Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26346-cd

A Structured Machine Learning Framework for Predictive Maintenance in Medical Imaging Systems

Bulent Alpay, Pankaz Das, Xi Wang

General Electric Healthcare, USA.
E-mail: bulent.alpay@gehealthcare.com, pankaz.das@gehealthcare.com, xi.wang1@gehealthcare.com

Medical imaging systems, such as computed tomography scanners, comprise highly integrated architectures of interdependent systems, subsystems, and components, where failures in any element can propagate across the system. Such disruptions lead to unplanned downtime, interruptions in clinical workflows, deferred patient examinations, and increased healthcare delivery costs. Traditional maintenance strategies are predominantly reactive, addressing failures only after they occur, which limits system reliability and compromises patient care. This reactive approach also results in inefficient resource utilization and potential unnecessary radiation exposure due to aborted scans.

We propose a structured machine learning framework for prognostics and health management, enabling predictive maintenance across imaging modalities. The approach integrates machine data (sensor data, system logs), operational data, and historical service records into a unified pipeline for Remaining Useful Life prediction. Advanced feature engineering techniques, including higher-order statistical transformations and drift detection metrics, capture latent degradation signatures within machine data. Degradation-specific feature selection, guided by hypothesis testing and domain expertise, isolates leading indicators from thousands of raw error codes, improving interpretability and robustness. The modeling strategy combines supervised learning with semi-supervised techniques, which address incomplete labeling in service ground truth data. Validation on historical and pilot deployments confirms scalability across product families and adaptability to software version variability.

This modular, scalable framework provides sufficient lead time for predictive service actions, reducing unplanned downtime and optimizing maintenance scheduling. By differentiating leading, lagging, and coincident log features, the approach enhances precision of predictions. Its adaptable design supports deployment across imaging modalities and extends to other industrial applications, offering a systematic and refined methodology for prognostics and health management.

Keywords: Predictive Maintenance, Machine Learning Framework, Remaining Useful Life Prediction, Prognostics and Health Management.

1. Introduction

The reliability of medical imaging equipment is critical to healthcare delivery, yet maintenance practices remain predominantly reactive. Adoption of Predictive Maintenance (PdM) is limited less by data scarcity and more by the semantic gap between machine-generated data and the physical state of machine health. Operating within safety-critical clinical workflows, catastrophic failures are actively prevented; as a result, datasets are right-censored and contain few true failure events Zio (2022); Compare et al. (2020).

A primary barrier to effective prognostics and health management (PHM) is the ground truth fidelity. Customer Relationship Management (CRM) service history and repair records document maintenance actions, not necessarily failure events. Interpreting all part replacements as failures introduces systematic label error (not random noise), misplacing event times, biasing hazards, and corrupting learned degradation signatures. Furthermore, the system logs generated by these devices are typically designed for debugging rather than prognostics. They are dominated by "lagging" indicators, error codes that trigger only after a functional limit has been breached, making them insufficient for early warning. We therefore employ semi-supervised label refinement to improve label fidelity prior to modeling.

This paper addresses these challenges by proposing a robust machine learning (ML) frame-

work designed to

- Refine Ground Truth: Utilize iterative semi-supervised learning to clean noisy service labels and identify high-fidelity degradation/failure events.
- Disentangle Log Complexity: Apply hypothesis driven feature engineering and statistical testing to categorize logs into leading, coincident, and lagging indicators.
- Isolate Degradation Modes: In complex imaging systems, isolating interacting failure modes from overlapping log and sensor symptoms constitutes an ill-posed inverse problem closely related to NP-hard mixture and source separation formulations. The proposed symptom elimination strategy imposes domain informed constraints that progressively remove explained symptom sets, transforming an otherwise combinatorial inference task into a sequence of well posed, degradation specific subproblems.

1.1. *Contribution*

This work introduces a structured ML framework for predictive maintenance in complex medical imaging systems, addressing two persistent challenges in industrial PHM: unreliable service-derived ground truth and the difficulty of extracting early degradation indicators from high-dimensional system logs. The framework integrates statistical feature discovery, iterative semi-supervised label refinement, and degradation-mode isolation into a unified pipeline designed for scalable deployment across imaging modalities and software versions.

Separately, label scarcity persists, true functional failures are rare, and most observations correspond to healthy or latent states. This class imbalance is distinct from the ground truth paradox and further complicates supervised learning (Zio 2022).

1.2. *Related work*

PHM and PdM have been widely studied for improving reliability and availability of complex systems Huang et al. (2024); Mallioris et al. (2024), with recent work emphasizing uncertainty, data quality, interpretability, and deployment constraints Zio (2022); Compare et al. (2020). Condition-based maintenance research has established diagnostics–prognostics–decision pipelines and highlighted the importance of feature extraction and data fusion for real assets Jardine et al. (2006). For Remaining Useful Life (RUL) estimation, statistical and stochastic-process approaches explicitly address censoring and heterogeneity, often building on survival foundations such as the Kaplan–Meier estimator and Cox proportional hazards models Kaplan and Meier (1958); Cox (1972); Si et al. (2011) . In operational service ecosystems, labels are frequently noisy or systematically biased; label-noise surveys and confident-learning methods motivate iterative label refinement rather than assuming service actions are ground truth failures Frenay and Verleysen (2014); Northcutt et al. (2021) . Finally, log-centric monitoring has grown rapidly in adjacent domains (e.g., sequence modeling for anomaly detection in system logs) and adaptive-window drift detection provides mechanisms to handle distribution shifts due to software updates and fleet evolution Du et al. (2017); Bifet and Gavaldà (2007). Compared to these threads, this work contributes a structured, modular PHM pipeline tailored to medical imaging systems that separates leading indicators from coincident/lagging logs, refines service-derived labels via iterative semi-supervision, and supports scalable deployment across versions and product families.

2. Data Challenge: Ground Truth and Log Design

2.1. *Ground truth paradox in service records*

In supervised learning, the model is only as good as its labels. In medical equipment maintenance, the labels derived from CRM service history and maintenance/repair record data are notoriously unreliable, creating a “ground truth paradox.”

- Ambiguity of Intervention: A service log stating “Replaced X-Ray Tube” does not confirm an X-ray tube failure. It may indicate a preventive replacement during a scheduled downtime,

or a misdiagnosis where the tube was replaced but the root cause lay in the High-Voltage (HV) generator. Interpreting all service-driven part replacements as failures results in systematically incorrect ground truth for degradation models, since some replacements occur prior to functional failure or without confirmed degradation.

- No Issue Found: A significant percentage of service calls result in no issue found, yet parts may still be swapped "just in case." These false positives pollute the training set, causing models to learn spurious correlations between healthy machine states and failure labels.
- Propagation and Amplification of Incorrect Ground Truth via Natural Language Processing (NLP): When CRM service records (SRs) are post-processed using ML or NLP techniques to derive failure labels, the resulting ground truth reflects both the original human decision biases encoded in service documentation and the intrinsic limitations of these methods in resolving interacting or concurrent degradation modes. This leads to systematic semantic mislabeling rather than stochastic error, further undermining degradation model validity.

2.2. *System log design*

System logs serve as the "voice" of the machine, but their taxonomy is often ill-suited for prediction. We categorize logs into three temporal classes relative to the failure event.

- Leading messages often precede observable system degradation and may serve as early indicators of potential failures. Their predictive value makes them critical for proactive maintenance strategies.
- Lagging messages, in contrast, typically appear after a fault has begun to impact system performance. These messages often exhibit a one-to-many relationship, where a single underlying issue generates multiple related logs across different components.
- Coincident messages are associated with specific symptoms and tend to occur simultaneously with other logs. They are numerous and diverse, often originating from various subsystems in response to a shared symptom, such as a computed tomography (CT) scan abort. This results in a complex one-to-many-to-many relationship, where a single symptom triggers a cascade of messages from multiple sources, complicating root cause analysis.

The challenge is exacerbated by cascading errors. In highly integrated systems, subsystems and components (SSCs), a single component failure (e.g., a slip ring) can trigger thousands of error logs across the gantry, image reconstruction system, and user interface. Distinguishing the causal "leading" log from the downstream "sympathetic" logs is a non-trivial causal inference problem.

3. Methodology: Structured ML Framework

The objective of this work is not to propose a new predictive algorithm but rather to present a structured framework that organizes data preparation, feature discovery, label refinement, and degradation modeling into a coherent PHM pipeline suitable for complex industrial systems. To address the limitations of unreliable ground truth and the complexity of system logs, we propose a multi-stage, iterative modeling framework that moves beyond static supervised training toward a dynamic process in which models progressively refine their training data.

The framework begins with initial ground truth construction, where labels are derived from service history and maintenance and repair records. In the second stage, feature engineering is performed using both system log data and sensor measurements, applying higher-order statistical transformations to capture temporal and distributional characteristics of machine health.

In the third stage, SSC specific leading, coincident, and lagging log features are identified using statistically grounded feature relevance and selection methods (e.g., permutation based testing, Boruta Kursa and Rudnicki (2010), ReliefF Kononenko (1994)) to assess temporal association with degradation and failure events. In the fourth stage, iterative ground truth refinement is carried out using a semi-supervised learning strategy to

Fig. 1. Overview of the structured ML-based PHM framework.

correct misaligned service-derived labels and recover latent degradation instances. Finally, degradation isolation is achieved through a symptom elimination strategy, which sequentially decouples interacting failure modes using the refined feature set. The overall framework is illustrated in Fig. 1, which is described in the following subsections.

3.1. *Ground truth creation and labeling*

In summary, service history and maintenance/ repair record data include technicians' understanding of an issue and field engineers' troubleshooting steps. In some cases, the symptoms of a functional failure of an SSC are well pronounced and isolated, such that either manual or NLP-based labeling of that service record (SR) is straightforward. For example, in Patient Handling subsystem of an MR system, cradle wheel wear out may lead to an increasing resistance on the cradle's in-out motion in time, which could also be observed either directly using the longitudinal motor's power consumption or indirectly using the time it takes the cradle to move a patient in and out of the bore as a proxy measurement. In this case, degradation of cradle wheels could be used as a label. However, when a technician reports this issue, the field engineer may not only replace these wheels but also clean all moving parts and replace the motor, because similar symptoms present when there is dirt in the rails, motor degradation or external power issues. This introduces a systematic bias to the ground truth, which limits its use without a correction step using the data and features. To overcome this issue, we use high level information in ground truth, i.e., instead of degradation/failure modes, using SSC names as labels to be corrected in the iterative semi-supervised stage. Note that, this approach still inherits other biases such as replacement of components simultaneously in multiple systems/subsystems and the quality-of-service record generation, which may be addressed using expert-judgement.

We excluded SRs involving multiple service actions, which may reflect mixed failure causes. For example, in analyzing CT detector failures, only SRs with detector replacements are retained, while SRs with any other parts involved are discarded. This ensures high-fidelity ground truth for modeling.

3.2. *Feature engineering*

Features from the raw system log or sensor data are generated in this stage. These are based on statistical representations of the underlying degradation phenomenon including drift indicators and expert-judgment. Aggregation of the raw data with respect to the temporal behavior (abrupt or incipient) of the modeled degradation that may include a sliding window representation may help generate the features of interest.

3.3. *Feature selection*

To isolate leading indicators from the noise of lagging and coincident indicators, causality and permutation-based hypothesis testing can be used.

Because RUL is inherently stochastic under varying operating conditions and usage profiles of a given SSC, prognostic models should provide actionable lead time for scheduling service interventions before unplanned downtime occurs. For each degradation mode, imaging systems may exhibit operationally mediated precursor signatures; therefore, feature relevance is evaluated within a predefined prognostic horizon. In this work, we operationalize the "near-failure" regime using a fixed RUL threshold (typically 7–21 days) to support hypothesis- and permutation-based testing, and we define a conservative "nominal/healthy" regime by retaining only observations with RUL $>$ 120 days, where failure is unlikely. This strict filtering reduces label noise in the negative class and sharpens the contrast needed to isolate leading indicators, while ongoing work tar-

gets SR–specific estimation of prognostic distance rather than a global horizon.

For imaging devices, we generally start with ∼3,000 distinct system log message types where ∼100-200 SSC-specific leading message codes are selected using these techniques, effectively. Note that in many cases, SSC experts were not aware of some of the error codes selected by these algorithms.

3.4. *Iterative Semi-Supervised learning*

We address the ground truth paradox using an iterative strategy. We use a classification algorithm for this purpose. In the first iteration, we train a classifier using a "Noisy Positive" set (ambiguous SRs). Service records labeled as failure of an SSC but predicted as "Healthy" with high confidence are flagged as misdiagnoses. These SRs are removed from the training set, as they may correspond to proactive component replacements, failure modes exhibiting conflicting degradation signatures, or instances of incorrect labeling.

Further iterations are conducted with this dataset as-is, to improve precision of the predictions, or adding unlabeled SRs or SRs with low confidence to improve labeling until the label set stabilizes effectively "distilling" high-fidelity ground truth from the messy service history.

3.5. *Symptom elimination and ensemble modeling*

We propose a symptom elimination logic based on the outputs of SSC-specific ML models. A new iteration is started with the SRs moved out of the training dataset through previous iterations to isolate other degradation modes via a symptom elimination strategy, which sequentially decouples interacting degradation modes using the refined feature set. An ensemble model is then created using all models coming out of iterations that address distinct degradation modes.

4. Case Study

The proposed framework was evaluated through a case study focused on predicting CT detector failures using high-dimensional system log data. The framework has been applied across multiple imaging system families in internal pilot deployments, demonstrating its practical applicability beyond the specific case presented here. The purpose of this case study is to illustrate the practical feasibility of the proposed framework using a real industrial dataset rather than to provide an exhaustive benchmarking study of predictive algorithms. Permutation-based feature selection was applied separately for each failure mode using service request grouped data, reducing nearly 3,000 syslog message codes to approximately 100 message codes per failure mode, effectively eliminating over 90% of irrelevant features. Domain experts confirmed that most top ranked features were semantically and operationally aligned with the target failure modes. The study further demonstrated that grouping data by SR substantially improved both computational efficiency and feature relevance, compared to applying permutation testing directly on raw, ungrouped data, which tended to retain a large number of lagging and coincident features.

Based on the predictive maintenance use case, model performance is evaluated using event-based measures designed to align closely with operational impact. For each SSC, only the first model prediction of an impending failure, replacement, or anomaly is considered. Predictions are evaluated within a predefined failure window following the prediction time. A prediction is classified as a True Positive (TP) if an actual failure or customer SR occurs within this window demonstrating the model's ability to anticipate actionable events, and as a False Positive (FP) if no such event occurs. Conversely, a False Negative (FN) corresponds to a failure or service request that occurs without any prior model prediction, with the system or component remaining classified as nominal up to the event.

Within this framework, precision is defined as TP/(TP+FP), serves as a key indicator of service cost efficiency. Higher precision corresponds to more targeted interventions, reducing unnecessary maintenance actions and improving utilization of SSCs. Recall is defined as TP/(TP+FN) captures the model's ability to minimize unplanned down-

time. Strong recall reduces missed failures, ensuring higher equipment up-time and minimizing disruptions to customer operations. Together, precision and recall provide a balanced view of performance, enabling informed trade-offs between targeted, cost-efficient maintenance interventions and operational risk, ultimately driving both cost optimization and improved service outcomes.

Using the selected features, predictive models were trained for detector failure classification and RUL estimation. Models incorporating feature selection achieved high recall ($> 90\%$) and low false positive rates ($< 1\%$) on both training and test data, indicating strong generalization. In contrast, models trained without feature selection exhibited significant performance degradation on test data, with increased false alarms and reduced recall, highlighting the risk of overfitting when irrelevant log features dominate the input space. The selected feature models also required substantially less computational time and showed more stable behavior during hyperparameter tuning. In production deployment, the detector model achieved approximately 91% recall and precision, with a median prognostic distance exceeding 21 days, demonstrating the practical value of the proposed ML framework for enabling reliable PdM from system log data.

5. Conclusion

This paper presented an iterative PdM framework for complex medical imaging systems that addresses unreliable service-derived ground truth and high-dimensional, ambiguous log data. The approach combines feature engineering, statistically grounded feature selection, semi-supervised label refinement, and degradation isolation. By separating leading, coincident, and lagging indicators and iteratively refining training data, the framework converts failure inference into tractable, degradation-specific modeling tasks.

The framework was validated using a CT detector case study, where feature selection reduced thousands of log-derived features to a small, interpretable set while improving generalization. Models with selected features achieved high precision and recall, low false positives, and a median prognostic distance of over three weeks, enabling actionable service lead time. Models without feature selection showed overfitting and poorer test performance, underscoring the importance of feature relevance and label refinement. More broadly, the framework is modular and extensible across imaging modalities, product families, and software versions. Although focused on log-centric PdM, the methodology generalizes to industrial systems with sparse failures and imperfect service records, providing a practical foundation for advancing PHM in safety-critical systems.

References

Bifet, A. and R. Gavaldà (2007, April). Learning from Time-Changing Data with Adaptive Windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448. Society for Industrial and Applied Mathematics.

Compare, M., P. Baraldi, and E. Zio (2020, May). Challenges to IoT-Enabled Predictive Maintenance for Industry 4.0. *IEEE Internet of Things Journal 7*(5), 4585–4597.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society*.

Du, M., F. Li, G. Zheng, and V. Srikumar (2017, October). DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas Texas USA, pp. 1285–1298. ACM.

Frenay, B. and M. Verleysen (2014, May). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems 25*(5), 845–869.

Huang, C., S. Bu, H. H. Lee, C. H. Chan, S. W. Kong, and W. K. Yung (2024). Prognostics and health management for predictive maintenance: A review. *Journal of Manufacturing Systems 75*, 78–101.

Jardine, A. K., D. Lin, and D. Banjevic (2006, October). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing 20*(7), 1483–1510.

Kaplan, E. L. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of Relief. In *Machine Learning ECML*.

Kursa, M. B. and W. R. Rudnicki (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software 36*(11).

Mallioris, P., E. Aivazidou, and D. Bechtsis (2024, June). Predictive maintenance in Industry 4.0: A systematic multi-sector mapping. *CIRP Journal of Manufacturing Science and Technology 50*, 80–103.

Northcutt, C., L. Jiang, and I. Chuang (2021, April). Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research 70*, 1373–1411.

Si, X.-S., W. Wang, C.-H. Hu, and D.-H. Zhou (2011). Remaining Useful Life Estimation: A Review. *European Journal of Operational Research*.

Zio, E. (2022, February). Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety 218*, 108119.