OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 04 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

TELEMETRY PRIMACY UNDER REGIME SHIFT: A CONTROLLED BENCHMARK OF COMMUNICATION-DEGRADATION EARLY WARNING FOR AUTONOMOUS AERIAL LINKS

***Fatima Ubaid**

Department of Engineering and Information Technology Foundation University, Rawalpindi, Punjab, Pakistan.

***Corresponding Author:** (13.cs.12@gmail.com)

DOI:(*https://doi.org/10.71146/kjmr899*)

Article Info

Abstract

Autonomous aerial systems rely on stable wireless links for control, telemetry exchange, and mission continuity. In practice, these links can degrade under changing load, interference, and difficult operating conditions, which makes early warning important. This paper studies communication-degradation early warning using a controlled 6G telemetry benchmark, while treating UAV-assisted and autonomous aerial links as the motivating application rather than as a direct data source. We formulate the task as next-interval risk prediction from recent telemetry and client-profile information, and evaluate generalization with a leave-one-regime-out protocol across four operating regimes. To test whether added temporal complexity is truly useful, we compare strong aggregate baselines with temporal and hybrid models. The results show that simple aggregate telemetry models remain the strongest overall. In particular, the logistic-regression aggregate baseline provides the strongest overall trade-off, reaching a mean AUROC of 0.876 and a mean balanced accuracy of 0.691 under regime shift. Hybrid stacking provides only a limited gain in worst-regime F1, improving it to 0.509, while temporal residual modeling alone is unreliable. These findings suggest that communication-degradation early warning under unseen conditions remains difficult, and that strong aggregate telemetry summaries are harder to beat than expected in this benchmark. The study provides a careful and deployment-aware reference point for future work on reliable early warning in communication-sensitive autonomous systems.

Keywords: *Communication degradation early warning; regime shift; aggregate telemetry baselines; hybrid stacking; autonomous aerial links.*

1. Introduction

Autonomous aerial systems depend on stable wireless links for control, telemetry exchange, and mission continuity. If those links weaken at the wrong time, the consequences can be immediate: delayed commands, lost situational awareness, degraded service, or even mission failure. In real settings, communication quality can change because of interference, congestion, mobility, variable channel conditions, or other disruptive operating contexts. For that reason, early warning is important. A system that can signal rising communication risk before failure becomes obvious may support safer monitoring, better operator response, and more reliable autonomy [1]-[4].

Recent work in 5G, 6G, and O-RAN has shown that network telemetry can support predictive management. KPI-based and QoS-oriented models have been used for service assurance, performance estimation, and fault monitoring, while O-RAN xApps have expanded interest in lightweight telemetry-driven analytics for near-real-time observation and control [1], [2], [5], [6]. These studies make an important point clear: modern wireless systems already produce the signals needed for proactive monitoring. However, most prior work has focused on terrestrial communication settings, known operating conditions, or average predictive performance. Much less is known about how well such models behave when the operating regime changes and the test condition is not seen during training [2], [5], [7].

A related body of work appears in UAV and autonomous aerial communication security. There, the focus is often on jamming, spoofing, intrusion detection, GNSS manipulation, or sensor-specific anomaly detection [8], [9]. This literature is important because it shows that communication disruption is not a theoretical concern; it is a practical safety issue for aerial systems. At the same time, much of that work depends on attack-specific traces, flight data, RF measurements, or navigation signals. That makes it difficult to separate a broader early-warning question from a narrower attack-identification problem. In many real settings, operators may first need a simple and reliable warning that communication quality is degrading, even before they know why [8], [9].

This paper studies that earlier and more practical question: **Can structured wireless telemetry support early warning of communication degradation under unseen operating regimes?** We examine this question using a controlled 6G telemetry benchmark with four regimes and client-level heterogeneity. Our task is framed as next-interval communication-risk prediction from recent telemetry and profile information, and evaluation is performed with a leave-one-regime-out protocol. This setup allows us to test whether a model trained on seen regimes can still provide useful warning signals when the environment changes [10], [11].

The study is also motivated by a practical modeling concern. Strong aggregate telemetry summaries are often simple, computationally light, and stable, while temporal and hybrid models offer richer context at the cost of added complexity. It is not obvious that this extra complexity is always helpful under regime shift. We therefore compare aggregate, temporal, and hybrid models under the same benchmark and evaluate not only ranking quality, but also balanced decisions, calibration, and low-budget warning capture. This follows a reliability-first view that has become

increasingly important in operational machine learning, where good ranking alone is not enough if probabilities are not trustworthy or warning decisions are hard to audit [12], [13].

It is important to state clearly what this paper does not claim. This is not a paper on attack attribution. It is not a direct anti-jamming or anti-spoofing detector for UAVs. It does not use UAV-native flight logs or RF traces. Instead, it uses structured 6G RAN telemetry as a controlled benchmark and treats autonomous aerial links as the motivating application domain. In that sense, the contribution is careful by design: we study a defensive early-warning problem that is relevant to communication-sensitive autonomous systems, without overstating what the dataset directly represents [8], [10].

The contributions of this paper are threefold. First, we conduct a controlled benchmark study that reframes next-interval risk prediction as communication-degradation early warning under regime shift. Second, we test whether temporal residual modeling and hybrid fusion add value beyond strong aggregate telemetry baselines under leave-one-regime-out evaluation. Third, we show that aggregate baselines remain the strongest overall, while hybrid stacking adds only limited residual value, mainly in worst-regime F1. These findings provide a practical reference point for future work on reliable early warning in communication-sensitive autonomous systems.

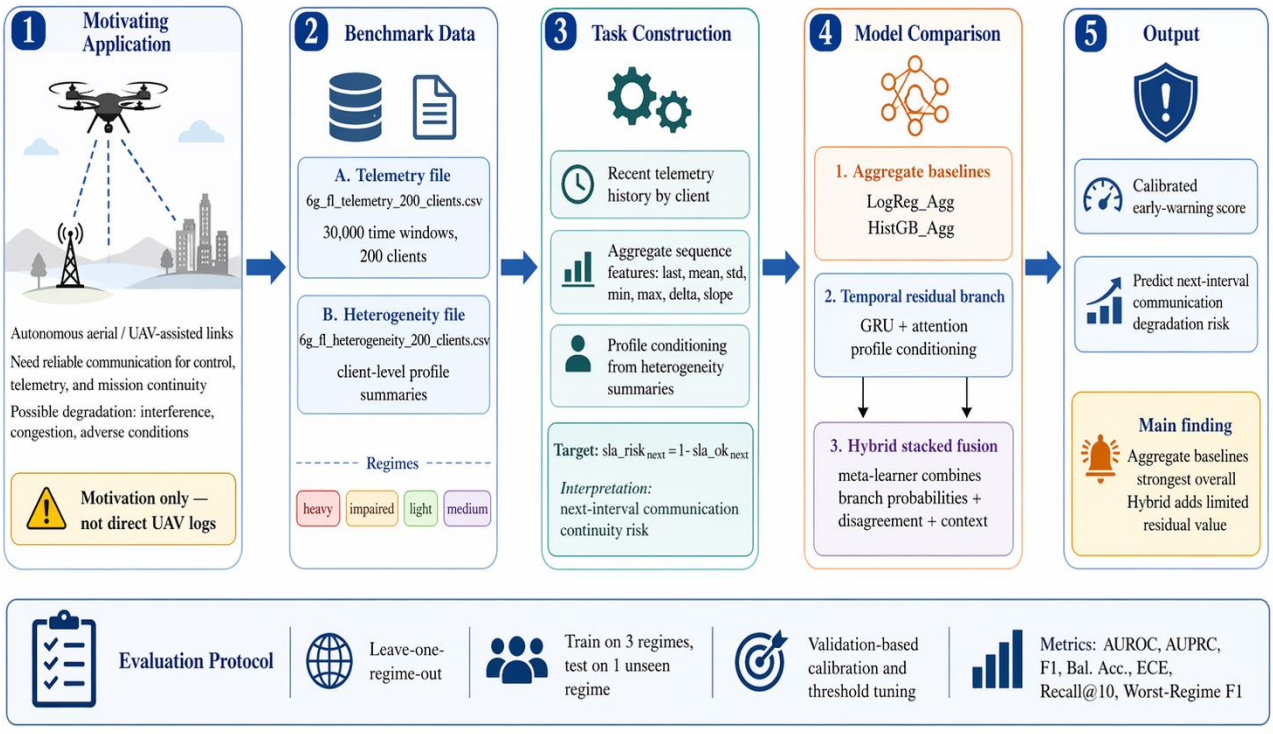

Fig. 1. System overview of the proposed benchmark for communication-degradation early warning, from motivating aerial-link use, benchmark construction, and model comparison to deployment-aware evaluation.

2. Related Work

2.1 Telemetry-Driven Network Prediction

Telemetry-driven prediction has become a central theme in intelligent wireless networking. In 5G, B5G, and 6G systems, KPIs such as latency, throughput, reliability, and channel quality are increasingly used to support proactive monitoring and automated control. Recent surveys show that performance evaluation and optimization in 6G are moving toward AI-assisted, cross-layer, and adaptive frameworks, but they also note that many current approaches remain simulation-heavy and only loosely aligned with deployment constraints [2]. In parallel, O-RAN research has highlighted the growing role of xApps and telemetry pipelines in near-real-time monitoring, analytics, and policy support [1], [6]. These trends make telemetry-based early warning a timely and relevant problem.

QoS and SLA forecasting form a closely related line of work. Prior studies have shown that machine learning can estimate QoS or service outcomes from network measurements, and federated formulations have also been explored in cellular vehicular settings where raw measurements are distributed and direct data sharing is undesirable [5], [14]. More recently, work on large-scale 6G telemetry has examined federated learning under client heterogeneity, focusing on communication cost, fairness, and tail-client behavior across many clients [10]. These studies confirm that wireless telemetry contains useful predictive signal. However, they mostly emphasize federated optimization, average predictive performance, or communication-efficiency trade-offs. They do not directly answer whether simple aggregate telemetry models, temporal models, or hybrid models are best suited for early warning under unseen regimes.

Another useful lesson comes from benchmark-oriented systems work outside wireless prediction. Recent evaluation studies in other domains have shown that strong, well-defined baselines and careful workload design often matter as much as model novelty. In particular, reproducible benchmarking work has stressed that fair comparison requires matched evaluation settings, correct ground truth, and metrics that reflect practical cost rather than average performance alone [13]. That lesson is directly relevant here, because early-warning models should be judged not only by ranking scores, but also by calibration, decision quality, and low-budget warning capture.

2.2 UAV and Autonomous Aerial Communication Reliability

The UAV literature makes clear that communication reliability is a real operational concern. Aerial systems may face communication degradation because of interference, spoofing, jamming, adverse channel conditions, or unstable navigation support [8], [9]. Reviews of UAV security and navigation resilience repeatedly identify jamming and spoofing as major threats and note that machine learning may help, especially when lightweight monitoring is required [8]. This body of work is important because it provides the application motivation for studying early warning in communication-sensitive autonomous systems.

At the same time, much of this literature is attack-specific or sensor-specific. Many studies rely on GNSS traces, inertial measurements, RF fingerprints, or network intrusion features. Those settings are valuable for threat-specific defenses, but they do not directly answer a broader systems

question: can a wireless link be flagged as risky early enough using only lightweight telemetry summaries? From a deployment perspective, that question matters because operators may first need dependable warning signals before moving to heavier or more specialized analysis. Our paper therefore stays at the defensive reliability layer. It asks whether structured wireless telemetry can support early warning of degraded communication states, not whether the root cause can be identified or attributed [8], [9].

This distinction is especially important for honest framing. The benchmark used here is not a UAV flight or RF dataset. It is a structured 6G telemetry dataset with multiple regimes and heterogeneous clients [10]. We therefore use UAV-assisted and autonomous aerial communication as a motivation domain, not as a direct observational claim. This choice keeps the problem practical while avoiding overstatement.

2.3 Distribution Shift, Calibration, and Robust Early Warning

A second relevant line of work concerns robustness under changing conditions. Models that perform well on random or in-distribution splits often degrade when the operating environment changes. This issue appears in many settings under names such as distribution shift, out-of-distribution generalization, and concept drift. Recent surveys in federated and dynamic learning confirm that shift and drift remain central obstacles to reliable deployment, especially when data are distributed, non-stationary, or both [7], [15], [16]. In wireless settings, this challenge is amplified by client heterogeneity, temporal variability, and unequal regime difficulty [10], [11].

For early warning, robustness is only part of the problem. Probability quality also matters. In operational systems, scores are often treated as risk estimates, and decisions are made by thresholding those scores. If the probabilities are poorly calibrated, the warning system may appear confident for the wrong reasons. Reliability-focused work has shown that post-hoc calibration, Brier score, Expected Calibration Error, and reliability diagrams provide useful checks on whether a model's scores can be trusted in threshold-based settings [12]. Although that work comes from a different application area, the operational lesson carries over directly: high ranking performance alone is not enough when the real task is auditable warning and intervention.

These observations motivate the gap addressed in this paper. There is still limited controlled evidence on whether lightweight sequential wireless telemetry, without UAV-native flight or RF traces, can support early warning of communication degradation under unseen operating regimes, and whether temporal or hybrid models truly outperform strong aggregate baselines in this setting. Our study addresses this gap through a controlled leave-one-regime-out benchmark that compares aggregate, temporal, and hybrid models using both predictive and operational metrics.

3. Dataset and Task Formulation

This study uses a structured 6G radio access network telemetry dataset designed for controlled machine learning experiments on distributed and heterogeneous clients. The dataset contains two CSV files. The first file, 6g_fl_telemetry_200_clients.csv, provides time-windowed telemetry observations. The second file, 6g_fl_heterogeneity_200_clients.csv, provides client-level summary information that reflects longer-term differences across clients. The dataset covers **200**

clients and approximately **30,000 rows** of telemetry data, which makes it suitable for studying communication risk prediction under heterogeneous operating conditions. The same dataset has also been used in recent 6G learning studies, which supports its value as a benchmark-style resource rather than a raw deployment log.

Each telemetry row corresponds to one client observed at one indexed time window. The telemetry file includes a client identifier, a time index, regime labels, and multiple numerical features describing recent network behavior. In the current benchmark, the operating regimes are **heavy**, **impaired**, **light**, and **medium**. These regimes are important because they allow the data to be organized into distinct operating conditions rather than treated as one mixed distribution. This makes the dataset useful for studying generalization under shift, which is central to the present paper.

The heterogeneity file complements the telemetry file by summarizing client-level characteristics. These profile variables are not short-term measurements from a single window. Instead, they capture broader client differences that may reflect differences in traffic behavior, stability, or service patterns. In practical terms, this file helps separate two kinds of signal in the benchmark: recent temporal observations from the telemetry stream and slower client-level differences from the profile summaries. That distinction is useful when comparing aggregate, temporal, and hybrid models.

The original supervised label in the telemetry file is sla_ok_next, which indicates whether the service remains acceptable in the next time interval. In this paper, we convert that label into a risk-oriented target:

$$\texttt{sla\_risk\_next} = 1 - \texttt{sla\_ok\_next}.$$

This transformation makes the prediction task easier to interpret from an operational perspective. **We interpret next-window SLA risk as a proxy for communication continuity risk.** In other words, the target is not treated as a direct indicator of a specific attack or failure mechanism. It is used as a practical warning label for whether communication quality is likely to degrade in the next interval. This framing is more suitable for an early-warning study than the original service-success label alone.

Formally, let $x_t^{(c)} \in \mathbb{R}^d$ denote the telemetry feature vector for client c at time window t, and let $p^{(c)} \in \mathbb{R}^m$ denote the client-profile vector from the heterogeneity file. For a sequence length L, the input to the predictor is the recent telemetry history

$$X_t^{(c)} = [x_{t-L+1}^{(c)}, x_{t-L+2}^{(c)}, \ldots, x_t^{(c)}],$$

Together with the corresponding client-profile information p(c)p^{(c)} p(c). The prediction target is the binary label

$$y_t^{(c)} = \texttt{sla\_risk\_next} \in \{0, 1\},$$

where $y_t^{(c)} = 1$ denotes communication-degradation risk in the next interval and $y_t^{(c)} = 0$ denotes no such risk. The learning problem is therefore to estimate

$$\hat{y}_t^{(c)} = f\left(X_t^{(c)}, p^{(c)}\right),$$

where $f(\cdot)$ may be an aggregate, temporal, or hybrid model.

A key feature of the dataset is that the regimes are not equally difficult. The benchmark shows visible differences in both sample counts and risk rates across regimes. For this reason, a simple average over all samples can hide important weaknesses. **Figure 2** therefore presents the regime risk profile, including both regime size and mean risk rate. This figure is important because it makes two things visible at once: the dataset is imbalanced across regimes, and the rate of risk is not uniform across operating conditions. These differences directly motivate the use of regime-aware evaluation later in the paper.

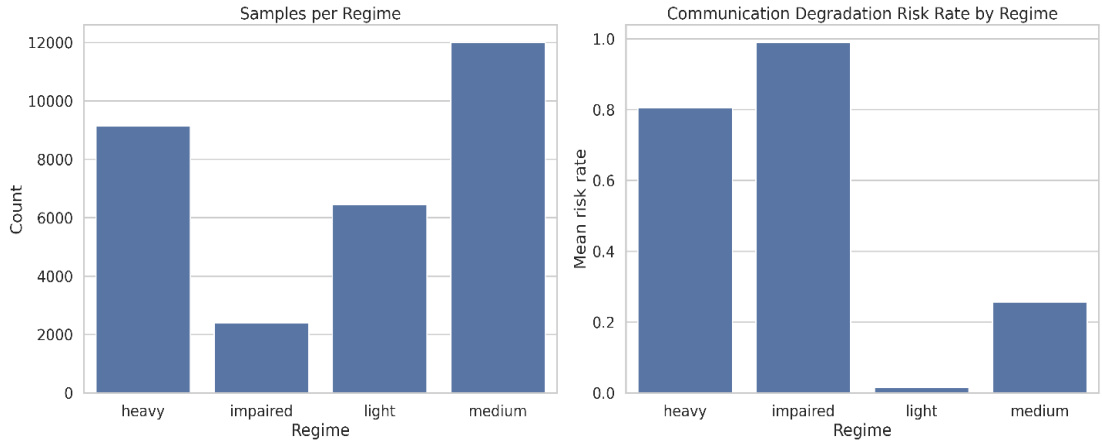

Fig. 2. Regime-wise sample counts and mean communication-degradation risk rates. The benchmark is imbalanced across regimes, and risk prevalence differs substantially by operating condition.

An honest limitation should also be stated here. **The data are structured 6G RAN telemetry rather than UAV-native communication logs, so the aerial-link connection is motivational rather than directly observational.** We use autonomous aerial and UAV-assisted communication as the motivating deployment context because such systems depend on stable links and may benefit from lightweight early warning. However, the present benchmark should be read as a controlled proxy study, not as direct evidence from aerial flight logs or RF attack traces. This distinction is important for keeping the claims of the paper careful and realistic.

Overall, this dataset provides a suitable testbed for the question studied in this paper: whether recent wireless telemetry, together with client-level heterogeneity information, can support early warning of communication degradation under unseen operating regimes. It is large enough to support systematic comparison across model families, yet structured enough to keep the evaluation controlled and reproducible.

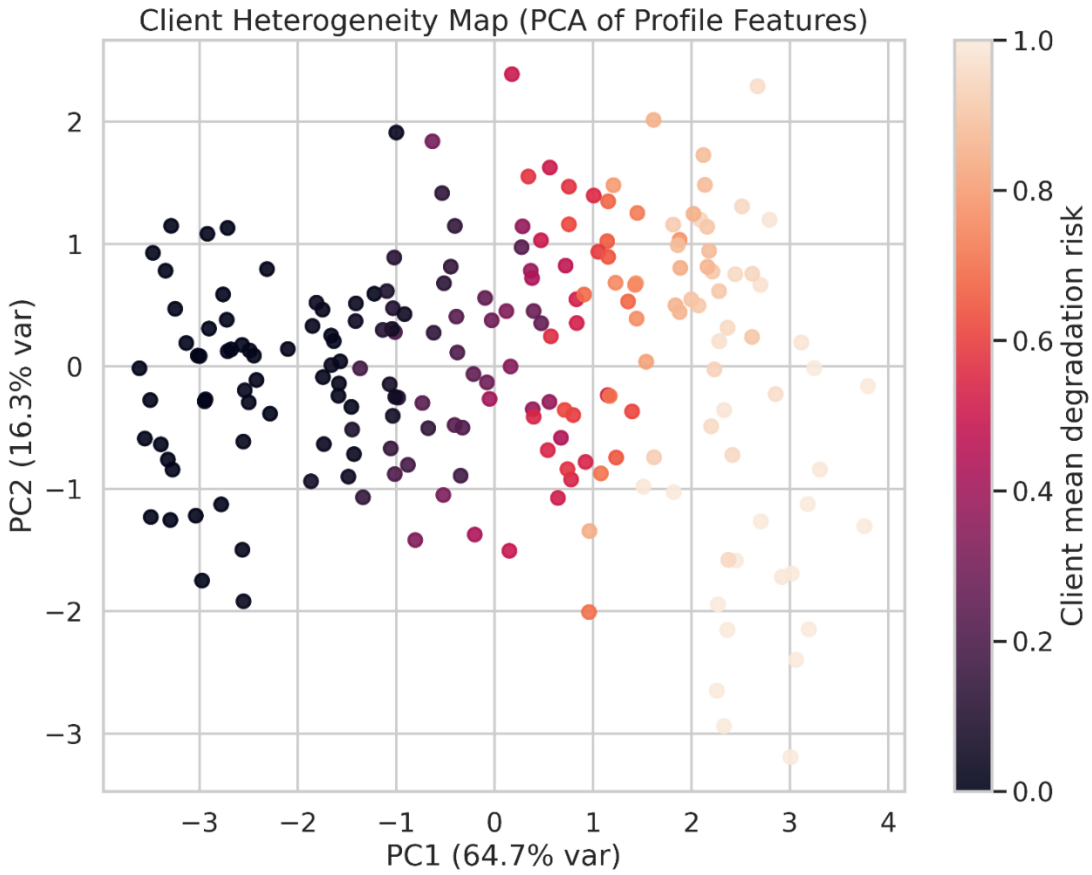

Fig. 3. PCA view of client-level heterogeneity profiles. The smooth risk gradient suggests that client summaries capture meaningful variation linked to communication-degradation risk.

4. Experimental Protocol

The goal of the evaluation is to test whether communication-degradation early warning remains reliable when the operating condition changes. For this reason, we do not use a random train-test split across the full dataset. Instead, we adopt a **leave-one-regime-out** protocol. In each fold, the model is trained on data from three regimes and tested on the remaining held-out regime. The four folds therefore correspond to holding out **heavy**, **impaired**, **light**, and **medium** in turn. This design is stricter than a mixed random split because it measures how well a model transfers to an operating condition that was not seen during training. Such evaluation is more appropriate for deployment-oriented early warning, where future conditions may differ from those observed during model fitting.

Within each fold, the data from the three seen regimes are divided into a training subset and a validation subset. The validation split is created only from the seen-regime portion, so the held-out regime remains completely untouched until final testing. This separation is important because model selection, threshold selection, and probability adjustment should not use any information from the unseen test regime. In practice, this setup gives a clearer estimate of whether the warning system can remain useful after an operating shift.

Sequence construction is also handled fold by fold. Since the task uses recent telemetry history, each example is formed from a fixed number of past windows for the same client, together with the client-profile vector from the heterogeneity file. We use **fold-specific sequence lengths** rather than one global value for all folds. This choice reflects the fact that regime difficulty is not uniform, and the amount of useful temporal context may differ across held-out conditions. In the final protocol, shorter histories are used for some folds, while the most difficult fold is allowed a longer history so that the evaluation is not biased by an unnecessarily restrictive temporal window.

Model development is performed in three stages. First, models are fitted on the training subset of the seen regimes. Second, the validation subset is used for score adjustment and threshold selection. Third, the final locked model is evaluated on the held-out regime. Probability calibration is included because early warning is not only a ranking problem. In an operational setting, predicted scores are often treated as risk estimates, so poorly calibrated outputs can lead to unstable or misleading warning decisions. For that reason, validation predictions are used to calibrate model outputs before final testing, and the same validation subset is used to select the decision threshold. This threshold is not chosen from the test regime. It is tuned only on seen-regime validation data, which makes the reported decisions more realistic and less optimistic. The reliability-first logic of this protocol is consistent with prior work that emphasizes calibration and operationally meaningful evaluation rather than ranking quality alone.

The evaluation compares aggregate, temporal, and hybrid models under exactly the same regime split. This is important because the purpose of the benchmark is not to maximize one headline score at any cost, but to test whether extra temporal or fusion complexity adds value beyond strong aggregate baselines. The protocol therefore treats all model families under one common setup: the same train-validation-test partitioning, the same held-out regime, and the same metric suite.

To reflect both predictive quality and practical warning utility, we report a broad set of metrics. Standard ranking quality is measured with **AUROC** and **AUPRC**. Thresholded decision quality is measured with **F1**, **balanced accuracy**, **precision**, and **recall**. Probability quality is measured with **Brier score** and **Expected Calibration Error (ECE)**. These metrics serve different purposes. AUROC and AUPRC show how well risky examples are ranked. F1, balanced accuracy, precision, and recall show whether the thresholder warnings are usable in practice. Brier and ECE show whether the predicted probabilities behave like trustworthy risk scores rather than arbitrary confidence values.

Because the paper is framed as an early-warning study, we also report **budget-based warning metrics**. In many real settings, operators cannot inspect every high-risk interval. They may only review a small fraction of the highest-risk cases. To reflect this constraint, we report **Recall@5**, **Recall@10**, and **Recall@20**, together with **Precision@5**, **Precision@10**, and **Precision@20**. These metrics measure how many true degradation events are captured if only the top 5%, 10%, or 20% highest-risk windows are reviewed. This view is more operational than using average metrics alone, because it reflects low-budget warning triage rather than full-coverage classification. The same logic is common in other deployment-oriented evaluation settings where the real question is not only whether a system ranks well, but whether it helps under limited review capacity.

In addition to average results across folds, we report **worst-regime F1** and **worst-regime Recall@10**. These summary measures are important because a model with good overall averages may still fail on the most difficult held-out regime. For early warning in communication-sensitive systems, such failures matter. A practical system should not be judged only by mean performance if it collapses under the very condition where early warning is needed most. Reporting the worst held-out regime therefore gives a stricter and more deployment-aware view of robustness.

Overall, this protocol is designed to answer a practical question rather than a narrow modeling question. It asks whether a warning model trained on seen operating conditions can remain reliable, calibrated, and operationally useful when the communication regime changes. By combining leave-one-regime-out evaluation, fold-specific sequence design, validation-based calibration and threshold tuning, and budget-aware warning metrics, the protocol provides a more realistic test of early-warning performance than a standard random split would provide.

5. Methods

This section describes the three model families used in the benchmark: aggregate baselines, a compact temporal residual branch, and a hybrid stacked fusion model. The goal is not to present a new state-of-the-art architecture. Instead, the design is intentionally comparative. We test whether added temporal and fusion complexity provide useful gains beyond already strong aggregate telemetry baselines.

5.1 Aggregate Baselines

We begin with two aggregate baselines: **LogReg_Agg** and **HistGB_Agg**. These models do not operate directly on the raw telemetry sequence. Instead, they use summary features derived from the recent telemetry window for each sample, together with the client-profile features from the heterogeneity table.

For a telemetry sequence of length LLL, we compute a set of aggregate descriptors for each telemetry variable. These descriptors include the **last observed value**, **mean**, **standard deviation**, **minimum**, **maximum**, **delta**, and **slope** across the sequence. Here, delta is the difference between the last and first value in the window, while slope is a simple rate-of-change proxy obtained by dividing that difference by the number of elapsed steps. These sequence summaries are then concatenated with the client-profile vector. The resulting representation provides a compact view of recent state, variability, direction of change, and longer-term client characteristics.

The first aggregate baseline, **LogReg_Agg**, uses logistic regression with class balancing. This model is simple, transparent, and computationally light. It also provides a useful reference for whether a linear decision rule over carefully designed telemetry summaries is already sufficient for the task. The second aggregate baseline, **HistGB_Agg**, uses histogram-based gradient boosting. This model is more flexible and can capture nonlinear interactions between summary features without requiring a large neural architecture. Together, these two baselines provide strong and complementary non-temporal references.

5.2 Temporal Residual Branch

The temporal model is designed as a **small residual branch** rather than as a large standalone sequence architecture. It uses a **GRU encoder**, an **attention pooling layer**, and **profile conditioning** through the client-level heterogeneity features.

Given a sequence of recent telemetry windows, the GRU produces a hidden representation for each time step. Attention pooling then assigns a learned importance weight to each hidden state and forms a weighted sequence summary. This helps the model focus on the most informative parts of the recent telemetry history rather than relying only on the final hidden state. In parallel, the client-profile vector is passed through a small multilayer projection layer. The pooled temporal representation and the profile representation are then concatenated and passed to a prediction head to produce the risk score.

This temporal branch is deliberately compact. The reason is methodological rather than computational. In this paper, the temporal model is not intended to dominate the benchmark through scale alone. It is included to test whether short-horizon temporal structure contributes useful residual signal beyond what is already captured by strong aggregate telemetry summaries. In that sense, it serves as a focused test of incremental temporal value rather than as a claim that deeper sequence modeling is always preferable.

5.3 Hybrid Stacked Fusion

The third model family is a **hybrid stacked fusion model**, denoted **Hybrid Stacked**. This model does not replace the aggregate or temporal branches. Instead, it combines them through a lightweight meta-learner.

The fusion process is based on the outputs of the base branches on the validation set. Specifically, the meta-learner receives three types of input. First, it uses the **branch probabilities** produced by the aggregate and temporal models. Second, it uses **disagreement features**, such as the absolute differences between branch probabilities, which indicate where the branches diverge in confidence. Third, it uses a small set of **simple context features**, including basic statistics derived from the recent telemetry window and profile information. These context features allow the fusion layer to condition its decision on local signal strength rather than relying only on raw branch scores.

The meta-learner itself is intentionally simple. It is trained on validation outputs and then applied to the held-out regime without access to test labels. This helps keep the comparison controlled and reduces the chance that the fusion model appears stronger simply because of excess flexibility.

The hybrid model is designed to test whether temporal residual signals add incremental value on top of already strong aggregate branches. This is an important distinction. The purpose of the hybrid is not to claim architectural novelty. Its purpose is to answer a practical question: if strong aggregate baselines are already effective, does adding a temporal branch and a small fusion layer meaningfully improve early warning under regime shift?

5.4 Calibration and Decision Thresholding

All model families produce risk scores that are further adjusted using validation data before final testing. This step is included because the task is not only to rank risky intervals, but also to support threshold-based warning decisions. Calibration is therefore applied on the validation subset, and decision thresholds are selected using validation performance rather than test outcomes. This prevents information leakage from the held-out regime and makes the final evaluation more realistic.

5.5 Methodological Positioning

Taken together, the three model families represent increasing levels of complexity: aggregate, temporal, and hybrid. This progression is deliberate. It allows the benchmark to test whether stronger temporal modeling and fusion are actually useful in this setting, rather than assuming that added complexity will automatically improve communication-degradation early warning.

6. Results and Discussion

This section reports the main findings of the benchmark and explains what they mean for communication-degradation early warning under regime shift. Overall, the results do not support the idea that added temporal or hybrid complexity automatically improves performance in this setting. Instead, the strongest and most stable results come from the aggregate telemetry baselines. At the same time, the hybrid model shows that a small amount of residual value can still be extracted from temporal signals, especially in the most difficult regime. The main outcome, therefore, is not an architectural breakthrough but a clearer understanding of where the usable predictive signal lies in this benchmark.

6.1 Main Comparison

Table 1 reports the mean results across the four leave-one-regime-out folds. Among all compared models, **LogReg_Agg** gives the strongest overall performance. It achieves a mean **AUROC of 0.876**, **AUPRC of 0.728**, **F1 of 0.748**, and **balanced accuracy of 0.691**. It also provides the best low-budget early-warning capture, with **Recall@10 of 0.323**, and the most favorable calibration among the top models, with **ECE of 0.036**. These results make it the most reliable overall choice in this benchmark.

Model	AUROC	AUPRC	F1	Bal. Acc.	Recall	ECE ↓	Recall@10	Worst-Regime F1	Worst-Regime Bal. Acc.
HistGB_Agg	0.864	0.718	0.741	0.638	0.785	0.044	0.299	0.484	0.500
Hybrid Stacked	0.834	0.720	0.749	0.651	0.850	0.065	0.289	0.509	0.500
LogReg_Agg	0.876	0.728	0.748	0.691	0.840	0.036	0.323	0.505	0.581
Temporal Residual	0.554	0.571	0.333	0.500	0.750	0.189	0.155	0.000	0.500

Table 1. Mean and worst-regime performance of aggregate, temporal, and hybrid models under leave-one-regime-out evaluation.

The **Hybrid Stacked** model produces a slightly different pattern. Its mean **F1 reaches 0.749**, which is marginally higher than that of LogReg_Agg, and its mean **recall reaches 0.850**, which is also slightly higher than the aggregate baseline. However, these gains are narrow. The hybrid model falls behind LogReg_Agg on **AUROC (0.834 vs. 0.876)**, **AUPRC (0.720 vs. 0.728)**, **balanced accuracy (0.651 vs. 0.691)**, **Recall@10 (0.289 vs. 0.323)**, and **ECE (0.065 vs. 0.036)**. In practical terms, the hybrid model improves sensitivity slightly, but it does not produce a decisive overall win.

The **HistGB_Agg** baseline also remains strong. Although it does not lead the table, it stays competitive across most metrics, with **AUROC of 0.864**, **F1 of 0.741**, and **ECE of 0.044**. This consistency reinforces the same conclusion: in this benchmark, aggregate telemetry summaries already carry a large share of the predictive signal.

By contrast, the **Temporal Residual** branch performs poorly when used on its own. Its mean **AUROC falls to 0.554**, **F1 to 0.333**, and **ECE rises to 0.189**. This large performance gap suggests that a small temporal branch alone is insufficient to provide robust warning decisions in this setting.

Aggregate telemetry baselines remained the most competitive overall, indicating that strong summary statistics already capture much of the predictive signal in this benchmark.

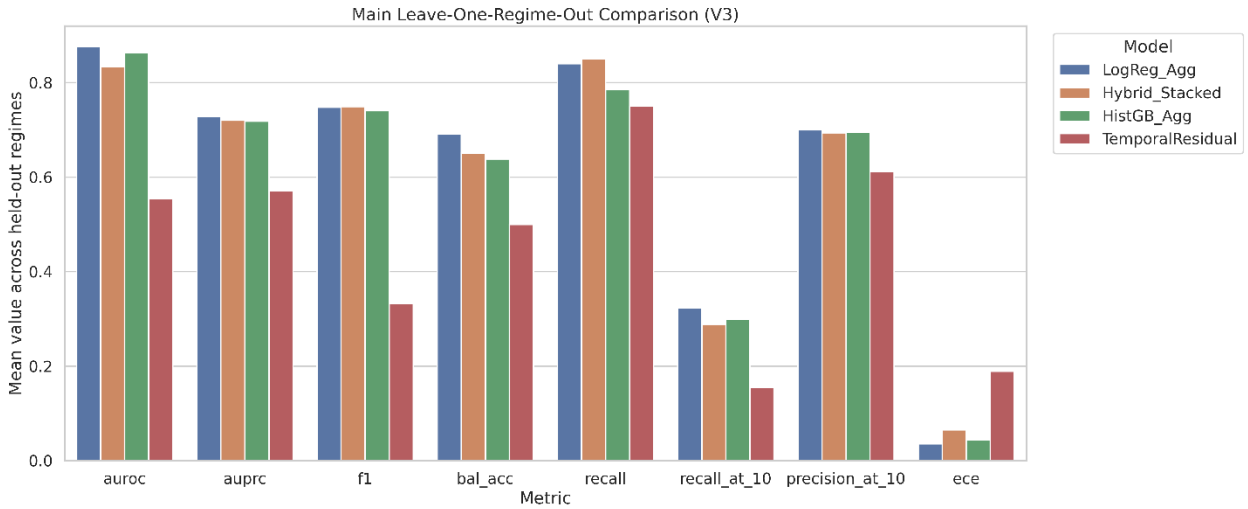

Fig. 4. Mean leave-one-regime-out performance across model families. Aggregate baselines remain strongest overall, while hybrid fusion offers only limited gains on selected metrics.

6.2 Worst-Regime Robustness

Average results alone can hide important weaknesses, so we also examine the most difficult held-out regime for each model. This view is especially important in a warning task, where failure under the hardest condition may matter more than small gains on easier folds.

The strongest result in this stricter setting comes from the **Hybrid Stacked** model, which achieves the best **worst-regime F1 of 0.509**. This is slightly higher than **0.505** for LogReg_Agg and **0.484**

for HistGB_Agg. This finding suggests that the hybrid model does extract some useful residual information that helps in the most difficult regime.

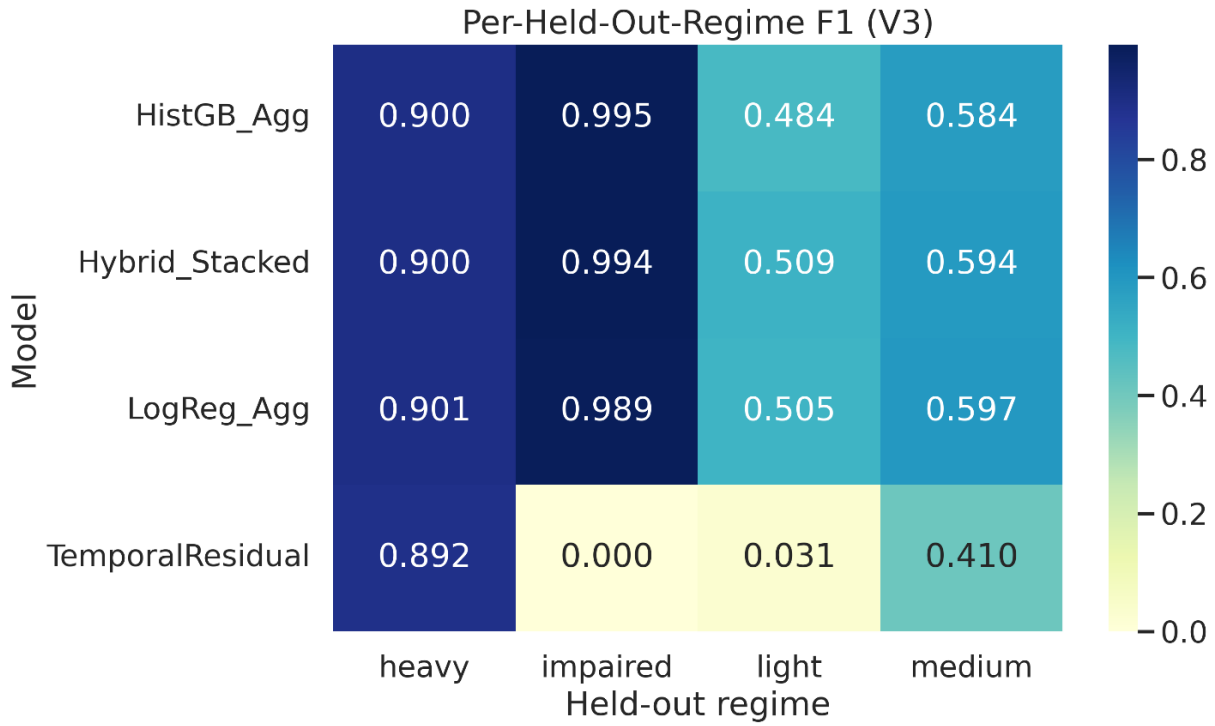

Fig. 5. Per-held-out-regime F1 scores. Hybrid stacking slightly improves worst-regime F1, but the gain is narrow and does not translate into a broad overall advantage.

However, that advantage is narrow and does not generalize across all worst-case metrics. For example, **worst-regime balanced accuracy** remains best for **LogReg_Agg at 0.581**, while Hybrid Stacked drops to **0.500**. A similar pattern appears in **worst-regime AUROC**, where **HistGB_Agg reaches 0.802**, **LogReg_Agg reaches 0.796**, and Hybrid Stacked falls to **0.763**. In addition, **worst-regime Recall@10 is identical for LogReg_Agg and Hybrid Stacked at 0.101**, which means the hybrid model does not improve low-budget warning capture in the hardest setting.

Taken together, these results show that hybrid fusion provides a **limited worst-regime gain**, but not a broad or consistent robustness advantage. It helps on one summary measure, yet it does not overturn the practical strength of the aggregate baseline.

6.3 Early-Warning Utility under Limited Review Budgets

Because the paper is framed as an early-warning study, budget-based performance is particularly important. In many operational settings, only a small fraction of the highest-risk intervals can be reviewed. For that reason, Recall@5, Recall@10, and Recall@20 provide a more realistic view of utility than average ranking quality alone.

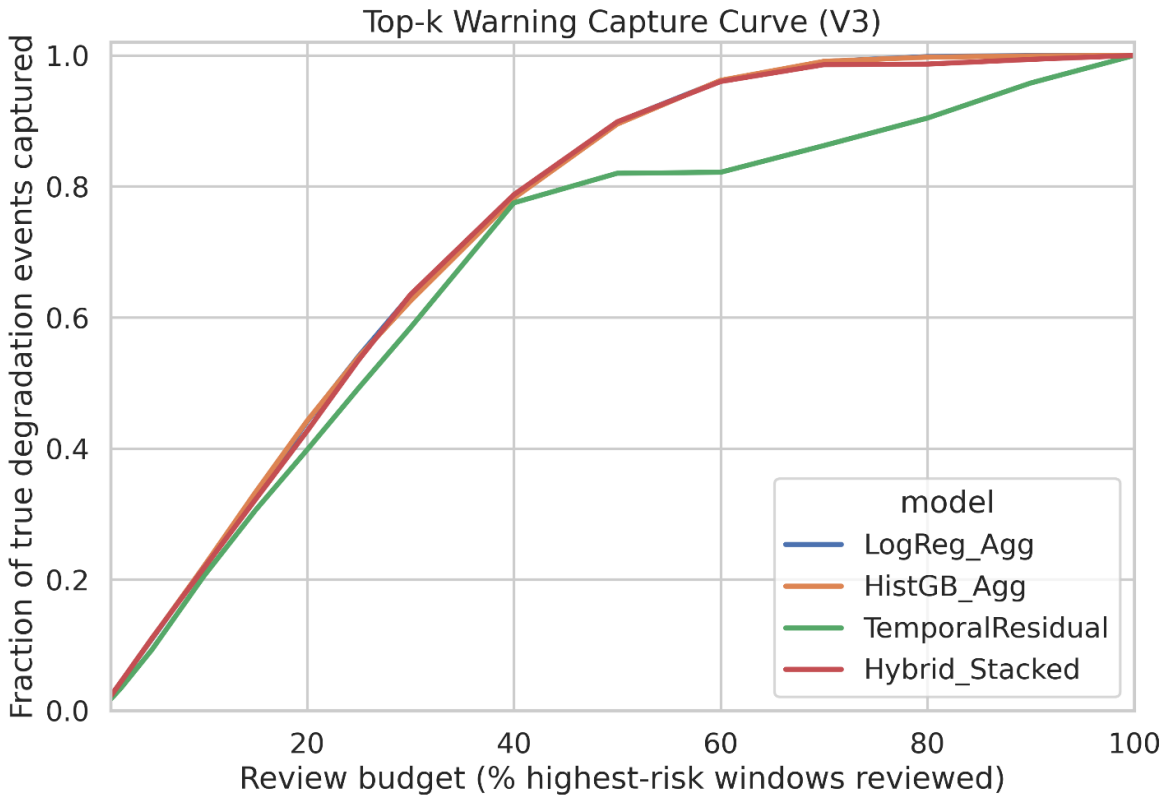

Fig. 6. Top-k warning capture curves under limited review budgets. Aggregate baselines capture more true degradation events at low review budgets, indicating stronger practical early-warning utility.

On these metrics, **LogReg_Agg again gives the strongest operational performance**. It achieves **Recall@5 of 0.239**, **Recall@10 of 0.323**, and **Recall@20 of 0.454**, while maintaining strong corresponding precision values. These results indicate that the model captures a larger share of true communication-degradation events when the review budget is limited.

The hybrid model performs reasonably well, but not better. Its **Recall@10 is 0.289**, which is clearly below both LogReg_Agg (**0.323**) and HistGB_Agg (**0.299**). The same pattern appears at **Recall@20**, where Hybrid Stacked reaches **0.401**, compared with **0.454** for LogReg_Agg. This means that although the hybrid model slightly increases full-threshold recall, it does not improve the more practically important low-budget warning capture.

This distinction is important. A model can look more sensitive after thresholding and still be less useful when only a small number of top-risk cases can be inspected. In the present benchmark, the low-budget warning setting continues to Favor the simpler aggregate baseline.

6.4 Calibration and Reliability

Probability quality is another important part of the evaluation. Since the task is framed as risk warning rather than pure ranking, predicted scores should behave like meaningful risk estimates. This is where calibration measures such as **Brier score** and **ECE** become important.

Among the strong models, **LogReg_Agg shows the best calibration balance**, with **Brier score of 0.076** and **ECE of 0.036**. HistGB_Agg is also well behaved, with slightly better Brier score (**0.075**) but somewhat worse ECE (**0.044**). The hybrid model is less reliable in this respect, with **Brier score of 0.079** and **ECE of 0.065**. The temporal residual model performs worst by a large margin.

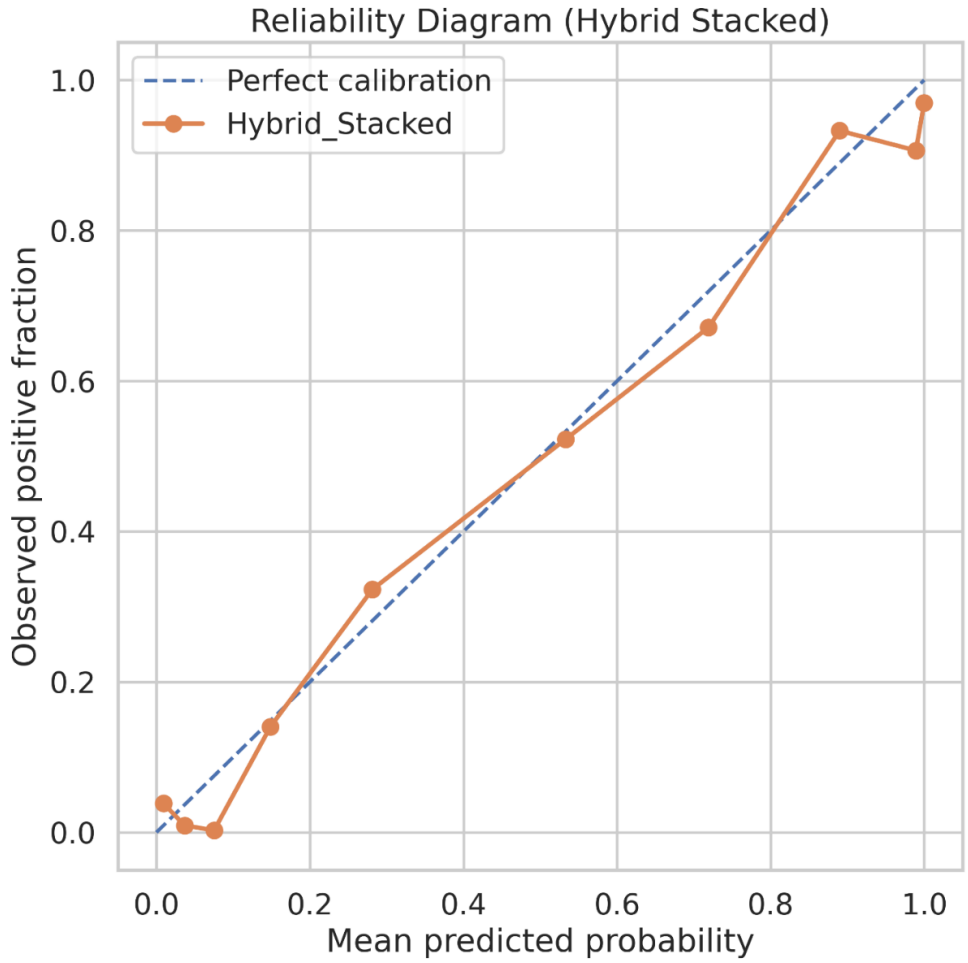

Fig. 7. Reliability diagram for the hybrid stacked model. The curve is reasonably aligned with the identity line, but calibration remains weaker than the strongest aggregate baseline.

These results reinforce a practical point. In a deployment-aware warning system, good ranking is not enough if the probabilities are unstable or poorly aligned with actual risk. In the current study, the aggregate baseline does not only rank well; it also produces the most trustworthy warning scores among the leading models. This makes it a stronger candidate for real decision support than a model that shows a small gain in one metric but weaker reliability overall.

6.5 Corruption Robustness

To examine whether the models remain useful under imperfect telemetry, we evaluate them under several stress scenarios, including random masking, burst masking, delayed telemetry, group-wise feature dropout, and additive noise. These settings do not represent direct attack traces, but they

provide controlled proxies for degraded or partially unreliable telemetry. This makes them relevant to the broader goal of defensive early warning, where the input stream may be incomplete, delayed, or noisy rather than fully clean.

Figure 8 shows that both **HistGB_Agg** and **Hybrid Stacked** remain relatively stable across the tested corruption settings, but neither model gains a decisive advantage. On average, **HistGB_Agg** retains slightly stronger **AUROC**, while **Hybrid Stacked** is slightly better on **F1** and **balanced accuracy** in several scenarios. However, these differences are small. The two models remain close under clean input, delayed telemetry, random masking, burst masking, and moderate noise. This suggests that hybrid fusion does not collapse under stress, but it also does not clearly displace the aggregate baseline.

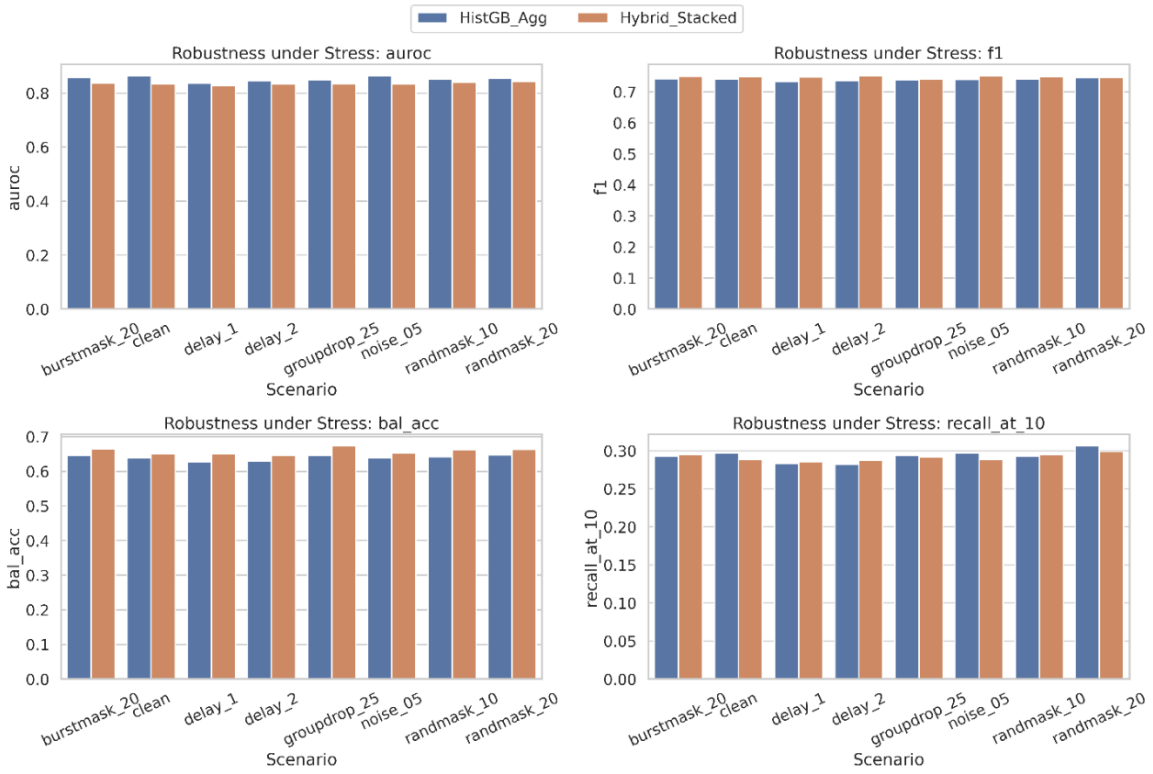

Fig. 8. Robustness under telemetry stress scenarios. Hybrid fusion preserves competitive performance in several cases, but the gains remain modest and do not displace aggregate baselines

A similar pattern appears in low-budget warning utility. Under corruption, **Recall@10** remains competitive for both models, but the gain from hybrid fusion is modest and inconsistent across scenarios. In some cases, the hybrid model preserves similar or slightly better threshold-based performance, while in others the aggregate baseline remains equally strong or stronger. This indicates that the hybrid model may absorb some branch disagreement under imperfect inputs, but the added value is limited rather than broad.

These findings are important for the overall interpretation of the paper. First, they show that the benchmark is not so fragile that small input degradation completely overturns the model ranking. Second, they suggest that robustness under telemetry stress is still driven largely by strong

aggregate structure. Third, they support a balanced conclusion: hybrid fusion can remain competitive under corruption, but its gains are modest, and aggregate baselines are still difficult to displace.

6.6 Interpretation and Practical Implications

The results suggest that this benchmark is driven more by **structured short-horizon summaries** than by complex temporal dynamics. Features such as recent level, variation, range, and direction of change appear to capture much of the usable information for next-interval communication-risk warning. This helps explain why both aggregate baselines remain strong and why the compact temporal branch performs poorly when used on its own.

The hybrid model offers a more nuanced picture. Its slight gain in mean F1 and worst-regime F1 suggests that temporal residual information is not completely useless. Rather, its value appears to be **incremental and selective**, not dominant. In other words, temporal signals may help at the margins, especially in harder cases, but they do not replace the importance of well-designed aggregate telemetry summaries.

From a practical perspective, this is an important outcome. It suggests that a simple and calibrated aggregate model may be a better operational choice than a more complex model whose gains are narrow and inconsistent. For early-warning systems in communication-sensitive autonomous settings, such simplicity is not a weakness. It can mean easier implementation, lower computational cost, stronger interpretability, and more stable calibration.

At the same time, the results also show that early warning under regime shift remains difficult. Even the strongest model leaves room for improvement, especially in the worst-regime and low-budget settings. This suggests that the benchmark is not yet saturated. Instead, it provides a realistic and useful testbed for future work on robust warning systems, better uncertainty handling, and evaluation on more realistic communication traces.

6.7 Summary of Findings

The results support three main conclusions. First, **LogReg_Agg is the strongest overall model** in this benchmark. Second, **Hybrid Stacked adds only limited value**, mainly through a small gain in mean F1 and worst-regime F1. Third, **Temporal Residual alone is not reliable**, which suggests that complex temporal modeling is not automatically beneficial for this task. Overall, the findings point to a clear and practical message: strong aggregate telemetry baselines remain difficult to beat under regime shift, and communication-degradation early warning in unseen conditions is still a challenging problem.

7. Limitations

This study has several limitations that should be stated clearly. First, the dataset used here is **structured benchmark data** rather than a raw field dataset collected from live operational systems. This is useful for controlled evaluation, but it also means that the results should be read as benchmark evidence, not as direct proof of real-world deployment performance. Second, the

data are **not UAV-native communication logs**. They do not contain flight records, onboard link measurements, mission context, or aerial control traces. For that reason, the connection to autonomous aerial systems is motivational rather than directly observational. Third, the benchmark does **not include RF-level attack traces**. It does not provide jammer measurements, spoofing signatures, spectrum captures, or attacker-labelled events. As a result, the models studied here should not be interpreted as direct detectors of jamming, spoofing, or hostile interference. Fourth, the target used in this paper is a **proxy label** derived from next-window SLA status. This makes it useful for studying communication-degradation early warning, but it does not support direct attribution of the cause of degradation. A predicted high-risk interval may reflect several possible conditions, not one specific threat mechanism. Fifth, this is a **controlled benchmark study**. The leave-one-regime-out setting is stricter than a random split and is valuable for testing generalization under shift, but it is still a simplified experimental setting. Real deployment environments may involve more complex shifts, longer-term drift, missing telemetry, and operational constraints that are not fully captured here.

For these reasons, we do not claim direct operational deployment readiness. The main value of this work is to provide a careful benchmark for comparing aggregate, temporal, and hybrid warning models under regime shift. Future work should test the same ideas on more realistic communication logs, UAV-relevant traces, and settings with richer operational context.

8. Conclusions

This paper studied communication-degradation early warning using a controlled 6G telemetry benchmark motivated by autonomous aerial and communication-sensitive systems. The results show that this benchmark is suitable for studying next-interval communication risk under changing operating conditions and for testing whether models remain useful when the regime shifts.

Across the evaluated models, the strongest overall results came from the aggregate telemetry baselines. In particular, the logistic-regression aggregate model provided the best overall balance of discrimination, decision quality, calibration, and low-budget warning capture. This suggests that well-designed telemetry summaries already contain much of the useful predictive signal in the present benchmark.

Hybrid stacking added only limited residual value. It produced a small improvement in mean F1, recall, and worst-regime F1, but these gains did not translate into a clear overall advantage over the best aggregate baseline. The temporal residual branch alone was not reliable, which further indicates that added sequence complexity is not automatically beneficial in this setting.

Overall, the findings support a careful and practical conclusion: communication-degradation early warning under regime shift remains difficult, and strong aggregate baselines are harder to beat than expected. Future work should test the same question on external datasets, UAV-native communication logs, RF-aware or multimodal traces, and more realistic operational environments.

References

[1] A. Elyasi, A. Ashdown, K. M. Rumman, and F. Restuccia, "O-RAN xApps: Survey and Research Challenges," *Computer Networks*, 2025, SSRN preprint, doi: **10.2139/ssrn.5236117**.

[2] M. M. Islam, K. Hasan, and S. H. Jeong, "Performance evaluation and optimization for 6G networks: A survey of KPIs, tools, and AI models," *ICT Express*, vol. **12**, no. **2**, pp. **390–416**, 2026, doi: **10.1016/j.icte.2025.12.012**.

[3] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Pérez, and C. Verikoukis, "Explainable AI in 6G O-RAN: A tutorial and survey on architecture, use cases, challenges, and future research," *IEEE Communications Surveys & Tutorials*, vol. **27**, no. **5**, pp. **2826–2859**, 2025, doi: **10.1109/COMST.2024.3510543**.

[4] K. Alam, M. A. Habibi, M. Tammen, D. Krummacker, W. Saad, M. Di Renzo, T. Melodia, X. Costa-Pérez, M. Debbah, A. Dutta, and H. D. Schotten, "A comprehensive tutorial and survey of O-RAN: Exploring slicing-aware architecture, deployment options, use cases, and challenges," *IEEE Communications Surveys & Tutorials*, vol. **28**, pp. **1637–1678**, 2026, doi: **10.1109/COMST.2025.3598406**.

[5] N. Baganal-Krishna, R. Lübben, E. Liotou, K. V. Katsaros, and A. Rizk, "A federated learning approach to QoS forecasting in cellular vehicular communications: Approaches and empirical evidence," *Computer Networks*, vol. **242**, art. **110239**, 2024, doi: **10.1016/j.comnet.2024.110239**.

[6] S. Hassouna, J. Kaur, B. Kizilkaya, J. U. R. Kazim, S. Ansari, A. A. Kherani, B. Lall, Q. H. Abbasi, and M. Imran, "Development of open radio access networks (O-RAN) for real-time robotic teleoperation," *Communications Engineering*, vol. **4**, no. **1**, art. **176**, 2025, doi: **10.1038/s44172-025-00524-0**.

[7] O. A. Mahdi, E. Pardede, S. Bevinakoppa, and N. Ali, "Federated learning under concept drift: A systematic survey of foundations, innovations, and future research directions," *Electronics*, vol. **14**, no. **22**, art. **4480**, 2025, doi: **10.3390/electronics14224480**.

[8] S. E. Meheretu, E. Nigussie, and G. B. Gebremeskel, "A systematic literature review on spoofing and jamming approaches in unmanned aerial vehicles navigation," *Journal of Aerospace Technology and Management*, 2025.

[9] U. Tariq, M. U. Zia, S. Khan, M. A. Khan, and others, "Systematic review of machine and deep learning models for unmanned aerial vehicles cyber threat defense," *Discover Artificial Intelligence*, vol. **6**, art. **216**, 2026, doi: **10.1007/s44163-026-00960-7**.

[10] N. Maitlo, M. H. Shah, A. Maitlo, G. Mustafa, K. Arshid, and N. Noonari, "Communication-fairness trade-offs in federated learning for 6G resource allocation: A 200-client study," *Inventions*, vol. **11**, no. **2**, art. **31**, 2026, doi: **10.3390/inventions11020031**.

[11] P. A. Mangi, S. Bibi, A. Nawaz, S. Bibi, and B. Raza, "When Clients Drift: Federated SLA-Risk Forecasting Across Unseen 6G RAN Regimes," *Spectrum of Engineering Sciences*, vol. 4, no. 4, 2026.

[12] B. Raza, A. Maitlo, Z. H. Shar, and I. Hyder, "Operational Android malware filtering: Calibrated probabilities and distribution-free guarantees," *Kashf Journal of Multidisciplinary Research*, vol. **2**, no. **12**, pp. **58–73**, 2025. **doi: 10.71146/kjmr778**

[13] S. Bibi, F. A. Rajput, M. Younis, S. Bibi, and B. Raza, "Vector+SQL retrieval with selectivity workloads: Measuring tail latency and quality under filtered Top-K," *VFAST Transactions on Software Engineering*, vol. **14**, no. **1**, pp. **335–349**, 2026, doi: **10.21015/vtse. v14i1.2353**.

[14] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. **1273–1282**.

[15] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems (MLSys)*, vol. **2**, 2020, pp. **429–450**.

[16] M. Polato, R. Bassoli, S. Lodi, and others, "Learning in federated and dynamic environments: A tutorial on drift, adaptation, and robustness," *Neurocomputing*, 2026.

[17] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. **132**, no. **12**, pp. **5635–5662**, 2024, doi: **10.1007/s11263-024-02117-4**.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. **1321–1330**.

[19] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, 2005, pp. **625–632**, doi: **10.1145/1102351.1102430**.

[20] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2002, pp. **694–699**, doi: **10.1145/775047.775151**.

[21] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. **61–74**.

[22] A. Bartsiokas, P. Gkonis, D. Kaklamani, and I. Venieris, "A federated learning-based resource allocation scheme for relaying-assisted B5G/6G networks," *Electronics*, vol. **13**, no. **3**, art. **390**, 2024, doi: **10.3390/electronics13030390**.

[23] M. K. Hasan, A. A. Habib, S. Islam, N. Safie, T. M. Ghazal, M. A. Khan, A. I. Alzahrani, N. Alalwan, S. Kadry, and A. Masood, "Federated learning enables 6G communication technology: Requirements, applications, and integrated intelligence framework," *Alexandria Engineering Journal*, vol. **93**, pp. **245–266**, 2024, doi: **10.1016/j.aej.2024.02.040**.

[24] A. Sajid, J. Lipman, M. Abolhasan, and others, "Towards SMPC-enabled O-RAN: A survey with deployment-oriented insights," *Computer Networks*, vol. **275**, art. **111816**, 2025, doi: **10.1016/j.comnet.2025.111816**.

[25] N. Omheni, H. Koubaa, and F. Zarai, "Artificial intelligence for 5G and 6G networks: A taxonomy-based survey of applications, trends, and challenges," *Technologies*, vol. **13**, no. **12**, art. **559**, 2025, doi: **10.3390/technologies13120559**.