

Gateway-level Early Warning of LoRaWAN Link Degradation Using LoED Measurements

Esma Fazilet Karagülle, Faruk Baturalp Günay, and Esra Odabaş Yıldırım

Atatürk University, Erzurum, Türkiye

 <https://doi.org/10.26636/jtit.2026.4.2789>

 Submitted: 8 Aug. 2026, Reviewed: 18 Sep. 2026, Accepted: 21 Sep. 2026

Abstract — LoRaWAN gateways observe packet-level metadata that may support reliability monitoring without modifying end devices or the protocol stack. We formulate short-term link degradation as a next-window warning task and aggregate LoED records into five-minute gateway windows described by CRC success rate, RSSI, SNR, spreading factor, traffic, frequency diversity, and historical features. A degradation event is a 10% or 20% decrease in the following window's CRC success rate. Class-balanced logistic regression (LR), random forest, XGBoost, and rule baselines are evaluated chronologically and with leave-one-gateway-out (LOGO) testing. LR obtains chronological PR-AUC values of 0.563 and 0.367 and exceeds both non-linear models in date-clustered randomization tests after Holm correction. Its LOGO means are also the highest, but the differences are not statistically significant, as only seven and five gateways contain both classes. Removing current CRC alone has a negligible effect, whereas removing all CRC history substantially reduces performance. Performance is further characterized across four window durations, probability calibration schemes, and explicit sequence models. A strictly chronological offline replay processes held-out windows with P99 scoring latency below 15 ms. The results support risk ranking within the LoED deployment, while gateway heterogeneity and calibration drift limit broader transfer claims.

Keywords — *early warning, edge monitoring, link degradation, LoRaWAN*

1. Introduction

LoRaWAN is widely used in low-power wide-area Internet of Things applications in which long-range connectivity and low energy consumption are prioritized over high data rates [1], [2]. However, in practical deployments, packet reception is not constant. Duty cycle restrictions, channel contention, interference, spreading factor distribution, gateway placement, and traffic density can jointly influence whether packets are received correctly [3], [4]. For monitoring applications, a short degradation in packet reception may cause missing measurements, delayed decisions, additional retransmissions, or temporary gaps in downstream analytics.

The operational difficulty is that a gateway sees heterogeneous packet observations rather than a direct explanation of the channel condition. The outcome of a single packet may be

shaped by instantaneous interference or by the behavior of a particular end device. A short sequence of packets received at a gateway, in contrast, can describe the recent state of that gateway. This motivates a window-level formulation in which recent packet observations are summarized and used to rank the risk of a reliability drop in the next short interval.

LoRaWAN management already includes mechanisms such as adaptive data rate (ADR), which adjusts the spreading factor and transmission power. However, empirical and analytical studies have shown that ADR behavior can be slow or deployment dependent when link conditions change [5]–[7]. Gateway-side monitoring should therefore be viewed as complementary to adaptation. It does not replace ADR or modify the protocol. Instead, it produces an early warning score that may help an operator decide when closer inspection, logging, threshold adjustment, or other network management actions are warranted.

We investigate whether gateway-observable metadata provides predictive information for short-term LoRaWAN link degradation warning in LoED deployment [8]. Packet records are aggregated into five-minute gateway windows described by reliability, signal quality, spreading factor, traffic, frequency diversity, and historical features. The target indicates whether the immediately following window experiences a specified drop in the CRC success rate.

The contribution is fourfold. First, the study defines the LoRaWAN degradation warning as a strict next-window prediction problem rather than a packet-level classification or retrospective link characterization. Second, all predictors are observable at a gateway, which allows implementation at the edge or on a network server. Third, learned models and rule baselines are evaluated under chronological and leave-one-gateway-out (LOGO) protocols with clustered uncertainty estimates. Fourth, alarm budgets, CRC ablations, window duration, calibration, gateway heterogeneity, explicit temporal models, and streaming behavior are examined to establish the scope and operational limits of the warning score.

The remainder of the paper is organized as follows. Section 2 positions the work with respect to LoRaWAN measurement studies, link adaptation, wireless link quality estimation, and edge intelligence. Section 3 describes the dataset, feature

Tab. 1. Position of the proposed study with respect to related research directions.

Research direction	Typical objective	Main limitation for this study	Position of this paper
LoRaWAN measurement studies	Characterize protocol behavior and real traffic variability	Mostly descriptive or retrospective	Uses real LoED observations for predictive warning
ADR and link adaptation	Select spreading factor or transmit power	Focuses on control, not early warning	Provides a risk score that can precede control actions
Wireless LQE with ML	Estimate link quality or packet delivery	Often packet/link-class focused	Predicts next-window CRC degradation
Edge intelligence	Move inference closer to the network edge	Often architectural or application level	Uses lightweight gateway-window features

construction, label definition, models, baselines, and evaluation protocols. Section 4 presents the experimental results. Section 5 discusses operational implications and limitations, while Section 6 concludes the paper.

2. Related Works

Measurement-based studies have shown that LoRaWAN reliability is shaped by both protocol constraints and deployment-specific conditions. The authors of [3] discuss duty cycle restrictions, interference, scalability limits, and traffic load effects in LoRaWAN networks, while research presented in [4] analyzes operational traffic from The Things Network and reports substantial spatial and temporal variability in real deployments. Paper [9] shows that LoRaWAN traffic traces may contain behavior that is difficult to interpret from the protocol configuration alone. These studies indicate that reliability monitoring should not rely solely on static link assumptions, as packet reception may vary with time due to gateway location, and traffic conditions.

The LoED data set contains packet observations from an urban LoRaWAN deployment, including reception time, gateway identifier, CRC status, RSSI, SNR, spreading factor, frequency, and payload size [8]. These fields are normally available at gateways and allow a degradation risk analysis to be performed without application payload inspection, device-side state, or protocol modification. Packet observations are aggregated into short gateway-specific windows rather than treated as independent classification instances, meaning that each monitoring state represents recent reception behavior.

LoRaWAN reliability has also been studied through link adaptation. ADR adjusts parameters such as spreading factor and transmit power according to network conditions, but previous work has shown that ADR may respond slowly when channel conditions change [5]. Therefore, several extensions and analytical studies have examined ways to improve fairness, robustness, or resource use in LoRaWAN parameters [6], [7]. These studies address an important control problem: how communication parameters should be configured once link conditions are observed or inferred. A complementary monitoring objective is considered here: rather than proposing a new ADR mechanism, the analysis tests whether gateway-observable metadata can provide an early warning that the

next short time interval is likely to experience a drop in CRC reliability.

Machine learning has been used more broadly for wireless link quality estimation and monitoring. In [10], the learning-based link quality estimators are reviewed, and the importance of feature design, class imbalance and evaluation methodology is emphasized. Related work on link quality classification also shows that reported performance may depend on the data distribution and evaluation protocol [11]. Temporal models such as LSTM have been considered for link quality estimation [12], and LoRaWAN-specific learning studies have used RSSI and SNR for tasks such as spreading factor prediction [13]. Recent surveys of data-driven LoRaWAN solutions also identify forecasting the future network state from current and historical observations as a key direction [14], while explainable lightweight learning models have been advocated for LoRaWAN link analysis [15]. Compared to these directions, the present study uses a narrower operational target. It predicts whether the CRC success rate of the next gateway window will decrease by a specified amount. The unit of analysis is therefore a gateway-window, not an individual packet, device, link quality class, or communication parameter decision.

The proposed formulation is also consistent with the principles of edge intelligence. Moving light analysis toward the edge may reduce latency and bandwidth use while supporting local decision making [16]. Paper [17] focusing on the LoRaCog framework similarly illustrates the value of assigning sensing and decision support functions to gateways in LoRa-related systems. In the present study, this idea is applied to reliability monitoring: a gateway or a network server can transform recently observed metadata into a degradation risk score and use that score for alarm prioritization. The emphasis is therefore on deployable risk ranking rather than on complex model design.

Together, previous works have characterized LoRaWAN behavior, improved link adaptation, and developed learning-based link quality estimators. Among the studies reviewed here, short-horizon degradation warning based only on gateway-observable LoED measurements has received less direct attention. This study addresses that specific problem by defining strict next-window degradation labels, applying temporally ordered and LOGO evaluation protocols, comparing learned models with simple rule-based baselines, analyz-

ing alarm-budget behavior, and testing the sensitivity of the model to CRC-derived features.

Table 1 summarizes the position of the proposed study with respect to the main related research directions.

3. Materials and Methods

Figure 1 summarizes the overall gateway-level early warning pipeline used in this study, from LoED packet records to alarm-oriented evaluation. The following subsections describe each stage in detail.

3.1. LoED Dataset and Preprocessing

The experiments use the full LoED dataset, which contains LoRaWAN packet observations collected on the gateway side in a dense urban environment [8]. Each packet record includes metadata such as reception time, gateway identifier, CRC status, frequency, spreading factor, RSSI, SNR, and payload size. Following the LoED encoding, CRC status 1 is assigned to successful reception and status 0 or –1 to unsuccessful reception; other or missing status values are excluded. Physically implausible RSSI and SNR values are removed before aggregation. The analysis accepts the SF6–SF12 values observed in the data and treats values outside that explicit analysis range as missing.

The processing pipeline used 188 daily CSV files and produced 112 214 five minute gateway-window samples before strict next-window alignment. Since the dataset is organized as daily files rather than as a continuous uninterrupted trace, target construction keeps only those window pairs for which the following gateway window exists exactly five minutes af-

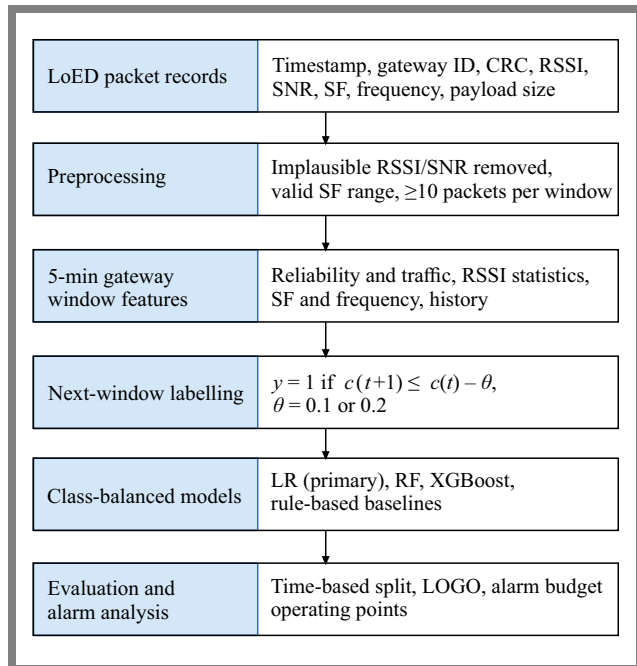


Fig. 1. Gateway-level early warning pipeline used in the proposed LoRaWAN monitoring framework.

Tab. 2. Data set and event summaries statistics.

Item	Value
Daily CSV files processed	188
Gateway-window samples before strict alignment	112 214
Valid current-next gateway-window pairs	110 157
Model-eligible pairs after one-step history	110 148
Gateways	9
Collection dates	187
Moderate degradation event rate ($\theta = 0.10$)	9.69%
Severe degradation event rate ($\theta = 0.20$)	2.70%

ter the current one. This prevents artificial labels from being created across gaps or between disconnected daily traces.

After alignment, 110 157 valid current-next gateway-window pairs remain across nine gateways and 187 collection dates, spanning 8 Feb. 2019 to 2 Sep. 2020. Construction of lagged predictors removes the first eligible pair for each gateway, leaving 110 148 model-eligible samples. Analyses that additionally require six exactly contiguous historical windows use a smaller common sample, which is reported with the corresponding sequence results. The data set scale is useful for studying rare degradation events, but it also creates class imbalance. The severe 20% CRC drop event appears in only a small fraction of the windows. Therefore, the evaluation emphasizes ranking and detection metrics rather than raw accuracy.

3.2. Gateway-window Feature Representation

Packet records are grouped by gateway and five-minute intervals. Windows with fewer than ten packets are excluded to avoid unstable CRC rate estimates. For every retained window, four feature groups are extracted. Reliability and traffic features include packet count, payload size statistics, and current CRC success rate. Radio features include summary statistics of RSSI and SNR. Transmission characteristics include spreading factor statistics, spreading factor ratios, and frequency diversity. Temporal context features include hour, day of week, previous CRC success rate, previous CRC change, rolling CRC mean and standard deviation, and rolling RSSI/SNR means.

The feature set is intentionally limited to information already observable in gateways. No application payload content, device-side state, or future packet information is used. Frequency diversity is defined as the number of distinct frequencies observed in a gateway window. Changes in this descriptive count may reflect channel selection and traffic behavior. The spreading factor ratios summarize the distribution of the observed link adaptation states. RSSI and SNR provide radio context, while packet count represents the support available to estimate reliability in a window.

All temporal features are computed using only the current or past windows. The next-window CRC success rate is used only to construct the label. The gateway identifier is treated as a categorical variable. In LOGO testing, unseen gateway categories are ignored by the encoder, preventing the held-out gateway label from being memorized.

3.3. Degradation Event Definition

Let $c(g, t)$ denote the CRC success rate for gateway g in current window t . The target is derived from the immediately following window $t + 1$, which must exist exactly five minutes later for the same gateway. For a threshold θ , the degradation event is defined as a decrease of at least θ in the next-window CRC success rate, as:

$$y(g, t; \theta) = \begin{cases} 1 & \text{if } c(g, t + 1) \leq c(g, t) - \theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Because window-level CRC rates are rational proportions represented in floating-point form, the implementation applies a numerical tolerance of 10^{-12} to the comparison in Eq. (1). This ensures that a decrease mathematically equal to θ is consistently included before and after CSV serialization.

Two thresholds are evaluated. 10% represents moderate short-term degradation, while 20% represents a rarer and more severe event. These thresholds are not presented as universal service-level objectives. They are used to test whether the proposed pipeline can identify both frequent moderate and less frequent severe degradation. In a deployment, θ should be selected according to the tolerance of reliability of the application.

Because the task is unbalanced, PR-AUC, recall and alarm-oriented operating points are more informative than accuracy alone [19]. A classifier that always predicts no degradation can appear accurate under the severe threshold, while failing to detect any event.

3.4. Models, Baselines, and Evaluation Protocols

Logistic regression with class balancing is used as the primary model because it provides a low-cost risk score that is straightforward to deploy and inspect. Numeric features are median-imputed, and logistic regression additionally uses standardization. Random forest and XGBoost are included as non-linear alternatives, with XGBoost serving as a common high performance tabular baseline [18]. Because class balancing changes the probability scale, calibration is evaluated separately rather than assumed from ranking performance.

Four rule-based baselines are evaluated: always no-drop, previous-drop rule, current-low rule and trend-down rule. Always no-drop represents the accuracy trap in imbalanced monitoring. The previous drop predicts the risk when the previous change already exceeded the same threshold. Current-low uses a low current CRC success rate, below 0.5, as a simple alarm. Trend-down predicts risk when the current CRC state is still acceptable, equaling at least 0.5, but the recent change is negative. These baselines test whether learning provides information beyond simple temporal heuristics.

Two evaluation protocols are used. The time-based split trains on the earliest 70% of collection dates and tests on the remaining 30%. This evaluates within-deployment forecasting over later dates. The LOGO protocol trains on all but one gateway and tests on the held-out gateway. Since some held-out gateways contain only one event class for a threshold, the

transfer summary is reported over binary test gateways where both event and non-event samples are present.

Operating point analysis treats the logistic regression output as a rank alarm score. For a fixed alarm budget, such as the top 5% or top 10% highest-risk windows, the analysis measures how many degradation events are captured. A complementary fixed-recall view measures the alert rate required to capture a target fraction of degradation events. This is closer to how a monitoring dashboard would use the model than a single universal probability threshold.

3.5. Implementation and Reproducibility Controls

All models are trained on the same feature matrix and threshold-specific labels. Logistic regression uses a class-balanced loss and the liblinear solver with a maximum of 5000 iterations. Random forest uses 300 trees, maximum depth 8, minimum leaf size 8 and class balancing. XGBoost uses 300 trees, maximum depth 3, learning rate 0.03, subsampling 0.9, column sampling 0.9 and a positive class weight computed from the training split. These settings are intentionally moderate to avoid turning the study into a hyperparameter search exercise.

Preprocessing is fitted only on the training portion of each split. This rule applies to imputation, scaling, and categorical encoding in the time-based split, LOGO evaluation, ablation experiments, and operating point analysis. The reported results therefore reflect the score that would be available when a model is trained on past dates or on other gateways and then applied to held-out evaluation data.

3.6. Extended Validation Design

Statistical uncertainty is evaluated at the level of collection units rather than individual gateway windows. For chronological evaluation, the 95% confidence intervals are estimated by non-parametrically resampling the 57 held-out collection dates as clusters [25]. Model comparisons use two-sided paired randomization tests. All scores from a collection date are swapped as a block between the two models, and the global PR-AUC or ROC-AUC difference is recomputed for 1999 random assignments. LOGO intervals use held-out gateways as the resampling unit, and model differences are assessed with exact two-sided sign-flip tests. Holm's procedure controls family-wise error within each threshold, metric, and comparison family [20].

Feature sensitivity is examined through seven LR specifications: complete representation, current CRC alone, omission of current CRC, omission of current and rolling CRC summaries, omission of all CRC-derived variables, current window radio variables alone, and delayed radio variables alone. Temporal-resolution sensitivity is evaluated by reconstructing the data at 1, 5, 10 and 15 minute intervals while retaining the ten packet minimum and exact next-window alignment. Because interval length changes the forecast horizon, prevalence, and eligible sample population, these settings are interpreted as related operational tasks rather than paired hyperparameter trials.

Probability calibration is quantified with reliability diagrams, Brier score and ten-bin expected calibration error (ECE) [22], [24]. A leakage-controlled recalibration experiment fits LR on the earliest 56% of dates, estimates sigmoid and isotonic mappings on the next 14%, and evaluates them on the untouched final 30%. In LOGO evaluation, the recalibration uses only dates from the training gateways.

Observable event precursors are compared after robust standardization within the gateway using Cliff's delta [21]. These associations are descriptive and are not treated as estimates of physical causes.

Two explicit sequence models use six exactly contiguous gateway windows, corresponding to 30 minutes of history. Sequence LR operates on the flattened lag vector, whereas a compact gated recurrent unit (GRU) uses 16 hidden units and early stopping on later training dates [23]. Finally, a strictly chronological offline replay processes final test gateway-windows in the timestamp order. Its latency measurement covers imputation, scaling, categorical encoding, LR inference, and sigmoid calibration, but excludes raw packet ingestion and window aggregation.

4. Results

Throughout this section, performance is reported using precision, recall, F1, ROC-AUC, and PR-AUC. Because degradation events are rare, PR-AUC, recall and the alarm-budget operating points are treated as the primary indicators, in line with the class imbalance considerations discussed in Subsection 3.3.

4.1. Time-based Evaluation

Table 3 reports the chronological test results for the three learned models. Logistic regression provides the strongest overall risk ranking. For the 10% threshold, it reaches 0.923 ROC-AUC, 0.563 PR-AUC, and 0.928 recall. For the rarer 20% threshold, it reaches 0.960 ROC-AUC, 0.367 PR-AUC, and 0.938 recall. Precision is lower at the severe threshold, because events are rare and the class-balanced model is tuned toward high sensitivity.

Non-linear alternatives do not consistently improve alarm ranking metrics. XGBoost is competitive at the 10% threshold, but logistic regression gives the highest PR-AUC at both thresholds. Date-clustered paired randomization tests indicate that LR exceeds RF and XGBoost for PR-AUC and ROC-AUC at both thresholds after Holm correction ($p_{\text{Holm}} \leq 0.007$). Together with its comparable performance across the extended analyses, this supports selecting LR as the parsimonious primary model for this data set.

Time-based results should be interpreted as within-deployment forecasting rather than universal generalization. They show that recent radio, traffic, and reliability history contain enough information to rank near-future degradation risk when the model has observed earlier windows from the same deployment period.

Tab. 3. Time-based early warning performance.

θ	Model	Precision	Recall	F1	ROC-AUC	PR-AUC
0.10	LR	0.298	0.928	0.451	0.923	0.563
0.10	RF	0.265	0.893	0.409	0.890	0.441
0.10	XGBoost	0.299	0.867	0.444	0.902	0.501
0.20	LR	0.133	0.938	0.232	0.960	0.367
0.20	RF	0.130	0.767	0.223	0.923	0.223
0.20	XGBoost	0.124	0.843	0.216	0.937	0.302

Tab. 4. LOGO performance on binary test gateways.

θ	Model	Precision	Recall	F1	ROC-AUC	PR-AUC
0.10	LR	0.314	0.496	0.347	0.786	0.407
0.10	RF	0.246	0.641	0.334	0.773	0.368
0.10	XGBoost	0.257	0.674	0.358	0.765	0.370
0.20	LR	0.265	0.708	0.318	0.868	0.376
0.20	RF	0.222	0.644	0.280	0.826	0.275
0.20	XGBoost	0.182	0.824	0.286	0.841	0.319

4.2. LOGO Evaluation

Table 4 reports LOGO performance on binary test gateways. This protocol is more difficult because the model is evaluated on a gateway location not observed during training. Logistic regression gives the highest PR-AUC and ROC-AUC for both thresholds. XGBoost produces higher recall, but with lower precision and lower PR-AUC, indicating more aggressive alerting.

The performance gap between the chronological and LOGO settings is a key operational finding. It shows that the LoRaWAN degradation warning is partially deployment-specific. LR has the highest observed mean LOGO PR-AUC at both thresholds, but paired gateway tests do not remain significant after Holm correction. A lightweight model can transfer some risk information across gateways, but the alarm threshold should be calibrated with local gateway data before operational use.

The decline is consistent with the shift in deployment. A gateway installed at a different location may observe different interference patterns, traffic levels, and RSSI/SNR distributions. The LOGO protocol exposes this change and restricts interpretation. Representation retains some discrimination, but direct threshold reuse across gateways is not sufficient.

4.3. Baseline Comparison

Table 5 compares the learned model with simple edge deployable rules under the time-based and LOGO protocols, respectively. Always no-drop can obtain high accuracy when degradation is rare, but cannot detect any event. Current-low, previous-drop, and trend-down rules capture limited parts of the event set and yield lower PR-AUC values. Logistic regression therefore extracts additional signal beyond trivial temporal rules while remaining simple enough for gateway-level deployment. For the rule baselines, precision and recall follow from the binary rule decision, whereas PR-AUC is

Tab. 5. Baseline comparison for alarm detection under time-based split and LOGO splits.

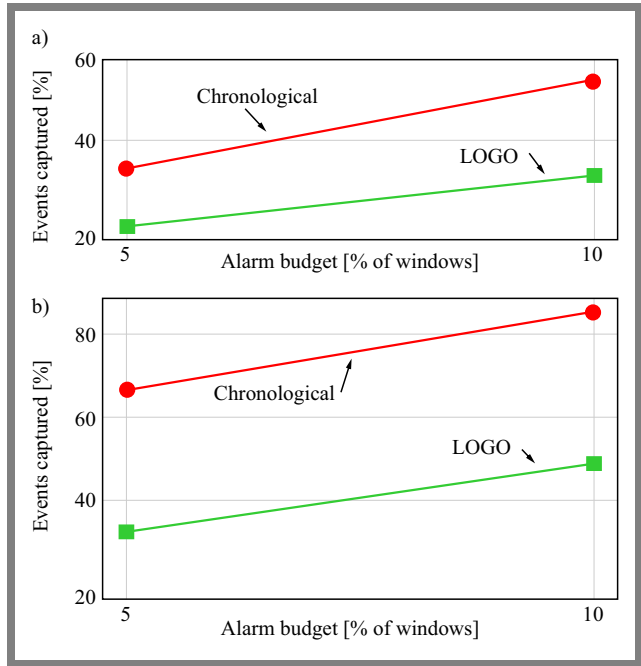
θ	Model	Precision	Recall	PR-AUC
Time-based split				
0.10	AlwaysNoDrop	0.000	0.000	0.099
0.10	CurrentLowRule	0.117	0.439	0.125
0.10	PrevDropRule	0.035	0.035	0.096
0.10	TrendDownRule	0.137	0.078	0.059
0.10	LR	0.298	0.928	0.563
0.20	AlwaysNoDrop	0.000	0.000	0.021
0.20	CurrentLowRule	0.016	0.286	0.022
0.20	PrevDropRule	0.000	0.000	0.021
0.20	TrendDownRule	0.019	0.051	0.011
0.20	LR	0.133	0.938	0.367
LOGO split				
0.10	AlwaysNoDrop	0.000	0.000	0.167
0.10	CurrentLowRule	0.051	0.140	0.110
0.10	PrevDropRule	0.047	0.047	0.159
0.10	TrendDownRule	0.130	0.149	0.110
0.10	LR	0.314	0.496	0.407
0.20	AlwaysNoDrop	0.000	0.000	0.079
0.20	CurrentLowRule	0.016	0.171	0.049
0.20	PrevDropRule	0.005	0.005	0.079
0.20	TrendDownRule	0.053	0.069	0.047
0.20	LR	0.265	0.708	0.376

computed from the associated graded score – a constant for always no-drop, the binary output for previous-drop, one minus the current CRC success rate for current-low, and the magnitude of the recent CRC change for trend-down. A weakly informative or anti-correlated score can therefore yield a PR-AUC below the event prevalence, as observed for the trend-down rule, which is consistent with short-term mean reversion in the window-level CRC success rate.

4.4. Alarm-oriented Operating Points

Since the task is unbalanced, the model is best interpreted as an alarm ranking score. Figure 2 and Tab. 6 show how many events are captured when only the highest risk windows are selected. In a time-based evaluation, alerting the top 10% highest risk windows captures 54.8% of moderate degradation events and 85.8% of severe degradation events. These values correspond to increases of 5.48 and 8.58 over random alerting. In the LOGO testing, the same 10% budget captures 31.6% and 48.7% of the events, respectively. The PR-AUC for severe events, equaling 0.367, is substantially above the 2.08% test prevalence, while the precision of 17.8% in a 10% alarm budget means that approximately one in six selected windows precedes a severe drop. This indicates the concentration of risk in the held-out period while quantifying the review burden created by false alarms.

Table 7 provides a complementary fixed-recall view. To capture 80% of severe time-based events, 7.9% of windows must be alerted, with 0.210 precision. For unseen gateways, the required alert rate increases, again showing that transfer requires

**Fig. 2.** Degradation events captured in alarm budgets for time-based and LOGO settings for drop threshold: a) 10% and b) 20%.**Tab. 6.** Operating points for logistic regression.

Split	θ	Budget [%]	Precision	Recall	Lift
Time	0.10	5	0.660	0.334	6.689
Time	0.10	10	0.540	0.548	5.476
Time	0.20	5	0.278	0.669	13.387
Time	0.20	10	0.178	0.858	8.575
LOGO	0.10	5	0.535	0.192	3.848
LOGO	0.10	10	0.475	0.316	3.164
LOGO	0.20	5	0.466	0.320	6.395
LOGO	0.20	10	0.361	0.487	4.873

Tab. 7. Fixed-recall operating points for logistic regression.

Split	θ	Alert rate for 80% recall	Precision	Achieved recall	Lift
Time	0.10	20.2%	0.391	0.800	3.966
Time	0.20	7.9%	0.210	0.801	10.081
LOGO	0.10	45.3%	0.293	0.803	1.795
LOGO	0.20	26.8%	0.231	0.800	3.112

deployment-specific calibration. These operating points are more actionable than a single default classification threshold. A network operator may decide that severe degradation is costly and accept a larger alert budget, or restrict alerts when manual inspection is expensive. The model score can therefore be used to rank windows for monitoring attention, even when absolute probability calibration remains imperfect.

4.5. Feature-level Interpretation and Ablation

The standardized logistic regression coefficients indicate that recent CRC behavior, SNR, RSSI, packet count, and frequency diversity are associated with degradation risk. Figure 3 shows the standardized coefficients of the largest

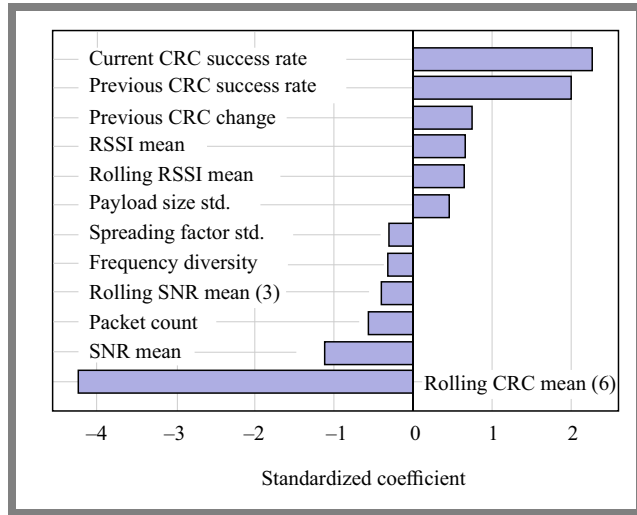


Fig. 3. Standardized logistic regression coefficients for $\theta = 0.10$. Gateway identifiers are excluded; coefficient magnitude does not establish causal importance.

Tab. 8. Sensitivity to removing current and rolling CRC features.

Split	θ	Feature setting	Recall	ROC-AUC	PR-AUC
Time	0.10	Full features	0.928	0.923	0.563
Time	0.20	Full features	0.938	0.960	0.367
Time	0.10	Without current/rolling CRC	0.931	0.904	0.513
Time	0.20	Without current/rolling CRC	0.910	0.947	0.318
LOGO	0.10	Full features	0.496	0.786	0.407
LOGO	0.20	Full features	0.708	0.868	0.376
LOGO	0.10	Without current/rolling CRC	0.398	0.735	0.371
LOGO	0.20	Without current/rolling CRC	0.629	0.841	0.335

magnitude for $\theta = 0.10$. The correlation and multicollinearity among the predictors limit the interpretation of the individual coefficients. They should not be interpreted as causal or physical importance estimates, but describe associations in a short-horizon warning model.

The ablation in Tab. 8 removes current CRC and rolling CRC features while retaining one-step previous CRC terms. Performance decreases but does not collapse. Time-based PR-AUC changes from 0.563 to 0.513 for $\theta = 0.10$ and from 0.367 to 0.318 for $\theta = 0.20$. LOGO PR-AUC changes from 0.407 to 0.371 and from 0.376 to 0.335, respectively. Thus, the CRC history is informative, but the warning signal is not determined exclusively by current and rolling CRC statistics. Because the label is a future drop in the CRC success rate, current and recent CRC states are plausible predictors. An important question is whether the model provides information beyond the current CRC level. Removing current CRC alone has a negligible effect, whereas removing all the CRC-derived history causes a large and statistically significant chronological reduction of PR-AUC. Radio and traffic context retain weaker predictive information.

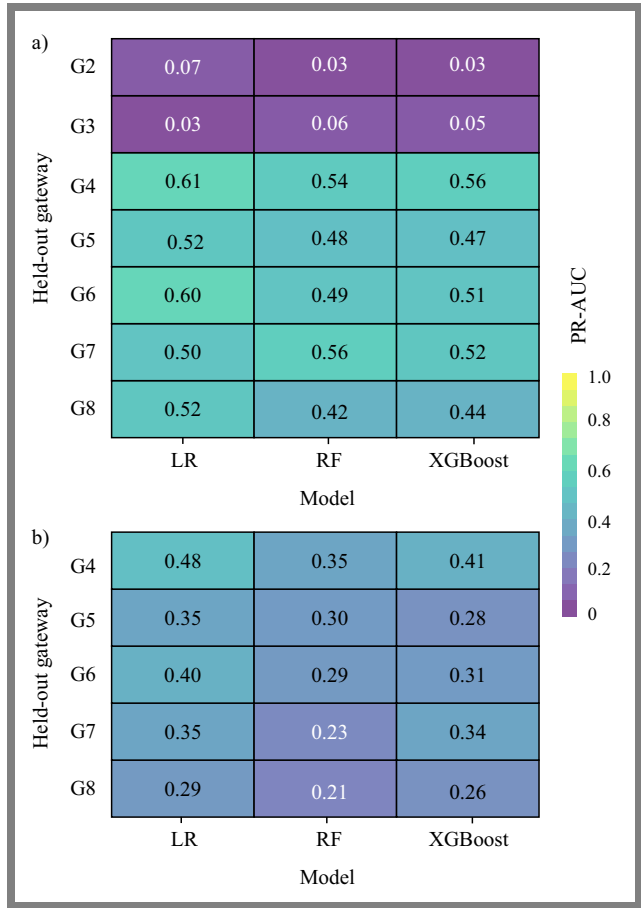


Fig. 4. PR-AUC under LOGO evaluation. Only gateways that contain both classes are shown. Drop threshold: a) 10% and b) 20%.

4.6. Statistical Uncertainty and Cross-gateway Variation

When considering within-date dependence, LR reaches a PR-AUC of 0.563 (95% cluster bootstrap CI 0.544–0.582) for $\theta = 0.10$ and 0.367 (0.328–0.407) for $\theta = 0.20$. Its paired PR-AUC and ROC-AUC differences relative to RF and XGBoost remain significant in date-clustered randomization tests after Holm correction ($p_{Holm} \leq 0.007$). The corresponding LOGO PR-AUC intervals are substantially wider, at 0.407 (0.233–0.564) and 0.376 (0.319–0.441). Although the LOGO means favor LR, neither comparison remains significant after multiplicity correction. This distinction indicates stable within-deployment separation but greater uncertainty when the gateway location is withheld.

The fold-level results shown in Fig. 4 reveal that cross-gateway variation extends beyond average discrimination. The prevalence ranges from 0.8% to 30.3% among the evaluable gateways. On the two lowest-prevalence gateways, LR retains ROC-AUC values of 0.800 and 0.692, yet its default decision threshold produces zero recall. Several higher-prevalence gateways instead exceed 0.91 recall. Thus, a useful ordering of risk can coexist with an unsuitable operating threshold when prevalence and feature distributions change.

Tab. 9. Anonymized gateway-level LR results for LOGO folds containing both classes.

Gateway	θ	n	Events	Prevalence	Recall	ROC-AUC	PR-AUC
G2	0.10	5180	42	0.008	0.000	0.800	0.065
G3	0.10	15 749	168	0.011	0.000	0.692	0.028
G4	0.10	6824	2070	0.303	0.977	0.790	0.609
G5	0.10	4083	895	0.219	0.356	0.793	0.518
G6	0.10	3359	810	0.241	0.914	0.825	0.604
G7	0.10	330	67	0.203	0.821	0.794	0.502
G8	0.10	37 046	6730	0.182	0.405	0.806	0.520
G4	0.20	6824	991	0.145	0.905	0.824	0.480
G5	0.20	4083	282	0.069	0.436	0.869	0.349
G6	0.20	3359	280	0.083	0.968	0.880	0.404
G7	0.20	330	20	0.061	1.000	0.899	0.352
G8	0.20	37 046	1431	0.039	0.229	0.869	0.294

Tab. 10. Chronological LR sensitivity to the duration of the aggregation window. Each duration defines a different forecast horizon and eligible population.

Window	θ	Test n	Events	Prevalence	ROC-AUC	PR-AUC
1 min	0.10	47 083	2056	0.0437	0.981	0.667
1 min	0.20	47 083	974	0.0207	0.981	0.514
5 min	0.10	35 776	3528	0.0986	0.923	0.563
5 min	0.20	35 776	744	0.0208	0.960	0.367
10 min	0.10	20 229	1561	0.0772	0.914	0.481
10 min	0.20	20 229	267	0.0132	0.958	0.289
15 min	0.10	14 040	801	0.0571	0.891	0.334
15 min	0.20	14 040	89	0.0063	0.882	0.041

4.7. Dependence on Reliability History

The expanded ablation analysis separates the current CRC measurement from its temporal history. Omitting current CRC changes chronological PR-AUC by less than 0.0001 at both thresholds, showing that the fitted classifier is not equivalent to a current-state threshold. By contrast, omission of every CRC-derived historical variable lowers PR-AUC from 0.563 to 0.280 for $\theta = 0.10$ and from 0.367 to 0.112 for $\theta = 0.20$. Both reductions remain significant in date-clustered randomization tests after Holm correction ($p_{Holm} = 0.006$). Models restricted to current or delayed RSSI and SNR retain limited chronological discrimination but transfer weakly across gateways. Therefore, recent reliability dynamics provide most of the observed predictive information, with radio and traffic measurements providing secondary context.

4.8. Sensitivity to Temporal Resolution

The duration changes both the warning horizon and the statistical support available for feature estimation. The one-minute windows achieve the highest chronological PR-AUC, reaching 0.667 and 0.514 for the moderate and severe outcomes,

respectively. However, the fixed requirement of ten packets removes one gateway and reduces the number of evaluable LOGO folds. At five minutes, all nine gateways are retained and the severe event LOGO PR-AUC has the smallest between-gateway standard deviation. Ten minute windows provide more evaluable severe event folds but greater PR-AUC dispersion, whereas 15 minute aggregation reduces severe event prevalence to 0.63% in the chronological test and PR-AUC to 0.041. These patterns favor five minutes as a balance between warning time, packet support, and cross-gateway stability, without implying a universally optimal interval.

4.9. Probability Calibration

Ranking discrimination does not translate directly into calibrated probability estimates. In the severe chronological test, the prevalence is 2.08%, while the mean output of the class-balanced LR is 23.43%. Its reliability curve lies below the identity line in most score strata, and the equal-frequency ECE is 0.214. Temporal sigmoid recalibration reduces the ECE to 0.019 for moderate degradation and 0.003 for severe degradation. The corresponding Brier scores decrease

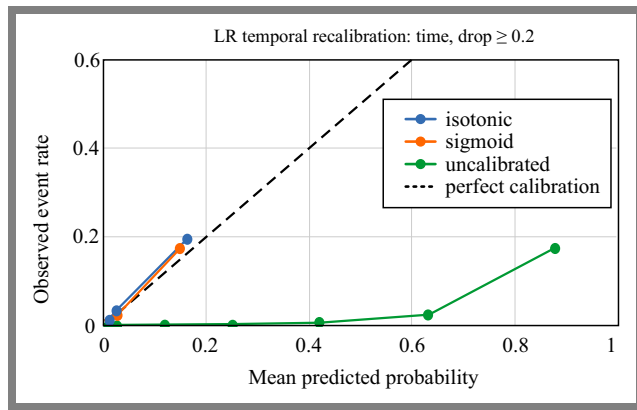


Fig. 5. Chronological reliability diagram for severe degradation before and after leakage-controlled calibration. The dashed diagonal denotes perfect calibration.

to 0.063 and 0.017. A similar improvement occurs in pooled LOGO predictions, although residual fold variation remains. Consequently, the uncalibrated output is suitable for ranking, whereas probability-based decisions require calibration data representative of the target gateway and operating period.

4.10. Gateway-observable Precursors

Within-gateway comparisons provide additional context for the learned risk function. Degradation windows are associated with lower packet counts and lower frequency diversity, with Cliff's delta values of -0.119 and -0.128 for moderate events and -0.230 and -0.179 for severe events. The current and lagged RSSI/SNR effects are smaller. The largest standardized association is the preceding CRC change, with delta values of 0.482 and 0.579 , indicating that many labeled drops follow a short-term reliability oscillation rather than a monotonic decline. Packet size differences may also reflect changes in traffic or device composition. These measurements characterize observable precursors – they cannot distinguish interference, collision, mobility, weather, or hardware mechanisms that are not recorded by LoED.

4.11. Explicit Sequence Modeling

Explicit use of 30 minutes of contiguous history produces only setting-dependent gains. On the common chronological sample, the GRU increases PR-AUC from 0.571 to 0.575 for moderate degradation and from 0.377 to 0.383 for severe degradation. Its LOGO mean increases from 0.418 to 0.424 for the moderate outcome, but decreases from 0.385 to 0.382 for the severe outcome. Sequence LR reaches the highest mean of severe event LOGO, 0.415 , while reducing moderate-event PR-AUC in both evaluation protocols. Neither sequence formulation consistently dominates the feature-based LR. The compact LR is therefore retained as the primary model because its discrimination is comparable across tasks and its inference path is simpler.

4.12. Offline Streaming Feasibility

The chronological replay processes 35 776 final test gateway windows in 16 277 timestamp batches. The mean preprocess-

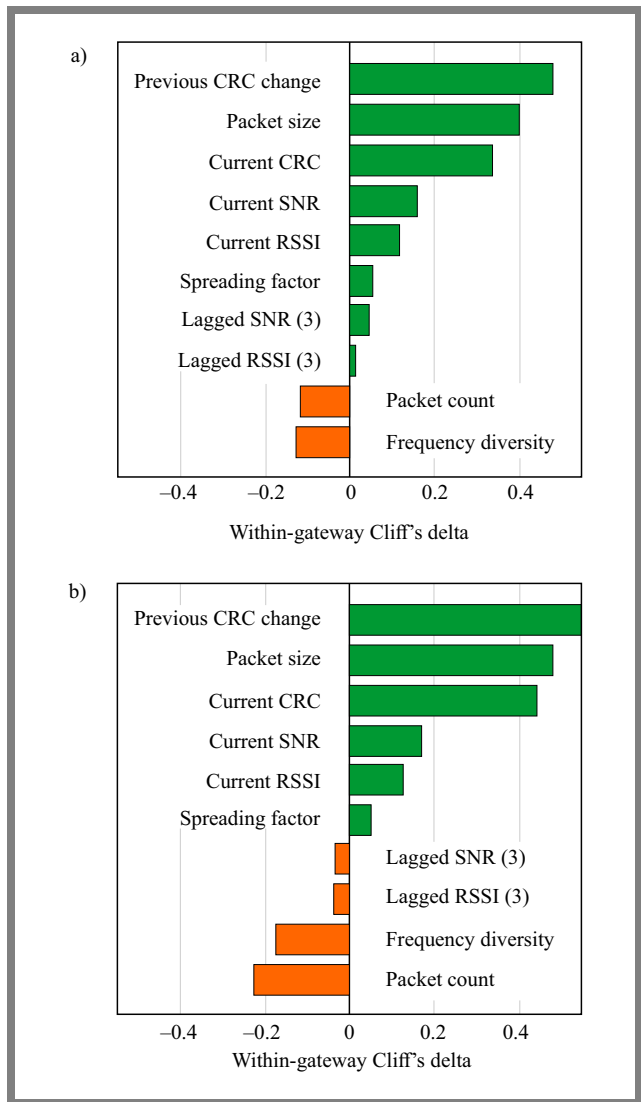


Fig. 6. Standardized precursor effects measured with the Cliff delta for drop threshold: a) 10% and b) 20%. Estimates are descriptive associations and do not identify physical causes.

ing, LR inference, and sigmoid calibration latency is 4.90 ms per batch for moderate degradation and 4.29 ms for severe degradation. The corresponding P99 values are 14.75 and 9.05 ms, respectively. A score cut-off selected from development data for a nominal 10% alarm budget yields test alert rates of 5.5% and 6.7% , capturing 35.2% and 73.4% of moderate and severe events. The scoring cost is negligible relative to a five minute interval on the evaluation computer, but the lower realized alarm rates expose temporal movement in the score distribution. Operational deployment would therefore require threshold monitoring in addition to efficient inference.

5. Discussion

Three observations emerge. First, a class-balanced linear model can rank short-horizon degradation risk within the LoED deployment using gateway-observable features, without payload inspection or end-device modification.

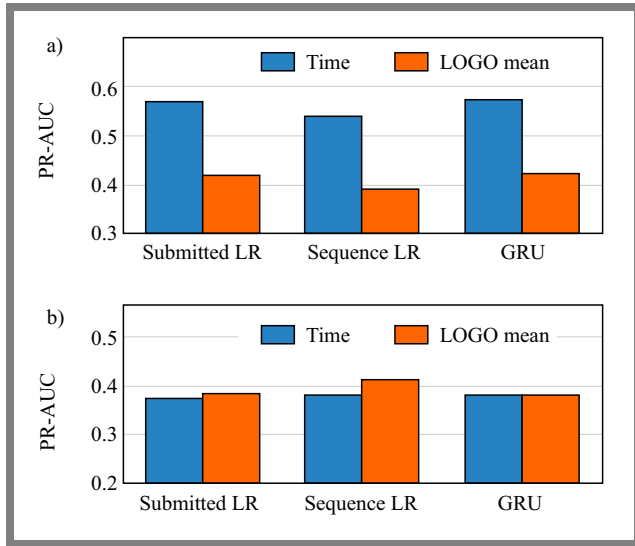


Fig. 7. PR-AUC comparison of the submitted feature-based LR, six window sequence LR and compact GRU on their common eligible samples. Drop threshold: a) 10% and b) 20%.

Second, the gap between chronological and LOGO evaluation shows that the mapping from metadata to degradation is deployment-specific. The model retains non-random ranking on unseen gateways, but large differences in prevalence, recall, and PR-AUC require local calibration, periodic threshold review, or gateway-specific operating points.

Third, alarm budget analysis is more informative than fixed threshold classification. Operators may not require a binary decision at a universal threshold. Instead, they may allocate an alarm budget, such as the top 5% or 10% highest risk windows, or choose an operating point that captures a target fraction of severe events. In this setting, a model with modest precision can still be valuable if it concentrates many events into a small set of windows.

From a system perspective, the proposed approach can be placed before the heavier adaptation mechanisms. A gateway or network server could score recent windows and raise a warning when the ranked risk exceeds a local operating point. The warning could trigger closer observation, temporary log, inspection of frequency/channel distribution, or application-level buffering. Therefore, the approach is best interpreted as an operational awareness component rather than as a closed-loop control policy.

The study has limitations. It reflects a single urban LoED deployment with nine gateways and cannot establish a generalization to other environments. Labels capture CRC success rate drops but do not identify their physical or network layer cause, and the observable precursor associations are descriptive rather than causal. Calibration improves under temporal holdout, but remains gateway-dependent. The streaming experiment is an offline replay that excludes packet aggregation and target gateway hardware. Finally, the small number of binary test gateways limits the power of corrected LOGO comparisons.

Although the replay supports the computational feasibility of window-level scoring, a live deployment would still need

to address raw packet aggregation, missing windows, model and threshold updates, persistent storage, and alarm fatigue. These issues define the next engineering step: prospective validation on gateway hardware with feedback from operators or automatic network management actions.

6. Conclusions

We presented a gateway-observable next-window warning formulation for LoRaWAN link degradation. Five minute LoED gateway windows were used to predict CRC success rate drops of 10% and 20%. LR achieved chronological PR-AUC values of 0.563 and 0.367 and exceeded RF and XGBoost in date-clustered randomization tests after correction for multiple comparisons. LR also produced the highest observed LOGO means, although the limited gateway count did not support a claim of statistical superiority.

A 10% chronological alarm budget captures 54.8% of moderate and 85.8% of severe events. CRC ablations show that the classifier does not simply reproduce current CRC, but depends substantially on short-term CRC history; radio-only information is weaker and transfers poorly. The five minute window duration offers a compromise between warning horizon, packet support, and gateway coverage. Leakage-safe sigmoid calibration improves the probability scale, while off-line replay demonstrates low window-level scoring latency and temporal alarm rate drift.

Therefore, gateway-observable metadata may support risk prioritization within deployments represented by LoED. It should not be interpreted as evidence of universal cross-deployment generalization or of a specific physical degradation mechanism. Prospective validation should include raw packet aggregation on target gateway hardware, independent deployments, periodic calibration monitoring, and integration with network management actions.

Acknowledgments

OpenAI ChatGPT and Codex were used to assist with language editing, manuscript restructuring, formatting, code organization, and implementation of supplementary analyses. The authors reviewed the generated code, reran the experiments, checked the numerical outputs, and remain responsible for the scientific design, interpretation, accuracy, and final content.

Data Availability

The LoED data set is publicly available through the source cited in the references. The preprocessing scripts, analysis code, and derived result tables associated with this study are available in a public reproducibility repository at <https://github.com/esma6/loed-lorawan-early-warning>.

References

- [1] LoRa Alliance, “LoRaWAN L2 1.0.4 Specification”, Technical Specification, 2020.
- [2] Semtech Corporation, “LoRa and LoRaWAN: A Technical Overview”, Technical White Paper, 2024 (<https://www.semtech.com/uploads/technology/LoRa/lorar-and-lorawan.pdf>).
- [3] F. Adelantado *et al.*, “Understanding the Limits of LoRaWAN”, *IEEE Communications Magazine*, vol. 55, pp. 34–40, 2017 (<https://doi.org/10.1109/MCOM.2017.1600613>).
- [4] N. Blenn and F. Kuipers, “LoRaWAN in the Wild: Measurements from The Things Network”, arXiv, 2017 (<https://doi.org/10.48550/arXiv.1706.03086>).
- [5] S. Li, U. Raza, and A. Khan, “How Agile is the Adaptive Data Rate Mechanism of LoRaWAN?”, *2018 IEEE Global Communications Conference*, Abu Dhabi, UAE, 2018 (<https://doi.org/10.1109/GLOCOM.2018.8647469>).
- [6] J. Finnegan, R. Farrell, and S. Brown, “Analysis and Enhancement of the LoRaWAN Adaptive Data Rate Scheme”, *IEEE Internet of Things Journal*, vol. 7, pp. 1717–1780, 2020 (<https://doi.org/10.1109/JIOT.2020.2982745>).
- [7] A. Hoeller, R.D. Souza, S. Montejó-Sánchez, and H. Alves, “Performance Analysis of Single-cell Adaptive Data Rate-enabled LoRaWAN”, *IEEE Wireless Communications Letters*, vol. 9, pp. 911–914, 2020 (<https://doi.org/10.1109/LWC.2020.2975604>).
- [8] L. Bhatia, M. Breza, R. Marfievici, and J.A. McCann, “LoED: The LoRaWAN at the Edge Dataset”, *Proc. of 3rd Workshop on Data Acquisition to Analysis*, pp. 7–8, 2020 (<https://doi.org/10.1145/3419016.3431491>).
- [9] P. Spadaccino, F.G. Crino, and F. Cuomo, “LoRaWAN Behaviour Analysis Through Dataset Traffic Investigation”, *Sensors*, vol. 22, art. no. 2470, 2022 (<https://doi.org/10.3390/s22072470>).
- [10] G. Cerar, H. Yetgin, M. Mohorcic, and C. Fortuna, “Machine Learning for Wireless Link Quality Estimation: A Survey”, *IEEE Communications Surveys & Tutorials*, vol. 23, pp. 696–728, 2021 (<https://doi.org/10.1109/COMST.2021.3053615>).
- [11] G. Cerar, H. Yetgin, M. Mohorcic, and C. Fortuna, “On Designing a Machine Learning Based Wireless Link Quality Classifier”, *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, 2020 (<https://doi.org/10.1109/PIMRC48278.2020.9217171>).
- [12] Y. Kanto and K. Watabe, “Wireless Link Quality Estimation Using LSTM Model”, *2024 IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Seoul, Republic of Korea, 2024 (<https://doi.org/10.1109/NOMS59830.2024.10575638>).
- [13] A. Prakash, N. Choudhury, A. Hazarika, and A. Gorrela, “Effective Feature Selection for Predicting Spreading Factor with ML in Large LoRaWAN-based Mobile IoT Networks”, *2025 National Conference on Communications*, New Delhi, India, 2025 (<https://doi.org/10.1109/NCC63735.2025.10983488>).
- [14] P. Maurya *et al.*, “A Comprehensive Survey of Data-driven Solutions for LoRaWAN: Challenges and Future Directions”, *ACM Transactions on Internet of Things*, vol. 6, art. no. 6, 2025 (<https://doi.org/10.1145/3711953>).
- [15] S. Hosseinzadeh *et al.*, “Explainable Machine Learning for LoRaWAN Link Budget Analysis and Modeling”, *Sensors*, vol. 24, art. no. 860, 2024 (<https://doi.org/10.3390/s24030860>).
- [16] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, “Wireless Network Intelligence at the Edge”, *Proceedings of the IEEE*, vol. 107, pp. 2204–2239, 2019 (<https://doi.org/10.1109/JPROC.2019.2941458>).
- [17] B.J. Bashar and H. Abdullah, “Analyzing Performance of Eigenvalue-based Spectrum Sensing within LoRaCog Framework”, *Journal of Telecommunications and Information Technology*, no. 2, pp. 67–74, 2026 (<https://doi.org/10.26636/jtit.2026.2.2576>).
- [18] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, *Proc. of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016 (<https://doi.org/10.1145/2939672.2939785>).
- [19] T. Saito and M. Rehmsmeier, “The Precision-recall Plot Is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets”, *PLOS ONE*, vol. 10, art. no. e0118432, 2015 (<https://doi.org/10.1371/journal.pone.0118432>).
- [20] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure”, *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979 (<https://doi.org/10.2307/4615733>).
- [21] N. Cliff, “Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions”, *Psychological Bulletin*, vol. 114, pp. 494–509, 1993 (<https://doi.org/10.1037/0033-2909.114.3.494>).
- [22] C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger, “On Calibration of Modern Neural Networks”, arXiv, 2017 (<https://doi.org/10.48550/arXiv.1706.04599>).
- [23] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation”, *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014 (<https://doi.org/10.3115/v1/D14-1179>).
- [24] G.W. Brier, “Verification of Forecasts Expressed in Terms of Probability”, *Monthly Weather Review*, vol. 78, pp. 1–3, 1950 ([https://doi.org/10.1175/1520-0493\(1950\)078<0001:V0FEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:V0FEIT>2.0.CO;2)).
- [25] A.C. Davison and D.V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge University Press, 1997 (<https://doi.org/10.1017/CB09780511802843>).

Esma Fazilet Karagülle, Ph.D. student

Department of Computer Engineering, Faculty of Engineering

 <https://orcid.org/0009-0002-7225-9970>

E-mail: esmafazilet.karagulle17@ogr.atauni.edu.tr

Atatürk University, Erzurum, Türkiye

<https://atauni.edu.tr>

Faruk Baturalp Günay, Ph.D., Assistant Professor

Department of Computer Engineering, Faculty of Engineering

 <https://orcid.org/0000-0001-5472-3608>

E-mail: baturalp@atauni.edu.tr

Atatürk University, Erzurum, Türkiye

<https://atauni.edu.tr>

Esra Odabaş Yıldırım, Ph.D., Assistant Professor

Department of Software Engineering, Faculty of Engineering

 <https://orcid.org/0000-0002-0936-1342>

E-mail: esra.odabas@atauni.edu.tr

Atatürk University, Erzurum, Türkiye

<https://atauni.edu.tr>