular, most existing frameworks typically derive anomaly scores from deviations relative to a single learned notion of nominal temporal behavior, without distinguishing channels that are intrinsically difficult to predict under nominal operation from channels whose deviations are more indicative of abnormal system behavior. The assumption that all residual deviations are equally informative becomes especially problematic in human-in-the-loop vehicle telemetry, where behavior-sensitive inputs can exhibit substantial nominal variability. The presence of feature-dependent nominal uncertainty motivate anomaly scoring formulations that adapt to feature-wise predictive uncertainty rather than treating all residual deviations as equally informative [3,23,25].

### *2.2. Context-Aware Anomaly Detection with Human-in-the-Loop Variability*

Anomalies are generally defined relative to an established notion of normal behavior [16]. In complex multivariate systems, however, abnormality often depends on operating context rather than absolute signal magnitude alone. Context-aware anomaly detection therefore evaluates deviations conditioned on system state, environmental conditions, or operating regime [16,26].

The distinction between nominal behavioral variability and genuinely abnormal system behavior is particularly important in human-in-the-loop vehicle systems, where driver control inputs constitute a major source of variability under healthy operating conditions [8]. Prior studies have shown that driving behavior differs substantially across individuals and also varies within individuals depending on context and functional state [27–31]. Although behavioral variability is informative for tasks such as driver identification and driving-style analysis, driver-dependent operating patterns introduce substantial heterogeneity into nominal vehicle telemetry from a fault-detection perspective.

Context-aware anomaly detection methods attempt to mitigate the influence of context-dependent variability by augmenting primary signals with auxiliary contextual information, such as system modes, environmental variables, or regime indicators [16, 26]. In time-series settings, context-aware anomaly detection has been implemented through regime-dependent models, conditional forecasting, and multi-stream architectures [32,33]. More recent work, such as the CADD framework [12], further incorporates estimated physical context to ground anomaly detection in mechanical invariants.

Comprehensive context modeling is generally infeasible in production vehicle systems. Many factors that shape driver behavior, including intent, internal state, and local interaction with traffic, are only partially observed or entirely unavailable. Moreover, the separation between exogenous driver inputs and endogenous system responses is often ambiguous in closed-loop vehicle telemetry, since many measured channels reflect mixtures of driver action, controller intervention, and environmental disturbance [8,9,34]. As a result, explicit feature augmentation alone does not eliminate regime-dependent uncertainty in signal predictability [9,12].

Recent intelligent transportation research has also explored broader AI-driven frameworks for traffic behavior modeling and vehicular-system monitoring. Hybrid TrafficAI [35] proposes a generative AI framework that integrates multimodal fusion, synthetic edge-case scenario generation, temporal-spatial attention, and LLM-based semantic reasoning for real-time traffic simulation, trajectory prediction, and anomaly detection. In a related intelligent transportation setting, hybrid ensemble-learning frameworks have also been investigated for attack detection in Internet-of-Vehicular-Things networks [36], demonstrating the use of adaptive AI architectures for robust vehicular-system monitoring. The aforementioned studies further highlight the growing importance of hybrid

and adaptive AI frameworks for modeling complex transportation-system behavior under heterogeneous operating conditions. The proposed U-GMM framework similarly addresses heterogeneous uncertainty in vehicle telemetry, but focuses specifically on adaptive anomaly scoring under partially observed human behavioral variability rather than traffic-level simulation or network-security monitoring.

In summary, explicit treatment of human-induced aleatoric variability in unsupervised vehicle fault detection remains limited although context-aware anomaly detection has been widely studied. Existing approaches do not directly address how anomaly scoring should adapt when large deviations arise from nominal but difficult-to-predict behavioral variability rather than from genuine system faults [3,12,23]. This limitation motivates uncertainty-aware scoring mechanisms that adapt to heterogeneous feature-wise predictability [25].

### *2.3. Hybrid and Uncertainty-Aware Anomaly Scoring*

Recent research in time-series anomaly detection has increasingly explored hybrid and probabilistic formulations, reflecting a broader shift from purely deterministic residual scoring toward more flexible anomaly-scoring strategies [2]. Representative examples include probabilistic latent-variable models such as OmniAnomaly [3] and hybrid adversarial-reconstruction frameworks such as USAD [21]. Probabilistic and hybrid anomaly detection frameworks demonstrate that anomaly detection can benefit from combining multiple modeling principles rather than relying on a single deterministic notion of normality [3,21].

Related work has also begun to incorporate predictive uncertainty more explicitly into anomaly detection, particularly in time-series settings where heteroscedasticity, noise, and distribution shift can reduce the reliability of residual-based scores. Recent probabilistic and reconstruction-based approaches have highlighted the importance of modeling uncertainty and latent variability in multivariate temporal data [3,23,25].

Existing hybrid and uncertainty-aware methods generally treat uncertainty as a property of a single predictive or reconstruction model, or as an auxiliary confidence quantity applied after anomaly evidence has already been formed. Existing approaches therefore do not explicitly address how anomaly scoring itself should adapt when nominal feature predictability varies substantially across telemetry channels and operating contexts. By contrast, U-GMM adaptively balances conditional temporal consistency and global marginal plausibility according to feature-wise predictive uncertainty at each feature and timestep.

## 3. Problem Formulation

In human-driven vehicle telemetry, observed signals are shaped by both deterministic vehicle dynamics and partially observed behavioral and contextual factors. As a result, nominal deviations in multivariate telemetry do not all carry the same diagnostic meaning: some arise from stochastic but valid driver behavior, whereas others are more indicative of abnormal departures from learned system dynamics. The coexistence of stochastic behavioral variability and deterministic system dynamics makes anomaly scoring in human-in-the-loop vehicle systems fundamentally different from anomaly scoring in more homogeneous dynamical settings.

We consider a multivariate time series

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\},$$

collected under nominal vehicle operation, where each observation $\mathbf{x}_t \in \mathbb{R}^D$ comprises both driver-controlled features and vehicle-state features. Given a look-back window of length $L$ ending at time $t$, denoted by $\mathbf{X}_{t-L+1:t} \in \mathbb{R}^{L\times D}$, the objective is to learn from nominal training data a model $\mathcal{M}_\theta$ that assigns an anomaly score to the subsequent observation $\mathbf{x}_{t+1}$:

$$s_{t+1} = S(\mathbf{x}_{t+1} \mid \mathbf{X}_{t-L+1:t}; \theta) \in \mathbb{R}. \tag{1}$$

An anomaly is declared when $s_{t+1} > \tau$, where $\tau$ is a threshold calibrated using nominal validation data.

### *3.1. Mixed Uncertainty Under Partial Observability*

A central difficulty is that important contextual factors, such as driver intent, traffic interactions, and local road conditions, are only partially represented in the observed telemetry. Let $\mathbf{z}_t$ denote such latent contextual influences. Even with access to the full observed history, the next-step uncertainty of some channels remains strictly positive:

$$H(\mathbf{x}_{t+1} \mid \mathbf{X}_{t-L+1:t}) > 0. \tag{2}$$

Uncertainty arising from partially observed driver behavior and operating context is especially pronounced in driver-controlled or behavior-sensitive channels, whose nominal evolution may be intrinsically stochastic and

regime-dependent. By contrast, many vehicle-state variables are more tightly constrained by physical dynamics and therefore exhibit comparatively lower conditional uncertainty under nominal operation.

### *3.2. Limitations of Residual-Based Scoring*

Most forecasting-based anomaly detectors define abnormality directly through predictive deviation, for example by using a score of the form

$$s_{t+1} = \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2, \tag{3}$$

where $\hat{\mathbf{x}}_{t+1} = f_\theta(\mathbf{X}_{t-L+1:t})$ is the predicted next observation. Residual-based anomaly scoring implicitly treats large deviations from the conditional mean as uniformly informative of abnormality. In mixed-uncertainty settings, however, elevated residuals may arise either from genuine departures from learned vehicle dynamics or from nominal but difficult-to-predict behavioral variability. Consequently, residual magnitude alone is an unreliable anomaly indicator.

An effective anomaly score for human-driven telemetry should therefore satisfy two requirements:

(1) Tolerance to nominal aleatoric variability:deviations arising from irreducible, context-dependent behavioral uncertainty should not dominate the anomaly score;
(2) Sensitivity to dynamic inconsistency: deviations that violate stable temporal or physical relationships should still be assigned high anomaly scores, even when their absolute magnitudes remain globally plausible.

Balancing robustness to nominal behavioral variability with sensitivity to genuine dynamic inconsistency motivates an anomaly-scoring formulation that distinguishes between conditional temporal inconsistency and global plausibility, since temporally unpredictable observations may nevertheless remain globally consistent with nominal operating behavior, rather than reflecting genuine system faults.

## 4. Methodology: Uncertainty-Aware Scoring

### *4.1. Conditional Probabilistic Forecasting*

The objective of the conditional forecasting component is to model the expected temporal evolution of vehicle telemetry given the recent system history. In human-driven vehicle systems, however, future observations are not always deterministically predictable because behavioral and contextual factors are only partially observed. As a result, observed history may be compatible with multiple nominal future trajectories. Deterministic forecasting may conflate normal behavioral variability with genuine departures from learned system dynamics, potentially inflating anomaly scores in behavior-sensitive channels.

To account for mixed uncertainty, the proposed framework models a full conditional distribution over the next observation rather than a single point estimate, as illustrated in Figure 1. The predicted mean captures the expected nominal trajectory, whereas the predicted variance provides a feature-wise signal of conditional predictability under the current temporal context. Channels that are consistently predictable given the available observed context tend to exhibit lower predicted variance, whereas behavior-sensitive channels may exhibit higher predicted variance under nominal operation because important contextual influences are only partially observed.

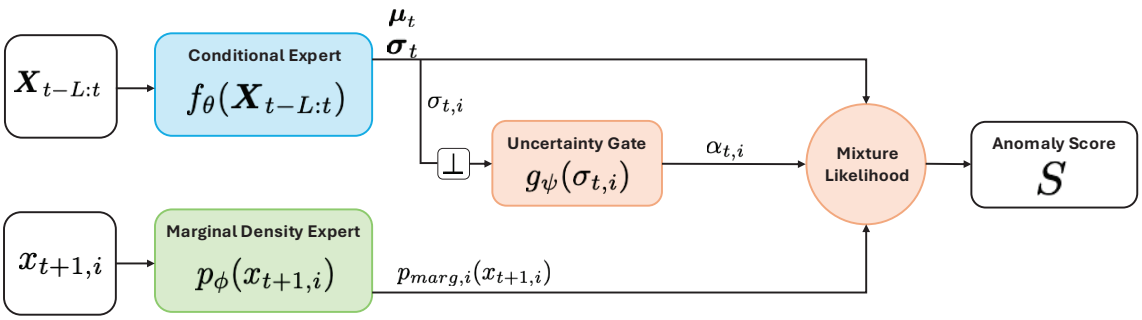


**Figure 1.** Architecture of the Uncertainty-Gated Mixture Model. The conditional forecasting component $f_\theta$ (LSTM) produces feature-wise mean and variance estimates $(\mu_{t,i}, \sigma_{t,i})$. The uncertainty gate $g_\psi$ computes the mixing coefficient $\alpha_{t,i}$ from the predicted variance using a stop-gradient operation ($\perp$). A per-feature mixture likelihood combines the conditional density with the marginal density model $p_\phi(x_{t+1,i})$ to produce the anomaly score $S$.

Accordingly, the proposed formulation models the full conditional density of the next observation given the recent history:

$$p_{\text{cond}}(\mathbf{x}_{t+1} \mid \mathbf{X}_{t-L+1:t}) = \mathcal{N}(\mathbf{x}_{t+1}; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\,. \tag{4}$$

To capture feature-wise heteroscedastic uncertainty while maintaining tractability, the covariance is modeled as diagonal, i.e.,

$$\boldsymbol{\Sigma}_t = \text{diag}(\boldsymbol{\sigma}_t^2). \tag{5}$$

The diagonal covariance assumption enables channel-specific uncertainty estimation with substantially lower complexity than a full covariance model, while remaining suitable for real-time multivariate telemetry forecasting.

The distribution parameters are produced by a Long Short-Term Memory (LSTM) network that maps the input window $\mathbf{X}_{t-L+1:t}$ to a latent hidden state $\mathbf{h}_t \in \mathbb{R}^H$. Two linear projection heads then generate the mean and standard deviation parameters:

$$\boldsymbol{\mu}_t = \mathbf{W}_\mu \mathbf{h}_t + \mathbf{b}_\mu, \qquad \boldsymbol{\sigma}_t = \text{softplus}(\mathbf{W}_\sigma \mathbf{h}_t + \mathbf{b}_\sigma) + \epsilon, \tag{6}$$

where $\mathbf{W}_\mu, \mathbf{W}_\sigma \in \mathbb{R}^{D\times H}$ and $\mathbf{b}_\mu, \mathbf{b}_\sigma \in \mathbb{R}^D$ are learnable parameters. Under the conditional Gaussian model, $\boldsymbol{\mu}_t$ captures the expected nominal trajectory, whereas $\boldsymbol{\sigma}_t^2$ quantifies feature-wise predictive uncertainty under the current temporal context.

The conditional forecasting model is trained exclusively on nominal data by minimizing the Gaussian negative log-likelihood (up to an additive constant). Under the diagonal Gaussian assumption, the objective decomposes into an element-wise sum:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{(T-1)D} \sum_{t=1}^{T-1} \sum_{i=1}^{D} \left[ \frac{(x_{t+1,i} - \mu_{t,i})^2}{2\sigma_{t,i}^2} + \log \sigma_{t,i} \right]. \tag{7}$$

The Gaussian negative log-likelihood objective jointly penalizes inaccurate predictions and poorly calibrated uncertainty estimates, allowing the model to represent context-dependent nominal uncertainty rather than forcing all channels toward the same deterministic predictability.

Nevertheless, conditional likelihood alone remains insufficient for anomaly scoring in mixed-uncertainty telemetry. Low conditional likelihood may arise either from genuinely abnormal system behavior or from nominal but difficult-to-predict behavioral variability under partial observability. Consequently, temporal surprise alone is not always a reliable indicator of abnormality. The ambiguity between genuinely abnormal behavior and nominal but difficult-to-predict variability motivates the incorporation of an additional marginal plausibility component that evaluates whether an observation remains compatible with the broader nominal operating distribution independently of immediate temporal predictability. Although the conditional forecasting component is instantiated here using an LSTM, the proposed uncertainty-gated scoring framework is model-agnostic and requires only a probabilistic forecasting backbone capable of producing feature-wise mean and variance estimates.

### *4.2. Marginal Plausibility Modeling*

While conditional forecasting evaluates temporal consistency relative to recent system history, observations that are difficult to predict may be assigned a low likelihood, despite remaining globally plausible under nominal operation. Unlike conditional forecasting, which evaluates how well an observation aligns with the expected temporal trajectory, the marginal component evaluates whether the observation itself remains plausible under nominal operation, regardless of whether it was temporally expected.

To capture the notion of global plausability, we introduce a context-independent marginal density model $p_{\text{marg},i}(x_i)$ for each feature.

Vehicle telemetry channels often exhibit complex and multi-modal marginal distributions. For example, driver-controlled variables such as pedal activation may concentrate around multiple operating regimes rather than following a simple unimodal distribution. To model such behavior, we employ normalizing flows [37]. Specifically, for each feature $x_i$, we define a bijective transformation $f_{\phi_i} : \mathbb{R} \to \mathbb{R}$ that maps the observed marginal distribution to a simple latent variable $z_i \sim \mathcal{N}(0, 1)$. By the change-of-variables formula, the marginal log-likelihood is given by

$$\log p_{\text{marg},i}(x_{t,i}) = \log p_{\mathcal{N}}(f_{\phi_i}(x_{t,i})) + \log \left| \frac{df_{\phi_i}}{dx_{t,i}} \right|. \tag{8}$$

In our implementation, $f_{\phi_i}$ is parameterized using monotonic rational quadratic splines (RQS) [38], which provide a flexible and differentiable family of invertible transformations. An independent flow is trained for each feature by maximizing marginal log-likelihood on nominal training data. Modeling each feature distribution with an independent normalizing flow is computationally efficient and aligns naturally with the feature-wise structure of the

proposed uncertainty-gated scoring framework, although the resulting marginal model does not explicitly capture cross-feature dependence.

The resulting marginal model assigns low likelihood to observations that fall outside the nominal operating envelope, even when large predictive residuals may arise from either genuine anomalies or nominal behavioral variability. Conversely, observations arising from temporally surprising but globally common nominal behavior retain high marginal probability. The marginal plausibility term therefore provides information that is complementary to conditional forecasting and is particularly useful when large predictive residuals arise from nominal but hard-to-predict variability.

The marginal component therefore does not replace temporal forecasting, but instead provides complementary plausibility evidence when predictive deviation alone becomes ambiguous under mixed uncertainty.

### *4.3. Uncertainty-Gated Anomaly Scoring*

The proposed anomaly-scoring formulation combines conditional temporal consistency and global marginal plausibility into a unified scoring framework. Conditional likelihood and marginal plausibility capture complementary notions of nominality: conditional likelihood evaluates whether an observation is temporally consistent with recent system evolution, whereas marginal plausibility evaluates whether the same observation remains globally compatible with nominal operation. Under mixed uncertainty, however, conditional deviations are not equally informative across all features and contexts. The proposed formulation therefore adapts the relative contribution of conditional and marginal evidence according to the feature-wise predictive uncertainty estimated by the forecasting component.

For each feature $i \in \{1, \ldots, D\}$ at time $t$, let

$$p_{\mathrm{cond},i} = \mathcal{N}(x_{t+1,i}; \mu_{t,i}, \sigma_{t,i}^2) \tag{9}$$

denote the conditional predictive density, and let

$$p_{\mathrm{marg},i} = p_{\mathrm{marg},i}(x_{t+1,i}) \tag{10}$$

denote the corresponding marginal plausibility term.

The two quantities are fused through a feature-wise mixture density

$$p_{\mathrm{mix},i} = (1 - \alpha_{t,i})\, p_{\mathrm{cond},i} + \alpha_{t,i}\, p_{\mathrm{marg},i}, \tag{11}$$

where $\alpha_{t,i} \in [0, 1]$ is produced by the uncertainty gate. The gating coefficient is conditioned exclusively on the predicted uncertainty $\sigma_{t,i}$ through a stop-gradient operation, so that routing decisions do not backpropagate into the variance estimates and artificially distort uncertainty calibration.

The feature-level anomaly score assigned to the target observation at time $t + 1$ is then defined as the negative log likelihood of the resulting mixture:

$$s_{t+1,i} = -\log p_{\mathrm{mix},i}. \tag{12}$$

For numerical stability, this computation is implemented using a `LogSumExp` formulation. The aggregate anomaly score is obtained by summing over features,

$$S_{t+1} = \sum_{i=1}^{D} s_{t+1,i}, \tag{13}$$

which is consistent with the feature-wise scoring structure adopted throughout the framework.

The uncertainty-gated anomaly score allows the anomaly score to adapt to heterogeneous nominal predictability. When conditional uncertainty is high, deviations from the forecast require additional contextual interpretation, since large predictive residuals may arise from nominal behavioral variability as well as anomalous behavior. When conditional uncertainty is low, deviations from the forecast contribute more strongly to the anomaly score because predictive residuals are more indicative of departures from expected system dynamics. By modulating the influence of conditional residuals according to predictive uncertainty, the proposed scoring rule reduces undue score inflation in behavior-sensitive channels while preserving sensitivity to dynamically inconsistent or globally implausible behavior.

Unlike fixed-weight hybrid scoring rules, the proposed formulation conditions the balance between temporal inconsistency and marginal plausibility on feature-wise predictive uncertainty.

### *4.4. Uncertainty Gating Mechanism*

The proposed uncertainty-aware scoring framework requires a feature-wise mechanism for determining how conditional temporal evidence and marginal plausibility should be combined. The proposed model therefore introduces a lightweight uncertainty gate that maps predicted conditional uncertainty to a context-dependent fusion coefficient $\alpha_{t,i}$. Conceptually, the gate determines how strongly predictive deviation should influence anomaly scoring under the current temporal context.

The gate is parameterized as a multilayer perceptron (MLP) $g_\psi$ followed by a sigmoid activation:

$$\alpha_{t,i} = \sigma_{\text{sigmoid}}\big(g_\psi(\perp(\log \sigma_{t,i}))\big), \tag{14}$$

where $\sigma_{t,i}$ denotes the predicted standard deviation from the conditional forecasting component and $\perp(\cdot)$ denotes a stop-gradient operation. The stop-gradient prevents the mixture objective from backpropagating into the variance estimates, thereby avoiding degenerate behavior in which the forecasting model artificially inflates uncertainty in order to shift responsibility toward the marginal plausibility component.

Detaching the uncertainty estimates from gradient propagation preserves the interpretability and calibration of the conditional uncertainty estimates. Specifically, the stop-gradient $\perp$ is critical for the structural stability of the mixture; without it, the forecasting network $\theta$ and the gating network $\psi$ may enter a competitive cycle where the forecaster artificially inflates $\sigma$ to delegate difficult samples to the marginal model, potentially leading to a collapse of the learned temporal representation.

#### 4.4.1. Gate Training Objective

With the conditional forecaster and marginal density model frozen, the gate is trained by minimizing the probability-space mixture negative log-likelihood

$$m_{t,i} = (1 - \alpha_{t,i})\, e^{-\ell_{\text{f},t,i}} + \alpha_{t,i}\, e^{-\ell_{\text{m},t,i}}. \tag{15}$$

$$\mathcal{L}_{\text{mix}} = \frac{1}{(T-1)D} \sum_{t=1}^{T-1} \sum_{i=1}^{D} -\log m_{t,i}. \tag{16}$$

where $\ell_{\text{f},t,i}$ and $\ell_{\text{m},t,i}$ denote the per-feature conditional and marginal negative log-likelihood terms, respectively. Equivalently, $\ell_{\text{f},t,i} = -\log p_{\text{cond},i}(x_{t+1,i})$ and $\ell_{\text{m},t,i} = -\log p_{\text{marg},i}(x_{t+1,i})$, so that the gate-training objective is algebraically consistent with the mixture score in (12).

To stabilize routing, we additionally apply together with sparsity and auxiliary supervision termsduring gate training. The precise regularization details are described in the experimental setup.

#### 4.4.2. Training Procedure

The three components are optimized sequentially. First, the conditional forecasting component is trained using the objective in (7) and then frozen. Second, the marginal density models are trained independently on nominal data and frozen. Finally, with both experts fixed, the uncertainty gate is trained using the probability-space mixture objective together with sparsity and auxiliary supervision terms to stabilize routing.

## 5. Experimental Setup

The proposed framework is evaluated on three real-world vehicle telemetry datasets under both nominal false-positive analysis and synthetic fault-injection testing. The experiments are designed to assess whether U-GMM reduces false positives induced by stochastic driver behavior while preserving sensitivity to mechanically meaningful anomalies.

### *5.1. Datasets and Split Protocols*

To evaluate the robustness of the proposed framework under diverse sources of nominal variability, three real-world vehicle telemetry datasets are utilized. The datasets are selected to capture complementary forms of behavioral and contextual uncertainty that commonly affect time-series anomaly detection systems.

The HCRL dataset [27] contains telemetry collected from 10 drivers operating the same vehicle over a fixed route, thereby isolating inter-driver behavioral variability under controlled vehicle and environmental conditions. The primary challenge in this dataset is cross-driver style shift, where unseen driving behaviors may distort anomaly score distributions and increase false-positive detections.

The Sonata dataset [39] contains telemetry from 4 drivers across heterogeneous urban trips, introducing combined driver, route, and environmental variability. Since external contextual factors such as traffic conditions,

road geometry, and pedestrian interactions are only partially observable from telemetry signals, nominal driver control actions may appear highly stochastic, resulting in large forecasting residuals in behavior-sensitive channels.

In addition, a single-driver multi-route OBD dataset [40] is employed to evaluate intra-driver contextual variability. Although subject variability is eliminated, changes in driving environments (e.g., highway versus urban traffic) alter the temporal predictability of control signals, producing localized temporal deviations that conventional detectors may incorrectly classify as anomalies.

For HCRL, driver-disjoint partitions are constructed using 6 drivers for training, 2 for validation, and 2 for testing. To reduce sensitivity to any single assignment, results are aggregated over 5 randomized split realizations. For Sonata, evaluation is performed over 4 driver-disjoint splits, such that each driver serves once as the held-out test subject while the remaining drivers are partitioned into training and validation subsets. Results are aggregated over all 4 split realizations. For the single-driver OBD dataset, subject-level partitioning is not applicable; instead, trips are divided into training, validation, and test subsets with proportions of 70%, 10%, and 20%, respectively. To reduce sensitivity to any particular trip assignment, results are aggregated over 10 randomized trip partition realizations.

### *5.2. Preprocessing and Window Construction*

All retained telemetry channels are normalized using statistics computed exclusively from the training split, and the same transformation is then applied unchanged to the corresponding validation and test data. Input sequences are constructed using sliding windows of length $L = 20$ with stride 1. Windowing is performed independently within each trip or file to prevent leakage across temporal boundaries. Since all datasets are sampled at 1 Hz, this corresponds to a 20-s look-back horizon and a 1 Hz anomaly-scoring frequency.

### *5.3. Implementation and Training Protocol*

The framework is implemented in PyTorch. The conditional forecasting component is instantiated as an LSTM-based probabilistic forecaster with hidden dimension 64 and 2 recurrent layers. The uncertainty gate is implemented as a lightweight multilayer perceptron. The marginal plausibility component is modeled using Rational Quadratic Spline (RQS) normalizing flows [38], with 8 spline bins, tail bound 10, and hidden dimension 64. The aforementioned settings were selected to balance expressive marginal modeling against computational cost.

All components are optimized using Adam with fixed batch size 128 and no learning-rate scheduler, so as to avoid masking structural learning behavior through adaptive optimization schedules. To stabilize optimization and prevent co-adaptation between components, training is performed in three stages:

(1) the conditional forecasting model is trained on nominal data using the Gaussian negative log-likelihood objective in (7) with learning rate $10^{-3}$;
(2) the marginal density models are trained independently by maximizing marginal log-likelihood with learning rate $5 \times 10^{-4}$;
(3) with both components fixed, the uncertainty gate is trained using the probability-space mixture objective together with sparsity and auxiliary supervision terms, using learning rate $3 \times 10^{-4}$.

Early stopping based on validation performance is applied at each stage.

#### 5.3.1. Gate Regularization

To stabilize routing, the gate is trained using the mixture objective together with a sparsity penalty and an auxiliary supervision term:

$$\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{mix}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{aux}}, \tag{17}$$

$$\mathcal{L}_{\text{reg}} = \lambda\, \mathbb{E}[\alpha]. \tag{18}$$

The gate output is denoted by $\alpha_{t,i} \in [0, 1]$. The regularization term $\mathcal{L}_{\text{reg}}$ acts as an $L_1$ sparsity penalty on the gate activations, discouraging trivial collapse toward the marginal pathway. In addition, $\mathcal{L}_{\text{aux}}$ denotes an auxiliary binary cross-entropy supervision term that encourages agreement between the learned gate activations and the relative discrepancy between conditional forecasting loss and marginal plausibility loss.

#### 5.3.2. Marginal Model Training Efficiency

Independent marginal density estimation for each feature can become computationally expensive when flow models are trained on every timestep of every file in larger datasets. In practice, moderate subsampling of nominal training data was sufficient to obtain stable marginal density estimates while reducing computation. The

independence of the feature-wise marginal models from temporal ordering allows the marginal component to remain effective under moderate subsampling.

### *5.4. Baselines*

The baselines are chosen in an explicitly ablative manner so as to isolate the contribution of each component of the proposed scoring formulation, rather than to maximize predictive performance through increasingly expressive backbone architectures. In particular, all compared models are built on the same LSTM forecasting backbone, allowing performance differences to be attributed to the anomaly-scoring formulation rather than to changes in model class or sequence capacity.

We compare U-GMM against two progressively simplified baselines:

(1) LSTM-AD (MSE): a deterministic LSTM forecaster trained using mean squared error, with residual magnitude used directly as the anomaly score. LSTM-AD isolates conventional point-prediction-based anomaly detection.
(2) Gaussian-LSTM: the probabilistic forecasting component described in Section 4.1, evaluated using conditional negative log-likelihood alone. Gaussian-LSTM isolates the effect of replacing deterministic point forecasting with heteroscedastic probabilistic prediction.

The resulting comparison is nested: LSTM-AD evaluates deterministic residual scoring, Gaussian-LSTM isolates the effect of conditional probabilistic modeling, and U-GMM isolates the incremental effect of augmenting conditional forecasting with marginal plausibility modeling and uncertainty-gated feature-wise score fusion. Accordingly, the scope of the comparison is not to identify the most powerful sequence backbone, but to evaluate the benefit of reformulating anomaly scoring under mixed uncertainty.

### *5.5. Evaluation Protocol*

#### 5.5.1. Nominal False-Positive Evaluation

False Positive Rate (FPR) is evaluated on nominal held-out test data. For a validation score sequence $\{S_t^{(\mathrm{val})}\}$, the detection threshold is defined as the empirical upper quantile

$$\tau_q = Q_q\Big(\{S_t^{(\mathrm{val})}\}\Big), \tag{19}$$

where $Q_q(\cdot)$ denotes the empirical quantile at level $q \in \{0.99, 0.999\}$. The threshold $\tau_q$ is calibrated on the validation split and then transferred unchanged to the corresponding held-out test split.

Evaluating nominal false-positive rate (FPR) is particularly important in human-in-the-loop vehicle systems, since deployed systems are expected to operate under nominal conditions for the vast majority of runtime. During nominal operation, normal behavioral variability may inflate anomaly scores and produce excessive spurious detections. The proposed framework is specifically designed to reduce such false-positive inflation while preserving sensitivity to genuinely abnormal system behavior.

#### 5.5.2. Synthetic Fault Evaluation

As the available datasets contain only nominal driving data, detection sensitivity is evaluated using controlled synthetic fault injection on the held-out test split. Faults are injected into randomly selected dynamically active windows of duration 20–30 timesteps, with up to three injected segments per file and no overlap between injected intervals. When appropriate, anomalies affect multiple channels so as to better reflect realistic coupled faults. Injection locations are chosen conditionally on signal activity, e.g., accelerator-related faults are introduced only during intervals in which the vehicle is actively accelerating, so that the resulting perturbations are both mechanically plausible and operationally relevant.

We consider three representative anomaly types:

- Actuator freeze:a representative driver-controlled input is held constant over an active segment, simulating a stuck-at or mechanically bound actuator;
- Boundary drift: a gradual drift is introduced into a physically constrained vehicle-state variable until it moves beyond its nominal operating envelope;
- Correlation inversion: one or more physically coupled response variables are manipulated so as to violate learned system relationships while remaining marginally plausible in isolation.

Synthetic fault injection is used to evaluate whether reductions in nominal false positives are achieved without reducing sensitivity to genuine anomalies. Detection performance is evaluated using AUROC as a threshold-independent measure of discriminative capability across varying decision thresholds. In addition, precision, recall,

and F1-score are reported under the quantile-based thresholds calibrated on the validation split in order to assess the practical trade-off between true-positive detection and false-positive suppression.

## 6. Results

We evaluate the proposed framework along three complementary dimensions: nominal false-positive behavior under cross-driver generalization, detection sensitivity under synthetic fault injection, and the behavior of the uncertainty gate under heterogeneous feature predictability.

### *6.1. Statistical Generalization and Threshold Stability Analysis*

We first evaluate the proposed framework on the HCRL dataset, which serves as the primary benchmark for cross-driver generalization. Detection thresholds are calibrated on validation subjects using fixed upper quantiles of the nominal anomaly-score distribution and then transferred unchanged to unseen test subjects. Table 1 reports the resulting False Positive Rates (FPRs), averaged over 10 randomized driver-disjoint split realizations. Across both operating points, U-GMM achieves the lowest held-out FPR, indicating improved threshold transferability under novel driver behavior.

At the 99.9th percentile, U-GMM attains a test FPR of $0.00039$, compared with $0.00099$ and $0.00096$ for Gaussian-LSTM and LSTM-AD (MSE), respectively. Relative to both baselines, the held-out false-positive rate achieved by U-GMM represents an approximate $2.5\times$ reduction in false positives at the stricter operating point. U-GMM also achieves the lowest held-out FPR at the 99th percentile, indicating that the reduction in false positives remains consistent across operating thresholds.

**Table 1.** False Positive Rate (FPR) on unseen HCRL test subjects using validation-calibrated quantile thresholds (0.99 and 0.999). Relative gain is reported with respect to U-GMM. Results are averaged over 10 randomized driver-disjoint split realizations.

| Model | Quantile | Threshold | FPR | Gain ($\times$) |
|---|---|---|---|---|
| U-GMM | 0.99 | 1.59 | **0.00287** | — |
| Gaussian-LSTM | 0.99 | 4.25 | 0.00580 | 2.02 |
| LSTM-AD (MSE) | 0.99 | 1.89 | 0.00522 | 1.82 |
| U-GMM | 0.999 | 3.36 | **0.00039** | — |
| Gaussian-LSTM | 0.999 | 37.04 | 0.00099 | 2.54 |
| LSTM-AD (MSE) | 0.999 | 4.33 | 0.00096 | 2.45 |

Bold values indicate the lowest false-positive rate (FPR) at each operating point.

A second important observation concerns threshold scaling at increasingly stringent operating points. When the target quantile is raised from $0.99$ to $0.999$, Gaussian-LSTM requires a substantial threshold increase ($4.25$ to $37.04$), indicating a heavy-tailed anomaly-score distribution. By contrast, U-GMM exhibits a much smaller escalation factor ($2.1\times$), comparable to the deterministic LSTM-AD baseline ($2.3\times$). The substantially smaller threshold escalation observed for U-GMM indicates that the proposed scoring formulation suppresses extreme score inflation caused by high-variance behavioral events, thereby improving the stability of threshold transfer under stringent false-positive constraints.

Taken together, our results indicate that uncertainty-gated fusion improves not only nominal false-positive performance, but also the concentration and transfer stability of the anomaly-score distribution under cross-driver generalization.

### *6.2. Detection Sensitivity Under Synthetic Fault Injection*

While the preceding analysis demonstrates improved false-positive stability, it is equally important to verify that uncertainty-gated scoring does not degrade sensitivity to genuine anomalies. We therefore evaluate detection performance under the injected fault scenarios described in Section 5.5. Aggregate results are summarized in Table 2.

U-GMM achieves the highest AUROC (0.763), as well as the best precision, recall, and F1-score among the evaluated models. In particular, recall increases to 0.857, compared with 0.831 for Gaussian-LSTM and 0.796 for LSTM-AD (MSE), indicating that the proposed gating mechanism preserves sensitivity to injected faults rather than suppressing true positive detections.

A comparison of raw detection counts is consistent with this trend. U-GMM produces substantially fewer false positives (FP = 8669) than LSTM-AD (FP = 12,511) while simultaneously identifying more true positives

(TP = 2100 versus 1950). Gaussian-LSTM achieves intermediate performance, but does not attain the same precision–recall balance as the uncertainty-gated formulation.

**Table 2.** Aggregate detection performance under injected fault scenarios, averaged over 5 cross-validation splits.

| Model | AUROC | Precision | Recall | F1 |
|---|---|---|---|---|
| U-GMM | **0.763** | **0.195** | **0.857** | **0.318** |
| Gaussian-LSTM | 0.752 | 0.185 | 0.831 | 0.303 |
| LSTM-AD (MSE) | 0.709 | 0.135 | 0.796 | 0.231 |

Bold values indicate the best performance achieved for each evaluation metric.

Taken together, our results indicate that uncertainty-gated fusion improves separability between nominal behavioral variability and mechanically meaningful anomalies. By reducing undue score inflation in stochastic but nominal channels, U-GMM lowers false-positive burden while retaining strong sensitivity to dynamically inconsistent or globally implausible fault patterns.

### *6.3. Interpretability of the Uncertainty Gate*

To examine whether the learned gating behavior is consistent with the intended mixed-uncertainty formulation, we analyze the distribution of gate activations $\alpha$ across two interpretable feature categories: directly actuated driver-input channels and vehicle-state response channels. Representative driver-input channels include accelerator pedal position, brake switch status, and steering wheel angle, whereas representative vehicle-state channels include engine speed, wheel velocity, and intake air pressure.

As shown in Figure 2, the learned gate assigns substantially larger activations to driver-controlled features than to vehicle-state features. The driver-controlled group exhibits a broad distribution with median $\alpha \approx 0.74$ and mean $\alpha \approx 0.56$, whereas the vehicle-state group remains tightly concentrated near zero, with median $\alpha \approx 0.008$ and mean $\alpha \approx 0.060$. The interquartile range for the driver-controlled group is approximately $[0.076,\ 0.994]$, compared with only $[0.005,\ 0.014]$ for the vehicle-state group. In addition, approximately $52.2\%$ of driver-controlled feature instances satisfy $\alpha > 0.5$, compared with only $5.0\%$ of vehicle-state feature instances. Experiment results indicate that the gating mechanism adapts systematically to heterogeneous feature-wise predictability rather than acting uniformly across channels.

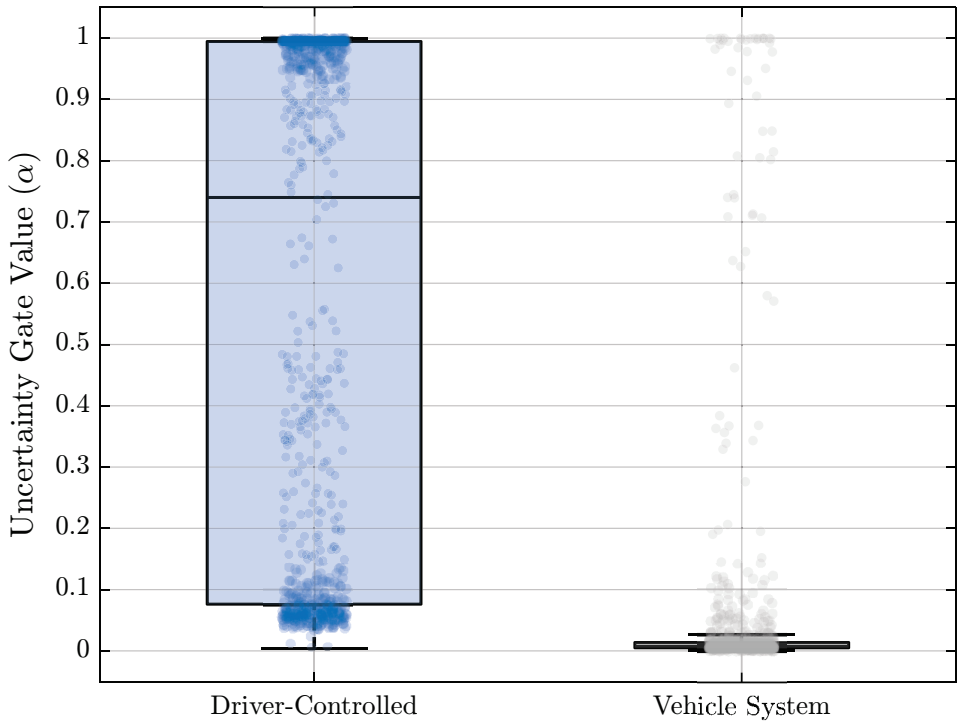


**Figure 2.** Distribution of gate values $\alpha$ for driver-controlled and vehicle-state feature groups. Driver-controlled channels exhibit substantially higher gate activations, indicating greater reliance on marginal plausibility in channels with lower nominal predictability, whereas vehicle-state features remain concentrated near zero.

The uncertainty gate additionally provides an interpretable indication of feature-wise nominal predictability. Higher gating responses indicate telemetry channels or timesteps for which conditional temporal forecasting is less reliable under nominal operation, typically due to partially observed contextual influences such as driver intent, traffic interactions, or environmental variability. For telemetry channels and timesteps associated with elevated predictive uncertainty, anomaly scoring therefore relies more strongly on global marginal plausibility rather than strict temporal forecasting consistency. Lower gating responses indicate more temporally predictable telemetry behavior, allowing anomaly scoring to place greater emphasis on conditional forecasting consistency.

Channels associated with direct human control inputs, such as accelerator or steering behavior, therefore tend to exhibit elevated predictive uncertainty, whereas tightly constrained vehicle-state variables such as engine-speed or wheel-speed relationships typically retain lower uncertainty and stronger temporal predictability. The gating mechanism consequently provides interpretable insight into which telemetry features behave deterministically under nominal operation and which remain intrinsically context-sensitive or behavior-dependent.

#### 6.3.1. Feature-Level Gate Ranking

The feature-level gate statistics further refine this interpretation. Table 3 shows that the highest average gate activations are assigned to driver-controlled or directly behavior-linked channels, including Steering Wheel Speed, Brake Switch, Clutch Operation, and Accelerator Pedal Value. By contrast, the lowest gate activations are associated with tightly constrained vehicle-state variables such as Wheel Velocity, Vehicle Speed, and Engine Speed. The systematic allocation of higher gate activations to behavior-sensitive channels, and lower activations to physically constrained vehicle-state variables indicates that the gate does not merely separate broad feature groups, but also captures meaningful differences in nominal predictability within and across telemetry channels.

**Table 3.** Feature-level average gate activations $\alpha$.

| **Highest Gate Values** | | **Lowest Gate Values** | |
|---|---|---|---|
| **Feature** | $\boldsymbol{\alpha}$ | **Feature** | $\boldsymbol{\alpha}$ |
| Steering Wheel Speed | 0.759 | Wheel Velocity (FL) | 0.009 |
| Brake Switch | 0.619 | Vehicle Speed | 0.009 |
| Clutch Operation | 0.530 | Engine Speed | 0.012 |
| Accel. Pedal Value | 0.478 | Long. Acceleration | 0.016 |
| Master Cyl. Pressure | 0.341 | Intake Air Pressure | 0.019 |

The group-level separation in Figure 2, together with the feature-level ranking in Table 3, is therefore consistent with the intended behavior of U-GMM: channels with intrinsically higher nominal uncertainty are scored more cautiously, while tightly constrained vehicle-state variables remain dominated by conditional forecasting evidence.

### *6.4. Case Study of Feature-Wise Gate Dynamics*

To further examine the behavior of the learned gate, we consider a representative pair of mechanically related signals: accelerator pedal position and engine speed. Figure 3 shows both the raw feature trajectories and their corresponding gate activations over a full driving segment, together with a zoomed local window.

While the two signals are clearly coupled in the raw telemetry, their gate behaviors differ substantially. Engine speed exhibits large amplitude variation but remains associated with gate values near zero throughout the sequence, indicating persistent reliance on conditional temporal prediction. By contrast, accelerator pedal position shows markedly more dynamic gate behavior, with elevated $\alpha$ values during active and rapidly varying control segments and lower values during more stable intervals.

The differing gate responses assigned to coupled telemetry signals indicate that gate activation is not determined by signal magnitude alone. Instead, the learned routing behavior is more consistent with feature-wise conditional predictability: dynamically constrained vehicle-state variables remain dominated by the forecasting term, whereas driver-controlled signals receive greater marginal weighting when their nominal evolution is less predictable. The zoomed segment further illustrates that this routing occurs asynchronously across features at the same time step, supporting the view that U-GMM performs genuinely feature-wise adaptive scoring rather than global regime switching.

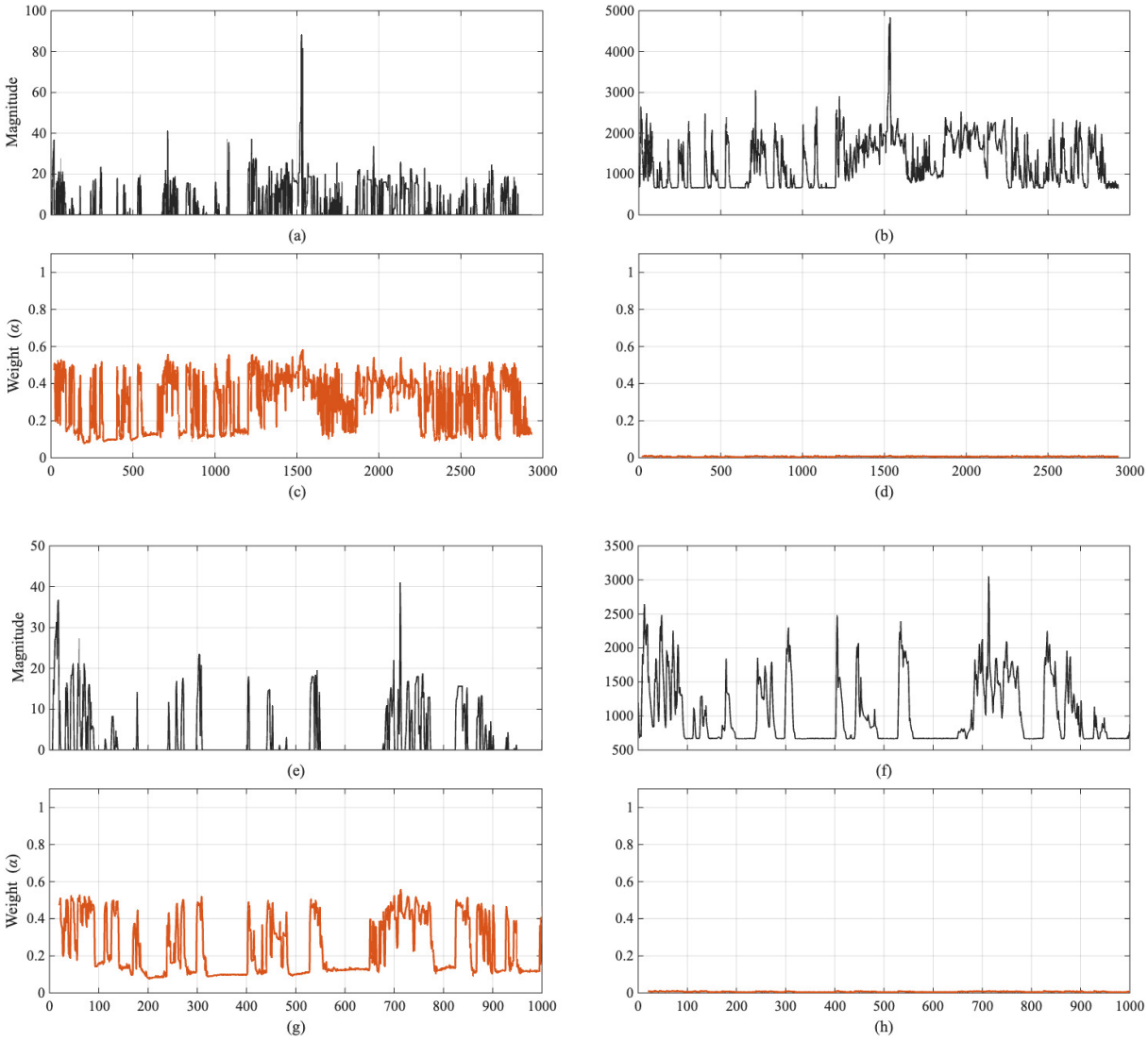


**Figure 3.** Feature values and corresponding gate activations over a full driving segment. Top row (Full Segment): (**a**) accelerator pedal signal, (**b**) engine speed signal, (**c**) accelerator gate $\alpha_{t,i}$, and (**d**) engine speed gate $\alpha_{t,i}$. Bottom row: (**e**)–(**h**) show a detailed temporal zoom of the exact signals presented in (**a**)–(**d**), respectively, highlighting the feature-wise asynchronous gating behavior within a 1000-timestep interval.

### *6.5. Ablation Study: Effect of Dynamic Uncertainty Gating*

To isolate the contribution of adaptive gating, we evaluate fixed mixing policies in which the gating coefficient is held constant at $\alpha \in \{0.3, 0.5, 0.7\}$. The comparison between adaptive and fixed gating assesses whether the performance gains of U-GMM arise simply from introducing a marginal plausibility term, or from dynamically adapting the balance between conditional and marginal anomaly evidence. Lower values of $\alpha$ place greater emphasis on conditional temporal forecasting, whereas larger values increasingly favor marginal plausibility. Fixed-gating configurations therefore impose a single global balance between temporal consistency and marginal plausibility across all telemetry channels and operating conditions. Heterogeneous telemetry uncertainty makes this assumption overly restrictive, since nominal predictability may vary substantially across telemetry channels, operating contexts, and timesteps.

#### 6.5.1. Static Gating Trade-Off

Table 4 summarizes performance under fixed-gating configurations. A monotonic relationship is observed between $\alpha$ and nominal stability: as $\alpha$ increases from 0.3 to 0.7, the False Positive Rate at the 0.95 operating point decreases from 0.0166 to 0.0103. The monotonic reduction in false-positive rate with increasing $\alpha$ indicates that stronger reliance on the marginal pathway suppresses stochastic variability more aggressively. The increased stability is accompanied by reduced fault sensitivity. While AUROC varies only modestly across configurations, the F1-score decreases from 0.154 at $\alpha = 0.3$ to 0.105 at $\alpha = 0.7$, corresponding to a decline of approximately 32%.

The observed reduction in F1-score with increasing $\alpha$ suggests that fixed reliance on marginal plausibility introduces an overly conservative bias, suppressing not only benign behavioral variability but also informative fault-induced deviations. Lower values of $\alpha$ preserve stronger temporal sensitivity by emphasizing conditional forecasting, whereas larger values increasingly favor conservative marginal plausibility scoring at the expense of fault discrimination.

**Table 4.** Ablation study with fixed gating coefficients $\alpha \in \{0.3, 0.5, 0.7\}$. FPR is reported at multiple quantiles; AUROC and F1 are evaluated on injected faults.

| Gate Value | 0.95 | 0.97 | 0.99 | 0.999 | AUROC | F1 |
|---|---|---|---|---|---|---|
| $\alpha = 0.3$ | 0.0166 | 0.0060 | 0.0002 | 0.0001 | 0.668 | 0.154 |
| $\alpha = 0.5$ | 0.0126 | 0.0053 | 0.0003 | 0.0001 | 0.674 | 0.136 |
| $\alpha = 0.7$ | 0.0103 | 0.0042 | 0.0003 | 0.0001 | 0.679 | 0.105 |

The convergence of all fixed-gating configurations toward nearly identical FPR values at the 0.999 operating point shows that static mixing alone is insufficient to control the extreme tail behavior of anomaly scores while maintaining useful discrimination under injected faults.

#### 6.5.2. Comparison with Dynamic Gating

U-GMM achieves both higher AUROC (0.763) and substantially higher F1-score (0.318) than all fixed-gating configurations, while also maintaining stronger false-positive control. The simultaneous improvement in detection performance and false-positive suppression demonstrates that the benefit of the proposed formulation does not arise merely from introducing a marginal plausibility term, but from adapting the relative contribution of conditional and marginal anomaly evidence according to feature-wise predictive uncertainty.

Dynamic gating allows the relative contribution of conditional and marginal evidence to vary across channels and timesteps. Behavior-sensitive channels with elevated predictive uncertainty are weighted more strongly toward marginal plausibility, while tightly constrained vehicle-state variables retain stronger temporal sensitivity. Static mixing policies cannot represent this context-dependent variability using a single global coefficient.

#### 6.5.3. Interpretation

Taken together, our results show that fixed mixing policies impose a global trade-off between nominal robustness and fault sensitivity, whereas dynamic uncertainty gating permits this trade-off to vary across features and time. The observed performance differences between fixed and dynamic gating therefore reinforce the central design claim of U-GMM: anomaly scoring in mixed-uncertainty telemetry should be adaptive rather than globally regularized through a single static mixture coefficient.

### *6.6. Cross-Dataset Robustness Under Behavioral Variability*

To assess whether the observed HCRL behavior generalizes beyond the primary benchmark, we evaluate U-GMM on the Sonata and OBD datasets. The datasets introduce complementary sources of variability, with Sonata emphasizing inter-driver and route heterogeneity and OBD emphasizing intra-driver contextual variability across multiple routes.

#### 6.6.1. False Positive Rate Stability

Table 5 reports nominal False Positive Rate (FPR) performance on both datasets. Across all datasets and operating points, U-GMM consistently achieves the lowest mean FPR. On Sonata, the improvement is especially pronounced at the 0.999 quantile, where U-GMM reduces FPR to $0.0014 \pm 0.0007$, compared with $0.0183 \pm 0.0333$ for Gaussian-LSTM and $0.0041 \pm 0.0028$ for LSTM-AD (MSE). In addition to the lower mean FPR, U-GMM also exhibits substantially reduced split-to-split variability, indicating improved threshold transfer stability under heterogeneous driving conditions.

Consistent behavior is also observed on OBD, which isolates contextual variability within a single driver across multiple routes. Although the relative gains are smaller than on Sonata, U-GMM again achieves the lowest FPR at both operating points. The improved false-positive control observed on OBD indicates that the proposed scoring formulation is beneficial not only under cross-driver shift, but also under route- and context-induced behavioral variability.

**Table 5.** Cross-dataset False Positive Rate (FPR) comparison on Sonata and OBD (mean ± standard deviation across split realizations).

| Dataset | Model | Quantile | Mean FPR ± s | Gain (×) |
|---|---|---|---|---|
| Sonata | U-GMM | 0.99 | **0.0139 ± 0.0064** | – |
| | Gaussian-LSTM | 0.99 | 0.0346 ± 0.0581 | 2.49 |
| | LSTM-AD (MSE) | 0.99 | 0.0217 ± 0.0143 | 1.56 |
| | U-GMM | 0.999 | **0.0014 ± 0.0007** | – |
| | Gaussian-LSTM | 0.999 | 0.0183 ± 0.0333 | 12.97 |
| | LSTM-AD (MSE) | 0.999 | 0.0041 ± 0.0028 | 2.92 |
| OBD | U-GMM | 0.99 | **0.0070 ± 0.0046** | – |
| | Gaussian-LSTM | 0.99 | 0.0112 ± 0.0070 | 1.60 |
| | LSTM-AD (MSE) | 0.99 | 0.0108 ± 0.0011 | 1.54 |
| | U-GMM | 0.999 | **0.0007 ± 0.0004** | – |
| | Gaussian-LSTM | 0.999 | 0.0021 ± 0.0022 | 3.00 |
| | LSTM-AD (MSE) | 0.999 | 0.0011 ± 0.0002 | 1.57 |

Bold values indicate the lowest false-positive rate (FPR) at each operating point.

#### 6.6.2. Detection Sensitivity

Table 6 summarizes detection performance under injected fault scenarios. The results indicate that U-GMM preserves competitive fault sensitivity while providing substantially stronger nominal robustness. On Sonata, Gaussian-LSTM attains the highest AUROC and F1-score, but U-GMM maintains equally high recall (0.964) under markedly improved false-positive stability. On OBD, U-GMM achieves the highest AUROC (0.893) while maintaining a precision–recall balance comparable to the deterministic baseline.

Taken together, the detection sensitivity results indicate that the benefits of uncertainty-gated scoring are not confined to the primary HCRL benchmark. Across both multi-driver and single-driver settings, U-GMM consistently reduces false positives and maintains useful fault sensitivity, supporting the claim that the proposed formulation improves robustness under diverse forms of behavioral variability.

**Table 6.** Cross-dataset detection performance under injected fault scenarios, averaged across split realizations.

| Dataset | Model | AUROC | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Sonata | U-GMM | 0.745 | 0.214 | 0.964 | 0.315 |
| | Gaussian-LSTM | **0.799** | **0.369** | 0.964 | **0.482** |
| | LSTM-AD (MSE) | 0.709 | 0.176 | **0.999** | 0.268 |
| OBD | U-GMM | **0.893** | **0.194** | 0.631 | 0.272 |
| | Gaussian-LSTM | 0.878 | 0.062 | 0.490 | 0.101 |
| | LSTM-AD (MSE) | 0.890 | 0.186 | **0.842** | **0.275** |

Bold values indicate the best performance achieved for each evaluation metric.

### *6.7. Comparison with Recent Deep Architectures*

To further evaluate the proposed formulation against more recent anomaly detection architectures, we compare U-GMM against representative probabilistic, adversarial, graph-based, and transformer-based methods, including OmniAnomaly [3], USAD [21], MTAD-GAT [24], and TranAD [25]. Unlike the earlier ablative baselines, the compared methods employ substantially more expressive latent-variable, attention-based, graph-structured, and transformer-based sequence modeling mechanisms. The comparison with high-capacity probabilistic, graph-based, adversarial, and transformer-based baselines therefore evaluates whether increased representational capacity alone is sufficient to reduce false-positive inflation under heterogeneous behavioral variability, or whether adaptive uncertainty-aware anomaly scoring remains necessary under mixed-uncertainty telemetry.

Table 7 shows that U-GMM consistently achieves the lowest nominal false-positive rates across all datasets and operating points, despite the substantially greater representational complexity of the compared architectures. The largest differences are observed on the Sonata dataset, where contextual and behavioral heterogeneity are most pronounced. In particular, OmniAnomaly exhibits substantially elevated split-to-split variance and severe threshold instability at the 0.999 operating point, indicating that expressive latent-variable forecasting alone does not prevent extreme anomaly-score inflation under heterogeneous nominal behavior.

The results further demonstrate that improved sequence modeling capacity alone is insufficient to stabilize anomaly-score distributions under mixed behavioral uncertainty. Although MTAD-GAT and TranAD incorporate graph-based and transformer-based temporal representations, respectively, nominal false-positive rates remain consistently elevated relative to U-GMM across all datasets. The consistently elevated nominal false-positive rates observed across the high-capacity probabilistic, adversarial, graph-based, and transformer-based baselines suggest that false-positive inflation arises not solely from limited representational capacity, but from heterogeneous nominal predictability that is not explicitly modeled during anomaly scoring.

**Table 7.** Nominal false-positive rate comparison against recent deep anomaly detection architectures (mean ± standard deviation across split realizations). Lower values indicate stronger nominal robustness.

| Dataset | Quantile | U-GMM | OmniAnomaly | USAD | MTAD-GAT | TranAD |
|---|---|---|---|---|---|---|
| HCRL | 0.99 | **0.00287 ± 0.00254** | 0.00599 ± 0.00227 | 0.00539 ± 0.00472 | 0.00484 ± 0.00420 | 0.00642 ± 0.00281 |
| HCRL | 0.999 | **0.00039 ± 0.00022** | 0.00069 ± 0.00022 | 0.00089 ± 0.00110 | 0.00088 ± 0.00057 | 0.00112 ± 0.00066 |
| Sonata | 0.99 | **0.01390 ± 0.00640** | 0.03611 ± 0.05531 | 0.01720 ± 0.00885 | 0.02403 ± 0.01501 | 0.02303 ± 0.01512 |
| Sonata | 0.999 | **0.00140 ± 0.00070** | 0.01855 ± 0.03429 | 0.00333 ± 0.00345 | 0.00423 ± 0.00309 | 0.00292 ± 0.00204 |
| OBD | 0.99 | **0.00700 ± 0.00460** | 0.01007 ± 0.00376 | 0.01165 ± 0.00249 | 0.01135 ± 0.00172 | 0.01060 ± 0.00126 |
| OBD | 0.999 | **0.00070 ± 0.00040** | 0.00098 ± 0.00031 | 0.00136 ± 0.00045 | 0.00107 ± 0.00022 | 0.00108 ± 0.00013 |

Bold values indicate the lowest false-positive rate (FPR) at each operating point.

Table 8 shows that several modern architectures achieve competitive or stronger synthetic fault-detection sensitivity than U-GMM on certain datasets, particularly in terms of F1-score and recall. OmniAnomaly, for example, achieves higher F1-scores on HCRL and Sonata, reflecting strong latent temporal representation learning. However, the stronger detection sensitivity achieved by some high-capacity architectures occurs alongside substantially elevated nominal false-positive rates and greater threshold instability under heterogeneous behavioral conditions.

**Table 8.** Synthetic fault-detection performance comparison against recent deep anomaly detection architectures. Best results are shown in bold and second-best results are underlined.

| Dataset | Model | AUROC | Precision | Recall | F1 |
|---|---|---|---|---|---|
| HCRL | U-GMM | **0.763** | 0.195 | 0.857 | 0.318 |
| HCRL | OmniAnomaly | <u>0.728</u> | **0.377** | <u>0.870</u> | **0.482** |
| HCRL | USAD | 0.638 | 0.142 | **0.878** | 0.227 |
| HCRL | MTAD-GAT | 0.711 | 0.271 | 0.740 | 0.323 |
| HCRL | TranAD | 0.725 | <u>0.279</u> | 0.815 | <u>0.387</u> |
| Sonata | U-GMM | <u>0.745</u> | <u>0.214</u> | <u>0.964</u> | <u>0.315</u> |
| Sonata | OmniAnomaly | **0.813** | **0.312** | **1.000** | **0.408** |
| Sonata | USAD | 0.697 | 0.091 | **1.000** | 0.146 |
| Sonata | MTAD-GAT | 0.739 | 0.170 | **1.000** | 0.246 |
| Sonata | TranAD | 0.731 | 0.139 | **1.000** | 0.208 |
| OBD | U-GMM | **0.893** | **0.194** | 0.631 | **0.272** |
| OBD | OmniAnomaly | 0.865 | 0.050 | 0.825 | 0.081 |
| OBD | USAD | 0.841 | 0.062 | **0.987** | 0.103 |
| OBD | MTAD-GAT | 0.851 | <u>0.092</u> | 0.790 | <u>0.146</u> |
| OBD | TranAD | <u>0.892</u> | 0.091 | <u>0.859</u> | <u>0.146</u> |

By contrast, U-GMM consistently maintains competitive AUROC and detection sensitivity while substantially improving nominal score stability across all evaluated datasets. The strongest advantage is observed under mixed contextual and behavioral variability, where the proposed uncertainty-aware scoring formulation suppresses spurious anomaly-score inflation without completely sacrificing fault sensitivity.

Taken together, the combined detection and false-positive results suggest that increased sequence-modeling capacity alone is insufficient to resolve false-positive inflation in human-driven telemetry. Although modern architectures improve temporal representation learning and fault discrimination, anomaly evidence remains heavily influenced by heterogeneous nominal predictability. The proposed U-GMM framework instead addresses the problem at the scoring level by adaptively balancing conditional temporal consistency against global marginal plausibility according to feature-wise predictive uncertainty.

## 7. Discussion

The experimental results indicate that the principal benefit of U-GMM lies not simply in increased predictive expressiveness, but in the way anomaly evidence is constructed under heterogeneous uncertainty. Across the primary HCRL benchmark and the auxiliary Sonata and OBD datasets, the proposed framework consistently reduces nominal false positives while preserving useful detection sensitivity. The ablation study further shows that this gain cannot be explained by the introduction of a marginal plausibility term alone: fixed mixing policies improve nominal robustness only at the cost of reduced fault sensitivity, whereas dynamic uncertainty-aware gating achieves a more favorable trade-off.

A broader implication is that anomaly detection in human-driven telemetry should not treat all predictive deviations as equally informative. Driver-controlled channels can remain difficult to predict even under nominal operation because relevant contextual factors are only partially observed. In such settings, increasing model capacity alone is unlikely to fully resolve false-positive behavior, since part of the variability is tied to nominal stochasticity rather than to insufficient representation power. The proposed formulation addresses this issue at the scoring level by adapting the balance between conditional temporal consistency and global plausibility according to feature-wise predictive uncertainty. In this sense, U-GMM should be viewed less as a more powerful forecaster and more as a scoring framework for mixed-uncertainty telemetry.

Although the present study focuses on vehicle telemetry, the proposed uncertainty-aware anomaly-scoring formulation is potentially applicable to other human-machine closed-loop systems in which nominal behavior arises from a mixture of deterministic system dynamics and partially observed human control behavior. Representative examples include drone co-piloting, robotic teleoperation, and assisted maritime navigation, where operator intent, environmental interaction, and autonomous control may jointly influence observed telemetry. In such settings, large predictive deviations do not uniformly indicate abnormality, since part of the variability arises from nominal but difficult-to-predict human interaction. The proposed framework may therefore provide a more robust anomaly-scoring strategy in broader mixed-uncertainty cyber-physical systems.

Evaluation across the three datasets (HCRL, Sonata, and OBD) demonstrates that different forms of nominal variability affect anomaly-detection behavior in distinct ways. In HCRL, cross-driver behavioral variability introduces forecasting deviations when previously unseen driving styles are encountered, causing conventional forecasting-based detectors to assign elevated anomaly scores to otherwise nominal behavior. In the Sonata dataset, the combination of driver, route, and environmental heterogeneity increases contextual uncertainty, particularly in behavior-sensitive telemetry channels influenced by partially unobserved traffic conditions. The resulting context-dependent uncertainty broadens the distribution of nominal anomaly scores and increases sensitivity to benign behavioral deviations in standard sequential models. In the OBD dataset, route-dependent contextual changes alter the temporal predictability of control signals even for the same driver, leading conventional detectors to incorrectly interpret localized predictive deviations as anomalous behavior.

The proposed U-GMM framework mitigates these effects by adaptively balancing conditional temporal forecasting against global marginal plausibility. By reducing over-reliance on uncertain temporal predictions in behavior-sensitive channels, the model suppresses false-positive inflation while maintaining sensitivity to genuinely abnormal vehicle dynamics.

## 8. Limitations and Future Work

While the proposed uncertainty-gated formulation improves robustness under behavioral variability, it does not eliminate the underlying source of unpredictability. The conditional forecasting component may still treat rare but valid behavioral patterns as statistically surprising; the gate primarily attenuates their influence at inference time. The proposed framework therefore manages behavioral uncertainty during anomaly scoring rather than reducing the uncertainty itself at the representation level.

A second limitation is the use of factorized output densities in both components. The conditional forecaster employs a diagonal Gaussian output distribution, while the marginal plausibility model uses independent one-dimensional flow densities for each feature. Although the use of factorized conditional and marginal densities promotes tractability, stability, and interpretability, it neglects structured cross-feature dependence that may be important in tightly coupled vehicle systems.

From a computational perspective, the marginal plausibility component can be more demanding to train than the conditional forecaster, particularly on longer and more diverse recordings. By contrast, the uncertainty-gating mechanism itself introduces relatively little additional overhead, since gating is implemented using a lightweight feature-wise MLP operating on predicted uncertainty estimates. The primary computational cost arises from fitting the independent flow-based marginal densities across large telemetry datasets. Factorized per-feature

marginal density modeling was intentionally adopted to improve computational efficiency and scalability relative to fully multivariate likelihood formulations, although this design sacrifices explicit cross-feature dependence modeling. Similarly, the conditional forecasting component employs a diagonal covariance formulation to maintain computational efficiency during probabilistic forecasting over high-dimensional telemetry signals while still enabling feature-wise uncertainty estimation. Preliminary experiments suggest that partial subsampling of nominal data can substantially reduce computational cost without significantly degrading performance, indicating that the marginal model primarily captures global operating envelopes rather than fine-grained temporal structure. A more systematic study of marginal data efficiency and scalable density modeling remains an important direction for future research.

Although the proposed framework demonstrates substantial false-positive reduction across datasets; the current evaluation focuses primarily on operational regimes represented in the evaluated datasets and synthetic fault scenarios. More extreme out-of-distribution conditions, such as severe weather or significant sensor degradation, were not explicitly evaluated. Nevertheless, because the proposed framework combines conditional temporal consistency with global marginal plausibility, it is expected to remain more robust than purely forecasting-based anomaly detectors under moderate distribution shift. Future work will investigate robustness under more extreme operational conditions and adaptive recalibration strategies for long-term deployment.

The proposed uncertainty-gating mechanism is expected to require retraining or domain adaptation when applied to substantially different human-machine systems. Although the general mixed-uncertainty formulation may transfer across domains, the learned association between predictive uncertainty and feature-wise anomaly relevance depends on the underlying telemetry structure, temporal dependencies, control interactions, and operational context of the target system. Different domains may therefore exhibit substantially different predictability characteristics and uncertainty distributions. Investigating transferability, domain adaptation, and cross-domain calibration of the gating behavior remains an important direction for future work.

Future work may also investigate stronger nominal representation learning for human-driven telemetry, including self-supervised pre-training, domain-adaptive learning, or meta-learning strategies that improve cross-driver invariance. Stronger cross-driver representation learning through self-supervised pre-training, domain adaptation, and meta-learning may reduce reliance on reactive gating and enable the proposed framework to focus more selectively on structurally implausible deviations.

## 9. Conclusions

This paper has addressed the problem of anomaly scoring in human-driven vehicle telemetry, where nominal deviations may arise either from stochastic behavioral variability or from genuinely abnormal departures from learned vehicle dynamics. Conventional residual-based detectors can assign disproportionately large anomaly scores to behavior-sensitive channels when nominal predictability varies substantially across telemetry features and operating contexts, leading to unstable threshold transfer and elevated false-positive rates under cross-driver deployment.

The proposed Uncertainty-Gated Mixture Model (U-GMM) combines conditional probabilistic forecasting with marginal plausibility modeling through a learned feature-wise uncertainty gate. By adapting the relative contribution of conditional temporal consistency and global plausibility according to predictive uncertainty, the proposed framework reduces undue score inflation in nominally stochastic channels while preserving sensitivity to dynamically inconsistent or globally implausible deviations.

Experimental results on multiple real-world vehicle telemetry datasets have shown that U-GMM improves false-positive robustness under cross-driver and cross-context variability while maintaining competitive detection sensitivity under injected fault scenarios. The ablation and interpretability analyses further indicate that the observed gains arise from adaptive uncertainty-gated fusion rather than from probabilistic forecasting alone, and that the learned routing behavior is consistent with heterogeneous feature-wise predictability in vehicle telemetry.

Overall, the results suggest that reliable anomaly detection in human-in-the-loop vehicle systems depends not only on predictive model capacity, but also on how anomaly evidence is constructed under heterogeneous nominal uncertainty. U-GMM should therefore be viewed less as a more powerful forecaster and more as an uncertainty-aware scoring framework for robust anomaly detection in behaviorally variable cyber-physical systems.

## 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 research was funded by the Royal Society of the UK.

## 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 [27], the Sonata dataset is available from IEEE DataPort [39], and the Automotive OBD-II dataset is available from RADAR4KIT [40]. Synthetic fault injections and derived experimental results were generated by the authors as part of this study.

## 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. Pang, G.; Shen, C.; Cao, L.; et al. Deep Learning for Anomaly Detection: A Survey. *ACM Comput. Surv.* **2021**, *54*, 38. https://doi.org/10.1145/3439950.
2. Darban, Z.Z.; Webb, G.I.; Pan, S.; et al. Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Comput. Surv.* **2024**, *57*, 1–42. https://doi.org/10.1145/3691338.
3. 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.
4. 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.
5. 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.
6. Muenchhof, M.; Beck, M.; Isermann, R. Fault Diagnosis and Fault Tolerance of Drive Systems: Status and Research. *Eur. J. Control* **2009**, *15*, 370–388.
7. Isermann, R. Model-Based Fault-Detection and Diagnosis: Status and Applications. *Annu. Rev. Control* **2005**, *29*, 71–85.
8. 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.
9. Zhang, C.; He, Z.; Wu, C.; et al. When Context Is Not Enough: Modeling Unexplained Variability in Car-Following Behavior. *arXiv* **2025**, arXiv:2507.07012.
10. 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.
11. 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.
12. 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, Padua Italy, 30 September–2 October 2024; pp. 531–545.
13. Nunes, P.; Santos, J.; Rocha, E. Challenges in Predictive Maintenance: A Review. *CIRP J. Manuf. Sci. Technol.* **2023**, *40*, 53–67.
14. Song, X.; Wu, M.; Jermaine, C.; et al. Conditional Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 631–645.
15. Li, Z.; van Leeuwen, M. Explainable Contextual Anomaly Detection Using Quantile Regression Forests. *Data Min. Knowl. Discov.* **2023**, *37*, 2517–2563.
16. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv.* **2009**, *41*, 1–58.

17. Zimek, A.; Schubert, E.; Kriegel, H.P. A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Stat. Anal. Data Min.* **2012**, *5*, 363–387.
18. Aggarwal, C.C.; Yu, P.S. Outlier Detection for High Dimensional Data. In Proceedings of the 2001 ACM SIGMOD international conference on Management of Data, Santa Barbara, CA, USA, 21–24 May 2001; pp. 37–46.
19. Malhotra, P.; Ramakrishnan, A.; Anand, G.; et al. LSTM-Based Encoder-Decoder for Multi-Sensor Anomaly Detection. *arXiv* **2016**, arXiv:1607.00148.
20. Munir, M.; Siddiqui, S.A.; Dengel, A.; et al. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access* **2019**, *7*, 1991–2005.
21. 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, online, 6–10 July 2020; pp. 3395–3404.
22. Ruff, L.; Vandermeulen, R.; Goernitz, N.; et al. Deep One-Class Classification. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 4393–4402.
23. Xu, J.; Wu, H.; Wang, J.; et al. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), online, 25–29 April 2022.
24. Zhao, H.; Wang, Y.; Duan, J.; et al. Multivariate Time-Series Anomaly Detection via Graph Attention Network. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 17–20 November 2020; pp. 841–850.
25. 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.
26. Hayes, M.A.; Capretz, M.A.M. Contextual Anomaly Detection Framework for Big Sensor Data. *J. Big Data* **2015**, *2*, 2.
27. Kwak, B.I.; Woo, J.; Kim, H.K. Know Your Master: Driver Profiling-Based Anti-Theft Method. In Proceedings of the 2016 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, New Zealand, 12–14 December 2016; pp. 211–218.
28. Hallac, D.; Sharang, A.; Stahlmann, R.; et al. Driver Identification Using Automobile Sensor Data From a Single Turn. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 953–958.
29. Fugiglando, U.; Massaro, E.; Santi, P.; et al. Driving Behavior Analysis Through CAN Bus Data in an Uncontrolled Environment. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 737–748.
30. Marchegiani, L.; Posner, I. Long-Term Driving Behaviour Modelling for Driver Identification. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 913–919.
31. Seraji, M.H.M.; Haghshenas, S.S.; Haghshenas, S.S.; et al. A State-of-the-Art Review on Machine Learning Techniques for Driving Behavior Analysis: Clustering and Classification Approaches. *Complex Intell. Syst.* **2025**, *11*, 386.
32. Xu, B.; Zhu, Q. Dynamic Probabilistic Latent Variable Model with Exogenous Variables for Dynamic Anomaly Detection. In Proceedings of the 2023 American Control Conference (ACC), San Diego, CA, USA, 31 May–2 June 2023; pp. 3945–3950.
33. Usmani, U.A.; Aziz, I.A.; Jaafar, J.; et al. Deep Learning for Anomaly Detection in Time-Series Data: An Analysis of Techniques, Review of Applications, and Guidelines for Future Research. *IEEE Access* **2024**, *12*, 174564–174590.
34. Schölkopf, B.; Locatello, F.; Bauer, S.; et al. Toward Causal Representation Learning. *Proc. IEEE* **2021**, *109*, 612–634.
35. Bilal, H.; Rehman, A.; Aslam, M.S.; et al. Hybrid TrafficAI: A Generative AI Framework for Real-Time Traffic Simulation and Adaptive Behavior Modeling. *IEEE Trans. Intell. Transp. Syst.* **2025**, 1–17.
36. Ullah, I.; Khalil, I.; Bai, X.; et al. An Ensemble-Based Hybrid Model for the Detection of Attacks in the Internet of Vehicular Things. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 17914–17927.
37. Wirnsberger, P.; Papamakarios, G.; Ibarz, B.; et al. Normalizing Flows for Atomic Solids. *Mach. Learn. Sci. Technol.* **2022**, *3*, 025009.
38. Durkan, C.; Bekasov, A.; Murray, I.; et al. Neural Spline Flows. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 7511–7522.
39. Park, K.H.; Kwak, B.I.; Kim, H.K. This Car Is Mine!: Driver Pattern Dataset Extracted from CAN-Bus. Available online: https://ieee-dataport.org/open-access/car-mine-driver-pattern-dataset-extracted-can-bus (accessed on 22 December 2025).
40. Weber, M. Automotive OBD-II Dataset. Available online: https://doi.org/10.35097/1130 (accessed on 16 November 2025).