POWERED III: Predictive Modeling of Transformer Outages Using Power Quality Data

Andrew Strelzoff, Sanja Cvijic
Charles River Analytics
Cambridge, Massachusetts, MA
{astrelzoff, scvijic}@cra.com

Nathan Peterson, Harold Sanborn
CERL/ERDC Army
Champaign, IL
{nathan.g.peterson, harold.sanborn}@usace.army.mil

***Abstract*—This paper presents POWERED III (Probabilistic Operations Warranted for Energy Reliability Evaluation and Diagnostics), a research effort that develops advanced Condition-Based Maintenance Plus (CBM+) modeling techniques for power infrastructure. The project leverages 18 months of power quality (PQ) trend data from PQube monitoring systems deployed at 25 substations, comprising distribution and sub-transmission assets including STATCOM installations—to predict grid-induced service interruptions. Using multi-window feature engineering (3-, 7-, 14-, and 21-day horizons) and gradient-boosted decision trees with calibrated probability outputs, the model achieves an AUC-ROC of 0.78, detecting approximately 80% of interruptions with minimal false positives. Calibrated probabilities exhibit characteristic buildup patterns over 3–21 days before events, providing actionable early warning for operational intervention. Post-hoc validation with utility engineers confirmed that predicted events represent protective shutdowns triggered by evolving grid stress conditions rather than scheduled maintenance or internal equipment failure. The system includes an operator-facing dashboard for real-time risk visualization. This work demonstrates practical integration of physics-informed feature engineering with machine learning for predictive maintenance, extending traditional CBM+ approaches from internal asset monitoring to grid-side power quality analysis.**

***Index Terms*-- power transformers, predictive maintenance, condition-based maintenance, power quality, machine learning, gradient boosting, early warning systems.**

I. Introduction

Reliable grid operation depends on detecting not only internal transformer degradation but also externally observable stress patterns in the surrounding power network. While traditional Condition-Based Maintenance (CBM+) focuses on internal diagnostics—dissolved gas analysis, thermal monitoring, and loading history—grid-induced protective shutdowns represent a distinct failure mode triggered by external power quality disturbances. These events, caused by voltage instability, or frequency impulses propagating through the network, may be predictable from grid-side observables days or weeks in advance.

Recent research demonstrates that data-driven approaches can capture degradation patterns from operational measurements. Zeng and Chu [1] employed a spatio-temporal graph neural network to model multivariate SCADA data and forecast transformer health indicators, showing that inter-sensor dynamics reveal evolving degradation. Similarly, Vita et al. [2] applied machine-learning classification to operational data from distribution systems to identify vulnerable transformers and schedule predictive maintenance, reinforcing the feasibility of data-centric CBM.

A comprehensive review by Al-Saedi et al. [3] concluded that computational-intelligence methods—including deep learning and IoT-enabled monitoring—are increasingly replacing offline diagnostics, marking a shift toward continuous, data-driven preventive maintenance. Parallel developments at the grid level underscore the value of multiscale temporal modeling. Salman et al. [4] combined wavelet transforms with LSTM and transformer architectures to forecast grid faults, demonstrating that hybrid temporal–spectral features improve prediction of system disruptions.

Building on these advances, our work bridges asset-centric and grid-centric perspectives with three key contributions. First, *grid-side predictive modeling*: rather than relying on internal diagnostics, we analyze externally observable power quality trends—voltage stability, reactive power asymmetry, and frequency volatility—to identify stress signatures that precede protective shutdowns, using multi-window feature engineering across 3-, 7-, 14-, and 21-day horizons. Second, *calibrated early warning*: the model produces calibrated probability estimates that exhibit characteristic buildup patterns days to weeks before interruptions, providing actionable lead time for operational intervention. Third, *operational validation*: collaboration with a commercial utility confirmed that predicted events represent genuine grid-induced protective shutdowns rather than scheduled maintenance, and that the 3-21 day warning horizon enables practical mitigation responses.

This paper describes the dataset comprising 18 months of power quality measurements from 25 substations, the multi-window modeling methodology, and predictive results demonstrating that grid-side observables provide meaningful advance warning of service interruptions.

979-8-3315-8138-1

II. Background and Project History

This work builds on a prior effort focused on transformer internal health modeling [5]. That project developed and validated a **Hybrid Artificial Intelligence (HAI)** framework combining probabilistic programming with data-driven learning to estimate transformer health and status (H&S) and remaining useful life (RUL). The HAI approach fused heterogeneous data (electrical, thermal, chemical, and environmental) to capture latent dependencies affecting degradation.

The probabilistic modeling component explicitly represented measurement and model uncertainty, producing full predictive distributions rather than single-point forecasts. This enabled robust inference even under sparse or noisy conditions. Predictive models of critical indicators were trained to forecast their evolution several weeks ahead, providing operators early visibility into emerging risk.

Building on these capabilities, the present study shifts focus from internal transformer diagnostics to grid-side precursors of service interruptions. Rather than modeling internal degradation mechanisms, this work analyzes externally observable PQ trends—voltage stability, reactive power asymmetry, and frequency volatility—to identify stress patterns that culminate in protective shutdowns. This shift reflects the operational reality that modern transformers are highly robust, with 25-year service lives, and that interruptions are more often triggered by external grid conditions than by internal component failures.

The dataset for this study comprises 18 months of PQ measurements from 25 distribution grid substations ranging from 13.8kV to 34.5kV identified by the utility as operationally problematic sites. These substations include a mix of distribution and sub-transmission assets, as well as reactive power compensation equipment (STATCOMs). The monitored events predominantly represent grid-induced protective shutdowns, as confirmed by post-hoc utility validation, rather than scheduled maintenance or catastrophic equipment failures. This reframing broadens the scope of Condition-Based Maintenance Plus (CBM+) to include grid-side observables, establishing a unified framework for anticipating interruptions caused by both internal and external factors.

III. Data Collection and Preprocessing

This study used 18 months of operational data obtained through a collaboration with a commercial utility. Identical monitoring units were installed at 25 transformer substations within a single grid network. To preserve confidentiality, the anonymized dataset is referred only as the *PDNV* dataset. The utility provided final review and expressed interest in applying these methods to grid stability analysis for large commercial and data-center customers.

The PDNV dataset comprised three integrated components. More than 20,000 categorized power-quality events captured transient disturbances in voltage, current, frequency, and power. Events were automatically classified by the sensor firmware according to standard categories consistent with IEEE 1159—voltage sags, swells, rapid voltage changes, and similar phenomena—and characterized by amplitude, duration, and affected phase. Although initial predictive models based on event counts and severity were explored, they exhibited limited foresight relative to models trained on continuous trend data. Inspection of high-resolution waveforms nonetheless revealed microsecond-scale "ringing" patterns that may propagate through intermediate systems, motivating future high-frequency studies.

Trend data formed the core of this analysis. Each monitoring unit produced approximately 600 channels of hourly measurements including voltage, current, active and reactive power, apparent power, frequency, and harmonic-distortion metrics. These hourly values represent extrema derived from much higher sampling rates, a practical compromise common to power-quality instrumentation

A custom preprocessing pipeline was developed to clean missing records, align all channels temporally, and generate fixed-length feature windows across 21-, 14-, 7-, and 3-day horizons. Event information was merged only at the modeling stage to avoid bias in preprocessing.

The PDNV dataset comprised two primary components: power-quality trend data and documented service interruptions. Trend data formed the core of this analysis. Each monitoring unit produced approximately 600 channels of hourly measurements, including voltage, current, active and reactive power, apparent power, frequency, and harmonic distortion metrics. These hourly values represent extrema derived from much higher sampling rates, a practical compromise common to power-quality instrumentation. Multi-window feature engineering was applied to capture degradation patterns across 3-, 7-, 14-, and 21-day horizons, enabling the model to detect both short-term disturbances and longer-term grid stress trends.

Ground-truth labels consisted of 78 documented service interruptions recorded by the monitoring system. These events were selected by the utility from substations identified as operationally problematic sites. While maintenance logs were unavailable for independent verification, several factors support the interpretation that these interruptions predominantly represent grid-induced protective shutdowns rather than scheduled maintenance. Analysis of interruption timing revealed no weekly, monthly, or seasonal patterns characteristic of scheduled maintenance activities. Additionally, a separate class of "reset" events, identified by the utility as scheduled operations, was excluded from the prediction target. The utility described these substations as experiencing "many problems", consistent with grid stress conditions rather than routine maintenance.

A custom preprocessing pipeline was developed to clean missing records, align all channels temporally, and generate fixed-length feature windows across the four temporal horizons (3-, 7-, 14-, and 21-day). Missing values within each window were imputed using forward-fill followed by backward-fill, with remaining gaps replaced by column medians for numeric features. Event information was merged only at the modeling stage to avoid bias during preprocessing.

Although initial predictive models based on categorized power-quality events (e.g., voltage sags, swells, rapid voltage changes) were explored, they exhibited limited foresight relative to models trained on continuous trend data. Inspection

979-8-3315-8138-1

of high-resolution waveforms revealed microsecond-scale “ringing” patterns that may propagate through intermediate systems, motivating future high-frequency studies. However, for this study, trend data provided the most reliable basis for predictive modeling.

IV. Modeling Approach

With over 600 measurement channels per monitoring unit, direct inclusion of all variables would introduce redundancy and excessive model complexity. To address this, a subset of 14 representative "bellwether channels" was selected using a domain-informed clustering and voting algorithm, ensuring that the reduced feature set preserved the variance structure of the full dataset while remaining interpretable. The utility’s technical review confirmed that the selected channels aligned with expert intuition.

For each bellwether, daily features were computed over **fixed-width trailing sliding windows** of 3, 7, 14, and 21 days, yielding four temporal resolutions of the same underlying grid condition. Each observation day was labeled by whether a service interruption occurred within the subsequent 21 days, forming a binary prediction problem suitable for early-warning applications. The resulting multi-window feature set combined short-term disturbance indicators with slower grid degradation trends, as recommended in recent power-quality research [6,7].

Twenty-four domain-specific features were engineered to capture voltage stability, load symmetry, dynamic response characteristics, and transient stress patterns. These features were computed directly from raw values leveraging CatBoost’s robustness to feature scaling.

The trend data used in this study was captured at a typical high sampling rate by the monitoring units but was only available as hourly aggregates. Each hourly value represents the maximum, minimum, or other extrema derived from the high-frequency data. This practical compromise, common to power-quality instrumentation, ensures that the most meaningful variations in power quality metrics are preserved while reducing data storage and processing requirements. Consequently, the engineered, shown in Table 1, features focus on capturing patterns in these extrema, which are most indicative of grid stress conditions.

Table 1: Model Engineered Features

Feature Type	Feature	Equation
Voltage Stability	Voltage range mean	$(1/N)\sum(V_{\max,t} - V_{\min,t})$
Voltage Stability	Voltage standard deviation	$\sigma(V_{\min})$
Voltage Stability	Voltage range volatility	$\sigma(V_{\max} - V_{\min})$
Power Factor	Power factor mean	$pf_t = P_{\min,t} / (S_{\min,t} + 10^{-6})$ $\bar{pf} = (1/N)\sum pf_t$
Power Factor	Power factor standard deviation	$\sigma(pf)$
Load Factor	Load factor mean	$P_{\min,t} / (P_{\max,t} + 10^{-6})$
Reactive Power	Reactive power asymmetry	$rr_t = Q_{\max,t} / (S_{\max,t} + 10^{-6})$ reactive_power_asymmetry = $\sigma(rr)$
Transient Stress	Current spike frequency	$\lvert\{t : \Delta I_t > P_{80}(\Delta I)\}\rvert / N$ where ΔI_t = 3-day rolling mean of $\lvert\Delta I\rvert$
Transient Stress	Transient power stress	stress_rate × stress_magnitude
Physics	Cumulative overload exposure	$\sum(P_t \text{ where } P_t > P_{95}(P)) / N$
Physics	Thermal cycling frequency	$\lvert\{t : \lvert\Delta P_t\rvert > P_{75}(\lvert\Delta P\rvert)\}\rvert / N$
Cross-Window w_L = Longer w_S = Shorter	Voltage instability ratio	$\sigma_{V,wS} / (\sigma_{V,wL} + 10^{-6})$
Cross-Window w_L = Longer w_S = Shorter	Voltage instability difference	$\sigma_{V,wS} - \sigma_{V,wL}$

The prediction model was formulated as a binary classification task optimized for the Area Under the Receiver Operating Characteristic (AUC-ROC) curve, which is insensitive to class imbalance [8]. The model predicts whether a service interruption will occur within 21 days based solely on historical measurements from the prior 3-21 days, ensuring strict temporal causality.

A shallow CatBoost gradient-boosted decision tree was selected to balance expressiveness and generalization. Default hyperparameters were retained without tuning to avoid overfitting to the limited dataset. Specifically, tree depth was limited to 6 with a learning rate of 0.03 and 1000 iterations. Limiting tree depth prevented memorization of idiosyncratic noise patterns observed in sparsely instrumented stations, a tradeoff consistent with recent approaches for mitigating overfitting in decision-tree models [9]. Feature preprocessing removed constant-valued columns and highly correlated features (correlation > 0.95) to reduce redundancy.

The dataset exhibits severe class imbalance, with positive-class (pre-interruption) days comprising approximately 0.6% of observations. Rather than resampling techniques, which can introduce bias in time-series contexts, class imbalance was addressed through optimization for AUC-ROC, which evaluates ranking quality independent of class proportions.

While gradient boosting naturally outputs class probabilities, these values are not intrinsically well-calibrated for operational use. Therefore, Platt scaling was applied post-training to map raw scores to calibrated probabilities using using scikit-learn's CalibratedClassifierCV with sigmoid calibration. This calibration ensures a broad monotonic relationship between predicted probabilities and observed event frequencies, producing conformally calibrated outputs that operators can interpret as genuine risk levels. For example, days assigned 70% probability should experience interruptions approximately 70% of the time. This interpretability is essential for operational decision-making, allowing operators to weigh predictive confidence when scheduling maintenance or evaluating risk.

Model evaluation used a strict temporal split: training data spanned January 2024 through January 2025, pooled across all 25 stations, with testing on February through April 2025. This temporal separation ensures that reported performance reflects true out-of-sample generalization rather than data leakage. performance metrics included AUC-ROC and confusion-matrix statistics, with AUC-ROC chosen as the optimization objective due to the severe class imbalance between normal and pre-shutdown days.

Although per-station models were infeasible given the limited interruption count at individual sites, future work with

979-8-3315-8138-1

longer observation periods will explore individualized station models to capture site-specific degradation patterns and mitigate potential bias toward frequently interrupted stations.

V. Results and Integration

The trained model was evaluated on a strictly time-based holdout set covering February 1 through March 30, 2025, comprising data from all 25 monitored substations unseen during training. Performance metrics confirm that the model generalizes beyond the training period and captures meaningful grid dynamics rather than memorizing past interruptions.

The overall AUC-ROC evaluated across all 25 substations of 0.78 indicates good discrimination between days preceding interruptions and normal operation days. At an operational threshold of 0.7, the model correctly identified approximately 80% of upcoming interruptions while maintaining a false-positive rate near 20%. This tradeoff reflects the project's goal of maximizing early-warning coverage without generating excessive alerts that could diminish operator trust.

Prediction timing analysis shows that alerts typically arise 3-21 days before service interruptions, with a median early-warning lead of approximately 8 days. Figure 2 illustrates this behavior for four representative substations, where calibrated failure probability accumulates progressively in the days preceding an interruption. This gradual buildup pattern—visible as successive peaks in the forecast—is consistent with evolving grid stress conditions and inconsistent with scheduled maintenance, which would not produce systematic power quality degradation prior to a planned event. During the evaluation period, the four representative substations recorded eight of ten actual interruptions with no false-positive alerts above the 0.7 threshold. The two missed events occurred at sites with sparse historical trend data, suggesting that prediction reliability improves with longer local baselines.

Post-event probability recovery provides additional evidence for the grid-stress interpretation: after each interruption, predicted risk returns to baseline levels, consistent with protective shutdowns resolving the underlying stress condition.

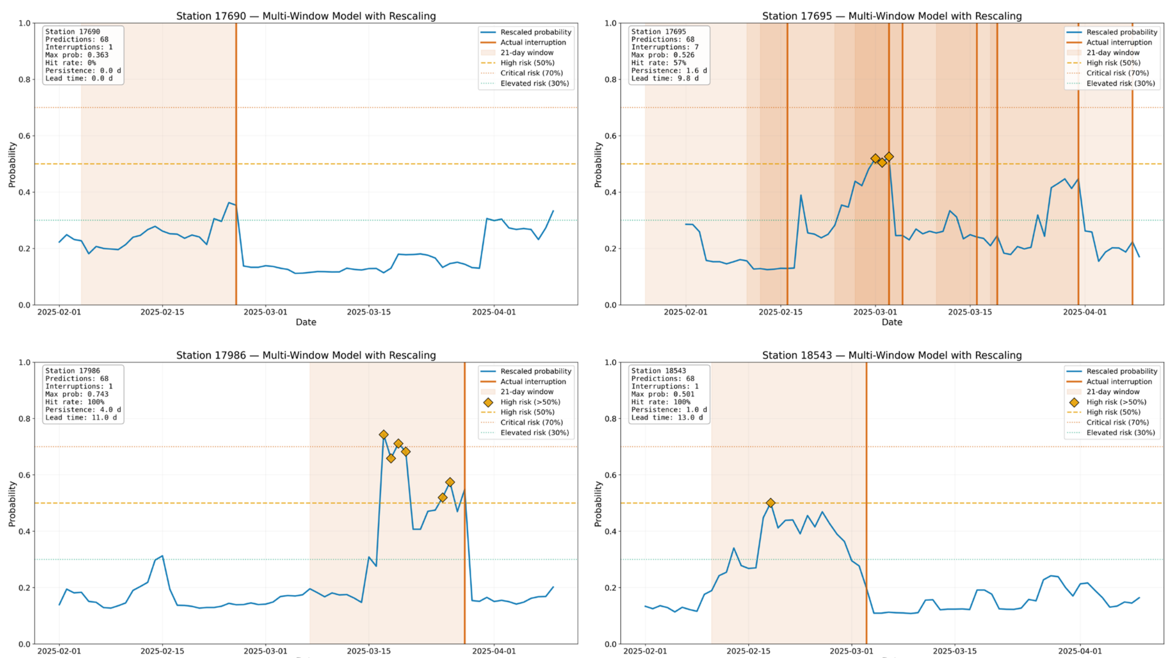

Figure 1: Multi-window model predictions shown for four stations

The model detected 8 out of 10 actual interruptions (80% true positive rate) across four representative substations with zero false positive alerts during the evaluation period. The two missed events (false negatives) occurred at locations with limited historical PQ data, suggesting that prediction accuracy improves with longer observation periods.

The low false positive rate is crucial for operational acceptance, as excessive false alarms can erode operator trust. By prioritizing high-confidence alerts, the system aims to provide actionable intelligence without causing operator fatigue.

Feature importance analysis reveals that voltage-related features (voltage standard deviation, voltage range mean, voltage regulation efficiency) and power quality metrics (power factor variability, reactive power asymmetry) are the strongest predictors of upcoming interruptions. Medium-term windows (7-14 days) provide the most predictive information, suggesting that interruptions are typically preceded by sustained abnormal patterns. However, all time scales contribute to predictions, with 3-day windows capturing acute disturbances and 21-day windows identifying gradual degradation.

VI. CBM+ Integration

POWERED III demonstrates that power quality event analysis provides complementary information to traditional CBM+ approaches by capturing transient disturbances that may indicate incipient faults before they manifest in steady-state measurements.

The project developed an operator-facing dashboard (Figure 3) that provides spatial and temporal visualization of model predictions, underlying PQ event data, and historical interruption records. The dashboard enables real-time monitoring across geographically distributed substations, allowing operators to examine failure probability predictions, detailed PQ event histories, feature importance, and historical prediction accuracy.

This visualization capability transforms raw model outputs into actionable intelligence that can guide maintenance prioritization and emergency response planning, particularly for utilities managing large numbers of distributed assets.

Figure 2: POWERED III interactive dashboard providing geographic visualization of monitored substations (top) with integrated time-series displays of model predictions, trend data, and event histories (bottom).

VII. Conclusion

POWERED III successfully demonstrates that power quality event analysis can provide meaningful advance warning of transformer service interruptions. The developed model achieves approximately 80% detection accuracy with minimal false positives, offering 3-21 day advance notice that enables proactive maintenance intervention.

By integrating PQ event monitoring into the CBM+ framework, the project extends traditional transformer health assessment with transient analysis of grid-side power quality dynamics. The hybrid AI approach—combining domain-engineered features with gradient-boosted tree learning—provides both interpretability and predictive performance.

The operational dashboard delivers these capabilities in a form that supports practical decision-making by grid operators. As power grids face increasing stress from renewable integration, electrification, and extreme weather events, predictive maintenance tools like POWERED III will become increasingly essential for maintaining grid reliability and preventing costly equipment failures.

VIII. References

[1] Y. Zeng and X. Chu, “Residual-aware health prediction of power transformers via spatiotemporal graph neural networks,” *PLOS ONE*, vol. 20, no. 1, pp. 1–16, 2025.

[2] M. Vita, F. Castelli-Dezza, and L. Piegari, “Predictive maintenance for distribution system operators in increasing transformers’ reliability,” *Electronics*, vol. 12, no. 6, p. 1356, 2023.

[3] R. Al-Saedi, A. E. Khalil, and M. El-Shimy, “Computational intelligence for preventive maintenance of power transformers: A review,” *Applied Soft Computing*, vol. 109, 2021.

[4] D. Salman, A. Kadhim, and M. Ismael, “Hybrid wavelet-LSTM-transformer model for fault forecasting in power grids,” *Int. J. Electr. Power Energy Syst.*, in press, 2024.

[5] S. Cvijic, N. Gupta, S. Lux, “Need for AI in Transformer Diagnostics and Prognostics ” *Proc. IEEE RAMS*, Orlando, FL, USA, 2023.

[6] Y. Hu, T. Zhang, and L. Chen, “Voltage stability indicators from time-series PMU data,” *IEEE Trans. Power Delivery*, vol. 37, no. 4, pp. 2951–2962, 2022.

[7] S. Alcaraz, M. Vega, and A. Maroto, “Power-quality indices for grid health monitoring,” *Energies*, vol. 14, no. 24, p. 8231, 2021.

[8] B. Fulcher and N. Jones, “Highly comparative time-series analysis: the empirical structure of time-series and their methods,” *J. R. Soc. Interface*, vol. 11, no. 101, 2014.

[9] M. Zlobin and V. Bazylevych, “A data-driven approach for balancing overfitting and underfitting in decision tree models,” 2025.

[10] R. Alonso, M. Chavarría, and J. A. Martínez, “Impact of large-scale data centers on power system stability,” *IEEE Transactions on Power Systems*, vol. 37, no. 2, pp. 1048–1058, 2022.

[11] T. Sundar, A. Pahwa, and C. Singh, “Integration of data centers into electric grids: challenges and opportunities for reliability and resilience,” *Electric Power Systems Research*, vol. 212, p. 108298, 2023.

979-8-3315-8138-1