NovGrid: Novelty-Aware Smart-Grid Resilience

KMA Solaiman
University of Maryland, Baltimore County (UMBC)
ksolaima@umbc.edu

***Abstract*—Cyber–physical systems (CPS) such as smart grids increasingly encounter novel behaviors that differ from labeled anomalies and may disrupt operations if left unmanaged. We present a feasibility pipeline for handling such novelties in time-series consumption data, combining anomaly detection, predictive residual modeling, and a simple triage illustration. Our goal is to demonstrate how existing tools can be integrated end-to-end in a pipeline that highlights challenges and opportunities for novelty-aware CPS resilience.**

We demonstrate the pipeline on residential consumption data from the State Grid Corporation of China (SGCC), where we inject CPS-inspired novelty patterns (e.g., drift, spike, surge, flip, outage) to stress-test detectors. We adapt off-the-shelf baselines, including density-based methods (e.g., LOF) and a predictive baseline (ELSTM), showing how they capture different novelty regimes. Finally, we illustrate how outputs may be aggregated into a novelty characterization step to highlight potential operator-facing triage.

Our results suggest that this integration of passive detection, predictive residuals, and simple scoring can highlight tradeoffs between novelty regimes and motivate future, deeper studies of novelty-aware CPS resilience.

***Index Terms*—novelty detection, anomaly detection, out-of-distribution detection, cyber–physical systems, smart grid, resilience, time series, triage**

I. Introduction

Learning-enabled cyber–physical infrastructure (CPS) increasingly underpins modern electric grids: AMI/IoT endpoints, DERs, and automated SCADA workflows expand observability and control, but also enlarge the attack surface and amplify the cost of model mismatch. Traditional anomaly detection (AD) methods target *known, labeled risks* such as electricity theft, often using supervised classifiers [3], [2], [4]. However, these approaches often struggle when faced with *novel behaviors*—previously unseen behavior regimes that break training assumptions. Novelty arises in practice from faults, new devices, or adversarial campaigns, and misclassifying it as anomaly may trigger brittle decisions and unsafe grid operations.

The need for novelty-aware detection is underscored by real incidents. False-data injection attacks (FDIAs) against power-system state estimation [22] and the 2015 Ukraine blackout [19], [20] illustrate how attacker-driven novelties can bypass traditional monitoring [22] and how coordinated, SCADA-level intrusions can directly manipulate operations. Botnet-driven oscillations (BlackIoT, MaDIoT) [28], [21] demonstrate how high-wattage IoT loads can create entirely new demand trajectories. Even gradual drifts, such as HVAC degradation [13], manifest as ramp-like novelties in load curves. Smart-grid cyber-physical attacks [12], [14], [15], [17] stress that unseen load patterns are operationally disruptive, reinforcing the importance of separating novelty from anomaly.

Beyond explicit attacks, the distribution of consumption itself drifts: adoption of EVs and heat pumps, extreme-weather events, and changing demand-response programs can induce non-stationarity. A large body of work shows that detectors trained on historical data degrade under concept drift; effective systems must quantify drift, adapt thresholds, and reassess confidence over time [23], [24]. In time-series specifically, forecasting-based residuals are a well-established basis for anomaly or Out-of-Domain (OOD) scoring, precisely because they surface *model–environment misalignment* without requiring new labels [25].

Existing methods remain limited in their functionalities. Density-based approaches such as LOF [1] can flag outliers without labels but are essentially *passive*, lacking temporal resolution. Reconstruction/likelihood methods [8], [9], autoencoders and VAEs [25], [29], [30] still conflate rare anomalies with unseen novelties. Forecasting-based approaches including recent advances like the Anomaly Transformer [31], treat residual error as an OOD signal [11], but few have been contextualized for CPS resilience. All of these systems aim for only the detection module, still lacking from generating actionable insights for planning.

Operators, however, need *actionable* signals, not just raw outlier scores. Compliance frameworks (e.g., NERC CIP) emphasize auditable processes and proportional response: high-confidence events should trigger mitigation; low-confidence deviations warrant monitoring rather than disruptive control actions. That operational need translates, algorithmically, into: (i) separating *known, labeled* risks (e.g., theft) from *unknown, unlabeled* behaviors (novelty); (ii) combining passive density cues with predictive divergence signals to catch both structural outliers and emerging shifts; and (iii) *characterizing* novelty by structure and *tiering* by severity to support triage and planning [26].

To illustrate this gap, we assemble a novelty-aware pipeline for residential electricity consumption data in a smart-grid setting. The pipeline integrates three complementary signals: (1) supervised anomaly filtering for known theft cases; (2) **passive novelty detection** via LOF trained only on normal users; and (3) **predictive novelty** via an ELSTM baseline, where *residual error* serves as a label-free divergence score. We then cluster flagged novelties and assign illustrative severity tiers (mitigate,

*This version reflects the author's final, independently completed implementation and results.

review, ignore) as a simple triage step.

Contributions. In this work, we make the following contributions.

- We present a preliminary, end-to-end pipeline for novelty handling in smart grid CPS, combining anomaly detection, predictive residuals, and a simple triage illustration.
- We adapt representative disruptive patterns as injected novelties to stress-test detectors in residential consumption data.
- We show how a density-based baseline (LOF) and a predictive baseline (ELSTM) capture different aspects of novelty, and illustrate how a severity scoring step may support operator-facing triage.

This work aligns with the ARRL workshop themes: **Adaptable** – we illustrate how a predictive baseline (ELSTM residuals) can expose model–environment misalignments and respond to previously unseen shifts; **Reliable** – we demonstrate that standard density-based methods (e.g., LOF) and clustering can provide complementary structural signals under uncertainty; and **Responsible** – we include a proof-of-concept severity scoring step to illustrate how novelty flags might be translated into operator-facing triage that avoids unnecessary interventions.

II. Related Work

Smart-Grid Anomaly and Non-Technical Loss (NTL) Detection: Electricity-theft and consumption anomalies are commonly cast as supervised or semi-supervised classification under heavy class imbalance. Tree ensembles (e.g., XGBoost) and sequence encoders (LSTM/GRU) achieve strong accuracy on seen theft behaviors [3], [2], [4], with imbalance remedies such as SMOTE reducing false negatives. However, these detectors degrade under distribution shift [23] and attacker adaptation, limiting robustness in practice [36]. Unsupervised or weakly-supervised approaches (clustering, autoencoders) relax label dependence but still conflate rare anomalies with true novelties [8], [9], [29], [30].

Novelty and OOD detection in time series.: Classical density and neighborhood methods such as LOF [1] and isolation forests [7] remain standard baselines for novelty detection without labels. Recent surveys systematize three dominant families for time-series novelty and anomaly detection: reconstruction/likelihood methods (autoencoders, VAEs), predictive-residual methods, and association/attention-based models [29], [30], [25]. Forecast-based approaches in particular use residual error as an out-of-distribution signal (e.g., Uber [11]), and are widely applied in industrial monitoring. Transformer-based variants such as the Anomaly Transformer [31] extend this idea by comparing attention associations rather than raw residuals, but remain label-free and distributional in spirit.

Our work builds on this forecast-residual paradigm [11], but explicitly illustrates its adaptation to smart-grid CPS with engineered context features and thresholds, and include a downstream characterization step.

Cyber-Physical Security in Smart Grids: A growing body of literature studies cyberattacks and failures in energy systems. Sun *et al.* [13] showed that gradual equipment degradation produces detectable ramps, motivating predictive monitoring. Rajkumar *et al.* [14] and Zhang *et al.* [12] survey cascading failures and load-altering attacks (LAA), linking them to systemic risks. Alanazi *et al.* [15] analyzed load oscillation attacks, while Mateu-Barriendos *et al.* [16] proposed power oscillation damping controllers for mitigation. Alomari *et al.* [17] and Jha [18] classify smart grid threats under the CIA triad (availability, integrity, confidentiality) and outline incident response strategies. These works motivate our effort to link the injected novelties to practical faults and cyberattacks, grounding novelty detection in real-world operator actions.

From Detection to Characterization and Triage: Most anomaly and novelty detectors stop at *detection*, returning raw scores without operator guidance. For CPS, however, operators require actionable, auditable signals. Clustering methods (PCA/t-SNE + k-means) have been widely applied to structure embeddings for interpretability [5], [6]. Severity scoring and tiered thresholds have been proposed in statistical quality control (EWMA, CUSUM, POT) [33], [32], [34] to turn score into decisions, but are rarely integrated with novelty detection pipelines. Our approach shows how clustering of novelty candidates with a severity scoring method that fuses predictive divergence and structural distance, suggests operator-friendly triage signals. This bridges ML-based detection with planning/control frameworks such as CityLearn [27], [35], where decision-aware adaptation is essential.

III. Methodology

A. Pipeline Overview

We structure the study around a multi-stage pipeline that integrates standard components for anomaly filtering, novelty injection, detection, and post-processing. The purpose is to demonstrate how such elements can be combined end-to-end, rather than to propose new algorithms at each stage. The pipeline proceeds as follows:

1) **Supervised anomaly filtering:** known labeled anomalies are removed using standard classifiers to approximate realistic operating conditions where obvious risks have already been screened.
2) **Passive novelty detection:** after CPS-inspired novelty patterns are injected as controlled stress tests, density-based methods such as LOF are applied to identify structural deviations without temporal modeling.
3) **Predictive novelty detection and thresholding:** in the same controlled setting, an ELSTM baseline is trained on normal behavior and absolute residual errors are used to illustrate how forecasting divergence can signal novelty. *Statistical rules* EWMA-based adaptive thresholds (per-user) and global percentiles (cold-start) - convert continuous scores into novelty flags.
4) **Characterization and triage:** flagged novelties are clustered into coarse families and assigned illustrative severity scores to make outputs more interpretable for operators.

This overview emphasizes the feasibility of integrating diverse techniques into a single workflow. The staged design

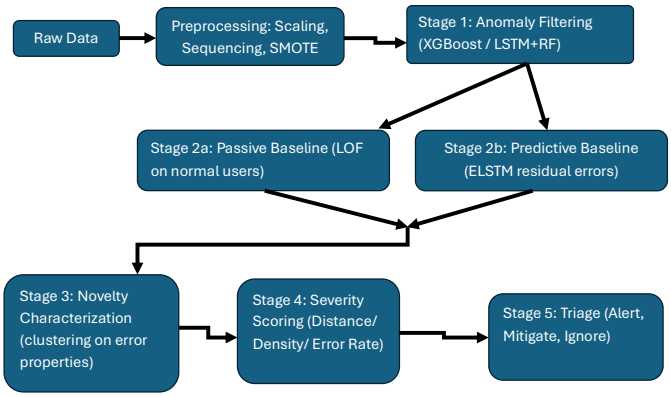


Fig. 1: End-to-end novelty-aware smart grid pipeline.

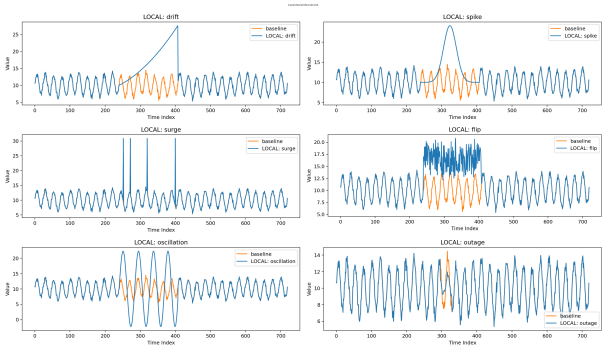


Fig. 2: Injected novelty families (drift, spike, surge, flip, oscillation, outage).

aims to (i) screen known, labeled threats, (ii) surface previously unseen behaviors (e.g., regime shifts from EV adoption or persistent equipment faults), and (iii) provide interpretable signals that can support operator action. It aligns with ARRL goals of *adaptability* (the ELSTM residual baseline highlights shifts), *reliability* (LOF and clustering offer complementary structural cues), and *responsibility* (the severity-scoring illustration encourages proportional responses).

B. Stage 1 — Supervised Anomaly Filtering

Before testing novelty detection, we apply a standard supervised anomaly filter to remove known labeled risks (`FLAG=1`) from the training pool. This ensures novelty detectors operate on realistic conditions where labeled anomalies are already screened. We treat this step purely as preprocessing, using the anomaly labels provided in the SGCC dataset.

We frame this as a supervised classification task and implement:

- **XGBoost**: a gradient boosting classifier applied to aggregated features of daily consumption. Gradient boosting is widely used in smart grid intrusion and fraud detection because of its robustness to heterogeneous features and class imbalance.

This module could be used as a pre-processor for the later novelty-aware components.

C. Novelty Injection Design

Novelty vs. anomaly: In this work, we formally separate novelty from anomaly. We define novelty as the sudden emergence of a short-lived behavior with no repetition (or extremely rare) across the observation period. Anomalies, in contrast, represent known, recurrent deviations (e.g., electricity theft) that exist in the labeled training distribution.

However, evaluating detection performance on a single instance is not reproducible. To bridge this gap, we define families of synthetic novelties which approximate categories of unseen behavior, and individual instances are injected once per trace. This preserves the "unseen" character of novelty while allowing systematic and repeatable evaluation across the dataset.

To demonstrate feasibility of detectors in our pipeline, we focus on following representative injected patterns:

- **Drift:** A slow upward or downward slide in the baseline load, mimicing long-horizon sensor drift or gradual efficiency loss in equipment; corresponds to permanent mean/slope changes [37].
- **Spike:** A sudden, isolated jump in demand embedded within otherwise stable traces, representing transient device faults or single-event anomalies.
- **Surge:** A sequence punctuated by irregular bursts, modeling erratic load injections that resemble misconfigured timers or sporadic demand spikes; captured via *step-like* and *slope-like* trends [39].
- **Flip:** Segments whose temporal order is distorted or reversed, designed to emulate replay or spoofing attacks that insert misleading historical values; sign-flip is a standard augmentation primitive [39].
- **Oscillation:** Repetitive swings of varying frequency and amplitude, echoing unstable control loops or cyclic demand patterns that stress grid balancing; matches *seasonal/periodic* behavior [37].
- **Outage:** Intervals where normal activity is abruptly suppressed or silenced, reflecting partial blackouts, rolling brownouts, or load-shedding scenarios.

Each injected pattern is parameterized by duration, amplitude, and starting day, with values sampled from reasonable ranges to generate varied instances. These are not intended as a comprehensive taxonomy; rather, they serve as illustrative stress-tests that allow us to observe how passive density-based detectors and predictive residuals respond under controlled novelty regimes. Our preliminary work introduced novelty with different injected patterns at the user consumption level. Although that behavior was referred to as "novelty," it more closely aligned with the anomaly class, since those consumption patterns were fully divergent from normal consumption data. A fuller classification of novelty types across datasets is left for future work.

D. Stage 2a — Passive Novelty Detection (LOF)

Once novelty patterns are injected, the next step is to evaluate whether unsupervised detectors can flag them without access to labels. We adopt the Local Outlier Factor (LOF) [1], a widely used density-based anomaly detector, as a baseline for *passive* novelty detection. LOF measures the relative density of each data point compared to its neighbors in feature space, assigning higher scores to points that lie in sparse regions. Unlike supervised classifiers, LOF does not require anomaly labels and thus reflects a realistic setting for detecting unknown behaviors.

In our pipeline, LOF is applied as-is after a PCA projection of daily consumption sequences (used only to stabilize distances). It is trained on users marked as normal (`FLAG=0`) and then evaluated on both held-out normal and novelty-injected users. We report LOF scores as a passive reference, where high scores indicate structural deviations from the normal training distribution.

As expected, LOF highlights novelties with sharp structural deviations (e.g., reversals or sparse bursts), but its lack of temporal modeling makes it fundamentally passive: it can identify unusual windows but cannot align novelties in time. In the pipeline, this serves to illustrate the limitations of passive density methods, motivating the use of predictive residuals in Stage 2b.

E. Stage 2b — Predictive Baseline (ELSTM Residuals)

To complement passive density methods, we **illustrate the use of a predictive residual model** as a baseline for novelty detection. The idea is straightforward: if a model trained only on normal behavior fails to forecast future consumption, its errors can serve as a *signal of novelty*.

We adapt a stacked LSTM architecture with sequence-to-one prediction. Input windows consist of **7 consecutive days of features** (daily consumption values plus calendar/temporal indicators), and the target is the next day's load. The network includes a single LSTM layer, followed by a dense projection layer and a regression output node. Training is performed only on users labeled as normal (`FLAG=0`), using mean squared error (MSE) loss.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

At inference, we compute the absolute prediction residual for each day:

$$e_t = |y_t - \hat{y}_t|$$

Thresholds are derived from residuals on held-out normal users, without using novelty labels, ensuring the procedure remains unsupervised. Larger residuals indicate stronger deviation from the training distribution. To convert residuals into novelty flags, we apply two simple thresholding strategies:

1) **EWMA-based adaptive threshold (per-user):** Exponentially Weighted Moving Average (EWMA) adapts threshold θ_t online while assigning higher weight to recent values, following:

$$\mu_t = \alpha e_t + (1-\alpha)\mu_{t-1} \quad (1)$$

$$\sigma_t^2 = \alpha(e_t - \mu_{t-1})^2 + (1-\alpha)\sigma_{t-1}^2 \quad (2)$$

$$\theta_t = \mu_t + L \cdot \sigma_t \quad (3)$$

2) **Global percentile threshold (cold-start):** a global pooled threshold is applied for assigning novelty flag when user-specific history is insufficient.

$$\theta_{\text{global}} = \text{Percentile}_{95}\left(\{e_t\}_{\text{train}}\right)$$

These thresholds illustrate how predictive errors can be turned into novelty flags. In practice, the ELSTM residual baseline is sensitive to regime shifts (global novelties) that persist over multiple days, but less effective at capturing short, sharp deviations (localized novelties).

Within the pipeline, this stage serves as an **illustration of predictive modeling for novelty scoring**. More advanced novelty-aware predictive approaches remain a subject for future work.

Predictive divergence examples are shown in Fig. 4, averaged over users for a representative novelty (ramp). Per-user sweeps across all injected patterns are shown in Fig. 3a, while aggregated divergence across novelty families remains above known anomalies (Fig. 5).

F. Novelty Characterization and Triage

Detection alone is not sufficient for operators, who require signals that can be interpreted and acted upon. To illustrate how novelty flags might be made more actionable, we include a simple characterization and scoring step within the pipeline.

First, windows flagged as novel are embedded into a two-dimensional space using PCA for stability followed by t-SNE for visualization. K-Means clustering is then applied to group samples into coarse structural families. This unsupervised grouping is not intended as a definitive taxonomy, but rather as an example of how structure can be surfaced from detector outputs.

In parallel, each flagged window is assigned a severity score that combines predictive residual magnitude with cluster centroid distance:

$$s = \alpha\, \text{z}(d) + (1-\alpha)\, \text{z}(\text{dist}_{\text{cluster}}), \quad \alpha \in [0,1], \quad (4)$$

where $z(.)$ denotes standardized values, d is the predictive error, and $dist_{cluster}$ is the Euclidean distance to the assigned cluster center.

Severity scores are then mapped into illustrative tiers: High, Medium, and Low. These tiers are presented as one possible way of aligning novelty outputs with operator-facing responses such as mitigation, manual review, or dismissal as benign drift.

This stage is included as a proof-of-concept illustration. It highlights how raw novelty scores may be transformed into interpretable triage signals. More systematic severity design and validation across real operational contexts is left for future work.

IV. Experimental Setting and Results

Dataset. We use the SGCC smart-grid dataset (2014–2016) with 1,037 users and 42,372 daily records. Each user has a consumption time series; a binary flag `FLAG=1` marks electricity theft. Normal data have `FLAG=0`. For novelty, we train on FLAG=0 users and evaluate on both normal and novelty-injected windows from these users. For efficiency, windows are partitioned per user to prevent overlap across training and evaluation.

Metrics. For anomaly baselines: Accuracy, Recall, Precision, F1, AUROC. For novelty: detection rate above threshold, and divergence margins vs. known anomalies. For characterization: cluster purity w.r.t. injected labels and severity separation.

Injected novelties. We generate six novelty families: Drift, Spike, Surge, Flip, Oscillation, Outage. Injection length and amplitude are randomized within realistic bounds; alignment to daytime hours preserves plausibility.

For stage 1, class imbalance was addressed using SMOTE-based resampling, and daily records with missing values were discarded. These preprocessing steps are applied only within the anomaly filtering module and do not affect subsequent novelty detection stages. We summarize the settings of our local/consumption novelty detection experiments below:

- **Sliding Window:** 7-day window length ($W = 7$).
- **Per-day Features:** 5 features per day: consumption, holiday flag, day-of-week (scaled), 3-day rolling mean, and day-to-day change ratio.
- **Model:** LSTM(64) + Dense(64, relu) + Dense(1). Optimizer Adam, loss MSE. Early stopping with patience=3. Input shape: (7,5).
- **Training Data:** Normal users only (FLAG=0). Max 30 windows per user. Saved as numpy arrays for efficiency.
- **Novelty Injection:** Injected into 5 random normal users, 30-day segment starting at day 100.
- **Thresholds:** Global pooled baselines from normal users. Threshold include: 95th percentile, mean+k·std, and EWMA (per-user). For local novelty detection, per-user IQR and 95th percentile were most effective, while EMWA provided adaptive and tail-sensitive alternatives.
- **Baseline Error:** Normal baseline prediction error distribution computed across all non-injected normal users.

A. Anomaly Detection Baselines

We first establish supervised anomaly detection baselines on SGCC theft labels (`FLAG=1`). Table I and II summarize results for XGBoost. XGBoost achieves moderate performance with recall of 42% but with low precision. On the SGCC test set, our xgboost model achieved an overall ROC–AUC of 0.80. These results shows XGBoost is a viable anomaly identifier, but is not the best for preprocessing anomaly.

TABLE I: Performance of XGBoost on Anomaly (Theft) class at decision threshold of 0.5.

Model	Recall	Precision	F1
XGBoost (Theft)	0.42	0.36	0.39

TABLE II: Class independent performance on SGCC at decision threshold of 0.5.

Model	Weighted Acc.	Macro Acc	Macro F1	AUROC
XGBoost	0.89	0.65	0.66	0.80

B. Passive Novelty via LOF

To go beyond labeled anomalies, we evaluate Local Outlier Factor (LOF) as an unsupervised *passive novelty detector*. LOF is trained only on normal users (`FLAG=0`) and then exposed to the novelty injections. As shown in Table III, LOF achieves low recall (21%) but low precision (35%), reflecting somewhat sensitivity to structural shifts but a tendency to over-flag. This makes LOF valuable for exploratory triage, but insufficient for precise or predictive detection.

Importantly, LOF remains a *passive* detector: it can identify whether a sequence looks unusual, but it does not perform windowed prediction or provide time-localized divergence scores. Thus, while useful for localized density anomalies, LOF cannot anticipate or quantify novelty in an active, predictive sense. This motivates our transition to sequence-based predictive divergence (ELSTM) for active novelty modeling.

C. Novelty Detection with Predictive Baseline

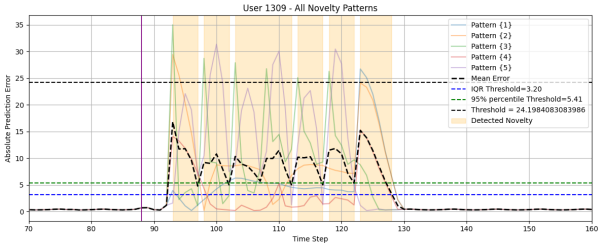
To further analyze how predictive divergence manifests under different novelty types and user behaviors, we structure the ELSTM results into three complementary lenses. First, we illustrate representative *case studies* (per-user and cross-user) to show how injected novelties emerge in time-aligned error trajectories. Second, we examine *threshold robustness*, contrasting global cutoffs with adaptive per-user thresholds. Finally, illustrate how predictive residuals behave across novelty families, highlighting feasibility of separating long vs. short novelties. Together, these analyses ground the predictive divergence signal both in visual case studies and in quantitative detection behavior.

1) Case Study: Per-User and Cross-User Novelty Profiles: To better understand how predictive divergence manifests across different novelty types and users, we analyze two complementary views.

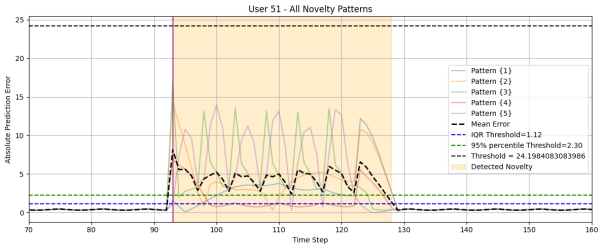
Per-user, all novelties. Figures 3a and 3b show two representative users (1309 and 51) with five novelty families injected. Within a single user baseline, each novelty produces a distinct error trajectory: ramps yield gradual elevation, bells produce sharp symmetric peaks, bursts generate jagged high spikes, reversals create sustained shifts, and sines induce oscillatory divergence. The mean error (black line) smooths across these but still rises above baseline during novelty intervals. Adaptive per-user thresholds (IQR, 95th percentile) scale to each user's variability and detect all novelties; in contrast, a static global mean+std cutoff (≈24.2) fails to flag subtle novelties (ramps, reversals). This demonstrates the necessity of per-user thresholding in heterogeneous populations.

TABLE III: Novelty detection performance of LOF.

Metric	Mean ± Std
Recall	0.205 ± 0.295
Precision	0.352 ± 0.427
F1	0.239 ± 0.323
AUROC	0.749 ± 0.190



(a) User 1309: all novelty families



(b) User 51: all novelty families

Fig. 3: Per-user case studies (Users 1309, 51) under five novelty families; adaptive thresholds outperform static global cutoffs.

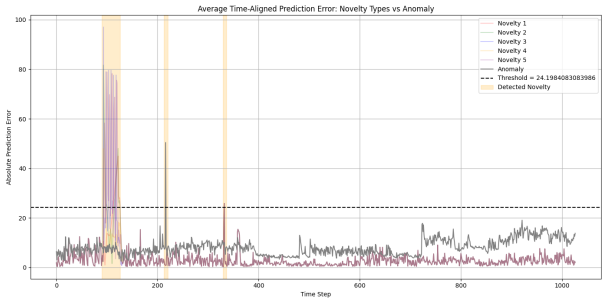


Fig. 4: Cross-user predictive divergence: novelties yield higher, structurally distinct errors than anomalies. Anomaly baseline shown in gray.

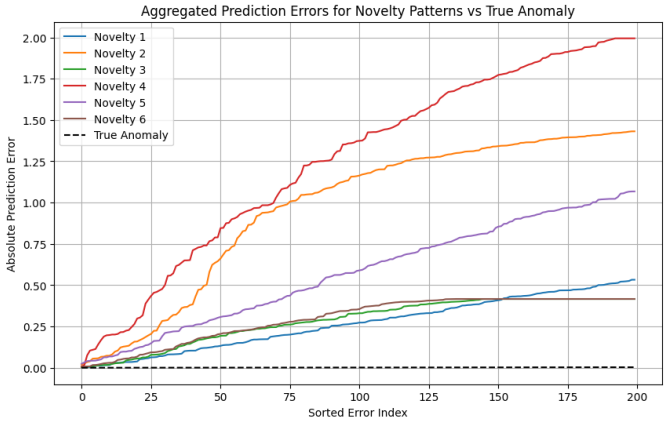


Fig. 5: Sorted prediction errors: novelties consistently exceed anomaly baseline.

Per-novelty, all users. Complementary results are shown in Figure 4, which plots average time-aligned error across five users for each novelty family. Here, every novelty type elicits spikes consistently above user baselines, while anomaly traces remain much lower. The error signatures differ across novelty families—ramps show smooth rise, bursts/sines show oscillations, reversals sustain elevated error—providing structural cues that enable characterization. This cross-user view confirms that ELSTM divergence generalizes across the population, not just individual cases.

Together, these case studies reinforce two points: (i) novelty types are structurally separable within a single user's trajectory, and (ii) they remain consistently detectable across multiple users when using adaptive per-user thresholds. This indicates that ELSTM residuals can yield discriminative, interpretable novelty signals in this dataset.

2) Error as a metric: Before analyzing thresholds under novelty, we verify that normal users remain well below detection cutoffs. Figure 3 shows that pooled normal errors almost never exceed per-user IQR or 95th percentile lines, ensuring low false positives.

To complement time-aligned views, Figure 5 plots sorted absolute prediction errors for each novelty family compared to true anomalies. Novelties are consistently shifted upward across the distribution, confirming separation beyond isolated spikes. In particular, Novelty 4 shows the strongest divergence, while Novelties 1 and 3 are closer to anomalies but still remain above the anomaly baseline. This distributional evidence reinforces that ELSTM errors encode structural novelty as a global shift, not just transient excursions. These results provide the specificity required to trust ELSTM divergence as a novelty signal.

Finally, we contrast novelty duration. Structured, long novelties (ramps, reversals, sines) consistently exceed thresholds across users, while short or sparse novelties (isolated bursts) often evade detection unless adaptive thresholds (IQR/EWMA) are used. This highlights the trade-off: structured drift is easy to catch, but sparse anomalies blend into natural variability.

3) Threshold Comparisons: We evaluated multiple thresholding strategies. Global pooled thresholds (mean+std) provided a static baseline, effective for cold start. Per-user thresholds (IQR and 95th percentile) adapted better to local dynamics, sharply increasing sensitivity to short-lived novelties. Figure 3 shows that the global cutoffs are not adaptable enough for individual consumption patterns of each user. Hence, we tried adaptive per-user thresholds with EWMA thresholds,

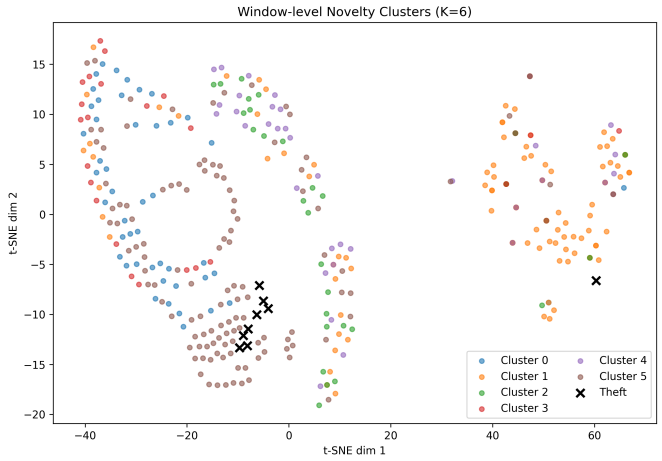


Fig. 6: t-SNE of LOF-flagged windows clustered via KMeans. Two distinct novelty families emerge in cluster 1 and cluster 5, though LOF alone struggles to perfectly separate anomalies from novelties.

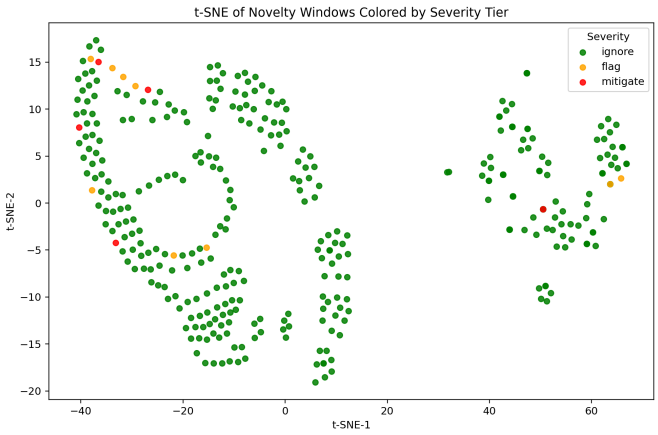


Fig. 7: t-SNE colored by severity tiers derived from divergence and cluster distance. High, medium, and low tier severe novelties emerge in the subspace.

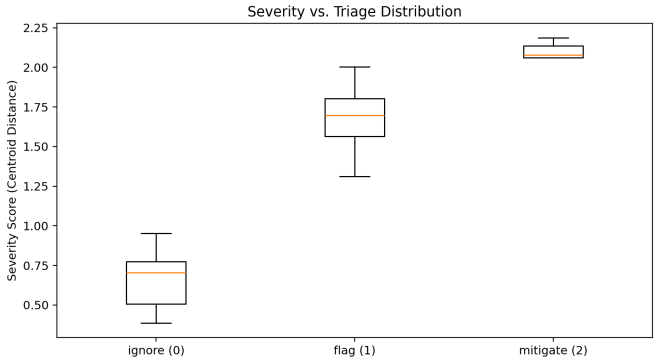


Fig. 8: Severity tiers vs. centroid distance. High, medium, and low tiers form statistically distinct groups with minimal overlap.

which adapted to gradual drifts in real time.

D. Characterization and Severity Results

Clustering analysis revealed that **LOF-flagged novelties** consistently separate into few distinct interpretable groups. These structural families align with respective novelty families and provide interpretable signals beyond binary novelty flags.

Severity scoring, as shown in Figures 7 and 8, produced a clear stratification across tiers. Although Figure 7 shows some overlap of points across tiers, this is expected since novelty severity is shaped not only by distance but also by underlying behavioral properties. The boxplot analysis in Figure 8 confirms that these tiers are statistically distinct with minimal overlap. As a feasibility test, we instantiated Eq. 4 with $\alpha = 0$, i.e., using only the cluster distance from the centroid among LOF-flagged novelties. Based on a percentile-based heuristic, novelties above the 85th percentile were categorized as *high risk*, warranting immediate operator intervention; those between the 75th and 85th percentiles as *medium risk*, requiring operator review and flagging; and the remainder as *low risk*, which can be safely ignored to reduce alarm fatigue.

Importantly, Figure 6 highlights that LOF alone struggles to perfectly separate anomaly (theft) and novelty windows, reinforcing the need for active predictive divergence (ELSTM). Novelty characterization illustrates how distinct structures may be surfaced, and how simple severity tiers could support operator triage in principle.

V. Discussion

Our findings highlight several implications for novelty-aware CPS monitoring. First, predictive divergence with ELSTM consistently separates injected novelties from known anomalies, suggesting that residual-based scores can serve as reliable, label-free signals in heterogeneous user populations. Second, the complementarity between LOF and ELSTM is noteworthy: while LOF achieves high recall on structural deviations without labels, ELSTM provides temporal resolution and reveals when novelty episodes emerge and persist. Third, The injected novelty families (ramps, reversals, oscillations, bursts) illustrate how structural patterns can be made interpretable and mapped into operator-facing signals. Finally, the contrast between long, structured novelties (readily detected) and short, sparse ones (often missed) underscores a key open challenge for future CPS monitoring systems.

Besides the ELSTM baseline, we conducted preliminary tests with Ridge regression, non-sequence LSTM, and global novelty injections. These baselines underperformed, with limited novelty detection capacity, hence we do not include them in full results.

A. Mapping to ARRL Themes

The staged pipeline can be interpreted in light of the ARRL workshop themes.

Adaptable. The ELSTM residual baseline highlights when model–environment misalignments occur without labels, indicating sensitivity to previously unseen shifts such as ramps or reversals. Results on both localized and global injections confirm that per-user adaptive thresholds (IQR, EWMA) can adjust online after a cold-start, offering pathways toward human-in-the-loop or feedback-driven adaptation.

Reliable. LOF complements ELSTM by robustly detecting localized novelties. Robust thresholds (per-user EWMA) and density cues reduce brittleness to heavy tails and user heterogeneity.

Responsible. Severity scoring tiers the introduced novelty families into actionable operator responses, aligning with operator protocols (e.g., NERC CIP) and avoiding unnecessary interventions, such as: maintenance calls for gradual ramps [13], inspections for HVAC malfunctions [14], or incident response protocols under the CIA triad for cyber attacks [17], [12]. This strengthens the alignment with responsible, trustworthy CPS design. We also emphasize transparency: severity scores are auditable; novelty families are interpretable.

B. Limitations and Future Work.

Our evaluation is scoped to the SGCC dataset with synthetic novelty injections, which, while CPS-grounded, cannot fully represent the diversity of real faults and attacks. Short, sparse novelties also remain harder to capture reliably than structured, long-duration shifts.

Going forward, we plan to broaden evaluation across multiple datasets and novelty families, report quantitative benchmarks (recall, FPR, divergence margins) beyond the case studies, and explore other adaptive thresholding methods (CUSUM, POT) for online deployment. An important direction is linking ELSTM outputs with severity scoring and control actions, enabling utility-facing demonstrations in live operational settings.

VI. Conclusion

We presented a feasibility pipeline for novelty-aware monitoring in cyber–physical systems, demonstrated on the SGCC smart-grid dataset. The pipeline integrates supervised anomaly filtering, CPS-inspired novelty injection, passive and predictive detection baselines, simple thresholding, and an illustrative triage step.

Our findings highlight how residual-based models can surface distributional shifts, how density-based methods capture structural deviations, and how simple severity scoring can suggest operator-facing responses. Together, these stages illustrate one possible pathway for integrating existing tools toward novelty-aware CPS resilience.

Future directions include testing across multiple datasets, refining severity design, and developing advanced predictive and structural detectors. By framing novelty handling as a staged pipeline, we aim to open avenues for systematic investigation of resilience in smart grids and other critical CPS domains.

References

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. SIGMOD*, 2000.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016.

[4] N. V. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *JAIR*, 2002.

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, 2008.

[6] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Info. Theory*, 1982.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. ICDM*, 2008.

[8] P. Vincent *et al.*, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, 2008.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.

[10] J. Zico Kolter and M. A. Johnson, "REDD: A public data set for energy disaggregation research," in *Proc. SustKDD*, 2011.

[11] N. Laptev, S. Amizadeh, and I. Flint, "Time-series extreme event forecasting with neural networks," in *Proc. KDD*, 2017.

[12] H. Zhang *et al.*, "Smart Grid Cyber-Physical Attack and Defense: A Review," *IEEE Access*, vol. 9, pp. 29641–29659, 2021.

[13] Z. Sun *et al.*, "Gradual Fault Early Stage Diagnosis for Air Source Heat Pump System Using Deep Learning Techniques," *Int. J. Refrigeration*, vol. 107, pp. 63–72, 2019.

[14] V. S. Rajkumar *et al.*, "Cyber Attacks on Power Grids: Causes and Propagation of Cascading Failures," *IEEE Access*, vol. 11, pp. 103154–103176, 2023.

[15] F. Alanazi *et al.*, "Load Oscillating Attacks of Smart Grids: Vulnerability Analysis," *IEEE Access*, vol. 11, pp. 36538–36549, 2023.

[16] E. Mateu-Barriendos *et al.*, "Power Oscillation Damping Controllers for Grid-Forming Power Converters in Modern Power Systems," arXiv:2409.10726, 2024.

[17] M. A. Alomari *et al.*, "Security of Smart Grid: Cybersecurity Issues, Potential Cyberattacks, Major Incidents, and Future Directions," *Energies*, vol. 18, no. 1, 141, 2025.

[18] R. K. Jha, "Cybersecurity and Confidentiality in Smart Grid for Enhancing Sustainability and Reliability," *Recent Research Reviews Journal*, vol. 2, no. 2, pp. 215–241, 2023.

[19] Cybersecurity and Infrastructure Security Agency (CISA), "Cyber-Attack Against Ukrainian Critical Infrastructure (IR-ALERT-H-16-056-01)," Dec. 2015. [Online]. Available: https://www.cisa.gov/news-events/ics-alerts/ir-alert-h-16-056-01

[20] R. M. Lee, M. J. Assante, and T. Conway, "Analysis of the Cyber Attack on the Ukrainian Power Grid," E-ISAC / SANS, 2016. [Online]. Available: https://nsarchive.gwu.edu/sites/default/files/documents/3891751/SANS-and-Electricity-Information-Sharing-and.pdf

[21] T. Shekari, S. M. Vedaei, A. Tajalli, A. Keshavarz, H. Mohsenian-Rad, and A. Mosenia, "MaDIoT 2.0: Modern High-Wattage IoT Botnet Attacks and Defenses in Power Grids," in *Proc. USENIX Security*, 2022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/shekari

[22] Y. Liu, P. Ning, and M. K. Reiter, "False Data Injection Attacks against State Estimation in Electric Power Grids," in *Proc. ACM CCS*, 2009, pp. 21–32. [Online]. Available: https://dl.acm.org/doi/10.1145/1653662.1653666

[23] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1–44:37, 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2523813

[24] F. Hinder, C. Webb, and S. M. Herzog, "One or Two Things We Know About Concept Drift—A Survey of Unsupervised Drift Detection," *Frontiers in Artificial Intelligence*, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2024.1330257/full

[25] R. Chalapathy and S. Chawla (eds., survey compilers), "Deep Learning for Time Series Anomaly Detection: A Survey," arXiv:2211.05244, v3, 2024. [Online]. Available: https://arxiv.org/html/2211.05244v3

[26] Industrial Defender, "What Is NERC CIP: The Ultimate Guide," 2025. [Online]. Available: https://www.industrialdefender.com/blog/what-is-nerc-cip

[27] J. R. Vázquez-Canteli, S. Dey, G. Henze, and Z. Nagy, "The CityLearn Challenge 2020," in *Proc. BuildSys '20*, ACM, 2020, pp. 320–321. [Online]. Available: https://doi.org/10.1145/3408308.3431122

[28] S. Soltan, P. Mittal, and H. V. Poor, "BlackIoT: IoT Botnet of high wattage devices can disrupt the power grid," in *Proc. USENIX Security*, 2018.

[29] Z. Z. Darban, M. Talebi, A. C. Lozano, and M. Ibrishimov, "Deep learning for time series anomaly detection: A survey," *ACM Comput. Surv.*, 2024.

[30] F. Wang, X. Zhang, and Q. Liu, "A survey of deep anomaly detection in multivariate time series," *Sensors*, vol. 25, no. 1, 2025.

[31] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," *NeurIPS*, 2021.

[32] NIST/SEMATECH, "6.3.2.4 EWMA control charts," NIST/SEMATECH e-Handbook of Statistical Methods, 2024. [Online]. Available: https://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm

[33] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[34] S. Solari, F. Egüen, and F. Losada, "Peaks over threshold: A methodology for automatic threshold estimation," *Water Resources Research*, 2017.

[35] K. Nweye, A. Wu, H. Park, Y. Almilaify, and Z. Nagy, "CityLearn v2: Energy-flexible, resilient, occupant-centric, and fair," arXiv:2405.03848, 2024.

[36] J. E. Zhang, A. M. Brahma, and N. Kandasamy, "Time series anomaly detection for smart grids: A survey," arXiv:2107.08835, 2021.

[37] Z. Darban, S. Tuli, G. Casale, and N. R. Jennings. Time-Series Anomaly Detection: A Survey. *ACM Computing Surveys*, 2022. Available at https://arxiv.org/abs/2211.05244.

[38] P. Wenig and T. Papenbrock. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. *Proceedings of the VLDB Endowment*, 12(12):1762–1777, 2019. DOI: https://doi.org/10.14778/3352063.3352069.

[39] P. Wenig and T. Papenbrock. Time Series Data Augmentation for Deep Learning: A Survey. *arXiv preprint*, arXiv:2109.04311, 2022. Available at https://arxiv.org/abs/2109.04311.

[40] S. Schmidl, P. Wenig, and T. Papenbrock. ADBench: Anomaly Detection Benchmark. *Proceedings of the VLDB Endowment*, 15(8):1779–1792, 2022. Available at https://www.vldb.org/pvldb/vol15/p1779-wenig.pdf.