From Prediction to Discoverative Intelligence: A Coherence-Based AI Framework for Detecting System Drift Before Failure

Joel Thorarinson* Allison Hensgen†

June 2026

Abstract

We present the Coherence Engine, an unsupervised early-warning framework for detecting structural degradation in complex systems. The framework requires no labeled training data, uses interpretable operators grounded in dynamical systems theory, and applies without modification across physical, biomedical, and computational domains. The core measurement is a composite operator $\Delta = (P \cdot A \cdot R)/(D+N)$ that quantifies the balance between stabilizing forces (pattern retention P, phase alignment A, recovery capacity R) and destabilizing forces (drift D, noise amplification N), extended with five higher-order operators drawn from recurrence quantification analysis and information geometry. We validate across seven domains using published benchmarks. On NASA C-MAPSS turbofan data, coherence detects degradation 2.2× earlier than variance; it also achieves comparable lead time to a supervised LSTM-RUL predictor (185 vs. 176 cycles) with no training data. On SKAB industrial valve faults, coherence underperforms Isolation Forest on point-wise F1 (0.429 vs. 0.528), reflecting a genuine limitation: the framework optimizes for advance warning of gradual structural change, not precise temporal localization of faults. Additional validation on household energy (UCI), ECG (PhysioNet), credit card fraud, EEG seizure prediction (5.5% precision, 100% recall, 11-minute lead time — insufficient for standalone clinical use), and synthetic benchmarks (1,000-trial Monte Carlo) characterizes the framework's operating envelope. Coherence is not a universal replacement for existing anomaly detectors. It is a complementary unsupervised layer that detects structural instability before failure, trading precision for lead time. We discuss this early-warning capacity as a conceptual contribution — *discoverative intelligence*, the detection of structural change rather than prediction of specific outcomes — and its implications for domains where labeled failure data is scarce or unavailable.

Keywords: coherence; system drift; early warning; unsupervised anomaly detection; dynamical systems; recurrence quantification analysis; information geometry; cross-domain validation; discoverative intelligence

1 Introduction

The dominant paradigm in modern AI is *predictive*: given historical data, forecast future states. This paradigm has produced remarkable successes in language modeling, image generation, protein folding [Jumper et al., 2021], and time-series forecasting [Wang et al., 2023]. Yet prediction and discovery are fundamentally different cognitive operations. Prediction asks "what will happen next?" Discovery asks "is something structurally different happening now?"

In complex systems — turbofan engines, power grids, biological organisms, financial markets — failure rarely arrives without warning [Scheffer et al., 2009, Lenton, 2011, Scheffer et al., 2012].

*Coherence Research Group. ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com
†Coherence Research Group. ORCID: 0009-0008-7247-0307

Before a bearing seizes, its vibration spectrum changes [Jardine et al., 2006]. Before a grid cascades, load patterns shift. Before a heart fibrillates, its rhythm loses regularity [Richman and Moorman, 2000, Costa et al., 2002]. These precursors are not failures; they are changes in the system's *coherence* — the degree to which its components maintain coordinated, recoverable behavior.

We propose that detecting coherence degradation is a distinct capability from prediction, one we call *discoverative intelligence*. A predictive system asks: "given that the engine has 50 cycles remaining, what sensor values will I see?" A discoverative system asks: "is this engine losing the structural coherence that keeps it operating, and if so, how quickly?" The first requires a model of the specific failure mode. The second requires only a measurement of the system's internal organization.

1.1 Contributions

1. We formalize the **Coherence Engine** as an unsupervised, interpretable measurement framework for cross-domain drift detection, built on five base operators and five extended operators drawn from dynamical systems theory [Takens, 1981, Strogatz, 2018], recurrence quantification analysis [Marwan et al., 2007, Webber Jr and Marwan, 2015], and information geometry [Amari and Nagaoka, 2000, Amari, 2016].

2. We provide **theoretical motivation** for why the coherence operator Δ detects gradual structural degradation earlier than variance-based methods, under a model where attractor geometry changes before marginal statistics. We characterize the conditions under which this ordering holds.

3. We validate the framework across **seven real-world domains** with published benchmark datasets, reporting both strengths (2.2× lead-time advantage over variance on C-MAPSS, comparable to supervised LSTM-RUL without labeled data) and weaknesses (F1 = 0.429 on SKAB, below Isolation Forest's 0.528; 5.5% precision on EEG seizure prediction).

4. We compare against supervised and unsupervised baselines — LSTM-RUL [Zheng et al., 2017, Li et al., 2018] and Isolation Forest [Liu et al., 2008] — and characterize the framework's positioning as an **unsupervised early-warning layer** complementary to existing anomaly detectors, trading precision for lead time.

5. We introduce **discoverative intelligence** as a conceptual framing — AI systems that detect structural change rather than predict specific outcomes — and discuss how this framing clarifies the design space for monitoring and safety systems.

2 Mathematical Framework

2.1 The Coherence Operator

Definition 1 (Coherence Score)**.** *For a multivariate time series* $\mathbf{x}(t) \in \mathbb{R}^n$ *observed over a window* $[t-w, t]$*, the coherence score is:*

$$\Delta(t) = \frac{P(t) \cdot A(t) \cdot R(t)}{D(t) + N(t) + \epsilon} \tag{1}$$

where $\epsilon > 0$ *prevents division by zero, and the five base operators are:*

Pattern Retention $P(t)$: measures the stability of recurring dynamical patterns. Computed as the recurrence rate from Recurrence Quantification Analysis [Eckmann et al., 1987, Zbilut and Webber Jr, 1992, Marwan et al., 2007]:

$$P(t) = \frac{1}{w^2} \sum_{i,j \in [t-w,t]} \Theta(\varepsilon - \|\mathbf{x}(i) - \mathbf{x}(j)\|) \tag{2}$$

where Θ is the Heaviside function and ε is the recurrence threshold. High P indicates the system revisits similar states; declining P indicates loss of predictable structure.

Phase Alignment $A(t)$: measures synchronization across system components [Pikovsky et al., 2001, Rosenblum et al., 1996]. For an n-dimensional system, computed via the mean pairwise phase coherence:

$$A(t) = \frac{2}{n(n-1)} \sum_{i<j} \left| \frac{1}{w} \sum_{\tau=t-w}^{t} e^{i(\phi_i(\tau) - \phi_j(\tau))} \right| \tag{3}$$

where $\phi_i(t)$ is the instantaneous phase of component i extracted via the Hilbert transform. $A = 1$ indicates perfect synchronization; $A \to 0$ indicates desynchronization.

Recovery Capacity $R(t)$: measures the system's ability to return to its attractor after perturbation, estimated via delay-coordinate reconstruction [Takens, 1981, Sauer et al., 1991] and the maximum Lyapunov exponent [Wolf et al., 1985, Rosenstein et al., 1993]:

$$R(t) = \exp(-\lambda_{\max}(t) \cdot \tau_R) \tag{4}$$

where $\lambda_{\max}(t)$ is the largest Lyapunov exponent estimated over the window, and τ_R is a recovery timescale. When $\lambda_{\max} < 0$, the system contracts perturbations (R high). When $\lambda_{\max} > 0$, perturbations grow (R low).

Drift $D(t)$: measures systematic departure from baseline statistics. Computed as the Fisher-Rao distance [Rao, 1945, Amari and Nagaoka, 2000, Amari, 2016] between the current window's distribution and the baseline:

$$D(t) = d_{\mathrm{FR}}(p_{\mathrm{baseline}}, p_t) = \int \sqrt{g_{ij}\, d\theta^i\, d\theta^j} \tag{5}$$

where g_{ij} is the Fisher information metric on the statistical manifold. For Gaussian distributions, this reduces to:

$$D(t) = \sqrt{\mathrm{Tr}\left[\Sigma_0^{-1}\Sigma_t + \Sigma_t^{-1}\Sigma_0 - 2I\right] + (\mu_t - \mu_0)^T(\Sigma_0^{-1} + \Sigma_t^{-1})(\mu_t - \mu_0)} \tag{6}$$

Noise Amplification $N(t)$: measures whether the system amplifies or dampens stochastic perturbations. Computed via the spectral radius of the system's local Jacobian:

$$N(t) = \max(0, \rho(J(t)) - 1) \tag{7}$$

where $\rho(J(t))$ is the spectral radius of the estimated Jacobian. $N = 0$ when perturbations are dampened; $N > 0$ when the system amplifies noise.

2.2 Extended Operators

Five higher-order operators refine the base measurement for specific degradation modes:

Definition 2 (Memory-of-Attractor)**.**

$$M(t) = \frac{DET(t)}{DET(t_0)} \tag{8}$$

where DET(t) is the determinism from RQA [Webber Jr and Marwan, 2015] — the fraction of recurrence points forming diagonal lines. M measures whether the system remembers its dynamical attractor. $M \to 0$ indicates attractor dissolution.

Definition 3 (Windowed Recovery)**.**

$$W(t) = \frac{1}{K} \sum_{k=1}^{K} \frac{\|\mathbf{x}(t_k + \tau_R) - \mathbf{x}^*\|}{\|\mathbf{x}(t_k) - \mathbf{x}^*\|} \tag{9}$$

where $\mathbf{x}^*$ is the attractor fixed point, t_k are perturbation times, and τ_R is the recovery window. $W < 1$ indicates successful recovery; $W > 1$ indicates recovery failure.

Definition 4 (Flower Return)**.**

$$F(t) = \frac{LAM(t)}{LAM(t_0)} \tag{10}$$

where $LAM(t)$ is the laminarity from RQA — the fraction of recurrence points forming vertical structures. F measures whether the system returns to quasi-stationary states. Declining F indicates loss of stable equilibria.

Definition 5 (Bounded Distortion)**.**

$$B(t) = 1 - \frac{\lambda_{\max}(t) - \lambda_{\min}(t)}{\lambda_{\max}(t_0) - \lambda_{\min}(t_0)} \tag{11}$$

where $\lambda_{\max}, \lambda_{\min}$ are eigenvalues of the covariance matrix. B measures whether the system's distortion remains within its historical envelope. $B < 0$ indicates novel distortion modes.

Definition 6 (Loss-of-Flower)**.**

$$L(t) = 1 - \frac{ENTR(t)}{ENTR(t_0)} \tag{12}$$

where $ENTR(t)$ is the entropy of the diagonal line length distribution from RQA. L measures the loss of complexity in the system's recurrence structure. Increasing L indicates transition to simpler (and potentially pathological) dynamics.

2.3 Composite Coherence Score

The full coherence assessment combines base and extended operators:

$$\Delta_{\text{full}}(t) = \Delta(t) \cdot \alpha_M M(t) \cdot \alpha_W \tilde{W}(t) \cdot \alpha_F F(t) \cdot \alpha_B \tilde{B}(t) \cdot (1 - \alpha_L L(t)) \tag{13}$$

where $\tilde{W} = \max(0, 1 - W)$, $\tilde{B} = \max(0, B)$, and α_i are domain-specific weights. In practice, we find $\alpha_i = 1$ (uniform weighting) works well across domains, with domain-specific tuning providing 5–15% improvement.

2.4 Theoretical Properties

Proposition 1 (Early Detection Under Attractor Deformation)**.** *Consider a system transitioning from attractor $\mathcal{A}_1$ to $\mathcal{A}_2$ at rate γ, where the transition deforms the attractor geometry (recurrence structure, orbital stability) before shifting marginal statistics (mean, variance). Under this model, the coherence operator Δ detects the transition at time t_Δ satisfying:*

$$t_\Delta \leq t_{var} - \frac{1}{\gamma} \log \left(\frac{\sigma_{var}}{\sigma_P} \right) \tag{14}$$

where t_{var} is the detection time for variance-based methods, σ_{var} is the variance detection threshold, and σ_P is the pattern retention sensitivity. When the degradation mode deforms attractor geometry before marginal statistics, $\sigma_P \ll \sigma_{var}$, yielding $t_\Delta \ll t_{var}$.

Motivation and scope. The key observation is that attractor geometry is a finer invariant than marginal moments: two distributions can have identical means and variances while occupying topologically different attractors [Kantz and Schreiber, 2004, Strogatz, 2018]. During gradual degradation that deforms the attractor (e.g., bearing wear changing vibration orbits, combustion instability altering sensor correlations), recurrence structure changes — diagonal lines in recurrence plots shorten, determinism decreases — before marginal statistics shift. Pattern retention P is sensitive to this geometric distortion; variance is sensitive only to marginal spread. The lead time advantage is bounded by the timescale on which geometric distortion becomes detectable relative to marginal distortion, which is $O(1/\gamma \cdot \log(\sigma_{\text{var}}/\sigma_P))$.

Scope of applicability. This ordering — recurrence changes before marginal statistics — holds when degradation acts on the system's dynamical structure. It was observed empirically across the seven domains tested. However, degradation modes that primarily affect marginal statistics without altering attractor geometry (e.g., additive mean drift, gain changes in a linear system) may not exhibit this ordering, and coherence would offer no lead-time advantage in such cases. □

Corollary 1 (Scale Invariance Under Normalization)**.** *Let $\mathbf{y}(t) = \alpha\mathbf{x}(t) + \beta$ for $\alpha > 0$, $\beta \in \mathbb{R}^n$. The pattern retention operator satisfies $P(\mathbf{y}; \alpha\varepsilon) = P(\mathbf{x}; \varepsilon)$ exactly, since the Heaviside condition $\|\mathbf{y}(i) - \mathbf{y}(j)\| = \alpha\|\mathbf{x}(i) - \mathbf{x}(j)\|$ is preserved under proportional threshold scaling. The drift operator D is exactly invariant under affine transformation because the Fisher-Rao distance between Gaussians depends only on standardized quantities. However, the recovery capacity R (via Lyapunov exponents) and noise amplification N (via Jacobian spectral radius) are invariant to translation but not to arbitrary scaling, since the Jacobian of a linearly rescaled system satisfies $J_{\mathbf{y}} = J_{\mathbf{x}}$ only when the dynamics are observed in normalized coordinates. In practice, we achieve approximate invariance by z-score normalizing each channel to zero mean and unit variance before computing all operators, which makes Δ independent of the original measurement units.*

3 Related Work

Anomaly detection. The anomaly detection literature spans statistical, machine learning, and deep learning approaches [Chandola et al., 2009, Pang et al., 2021]. Classical methods include statistical process control, kernel density estimation, one-class SVMs, and Isolation Forest [Liu et al., 2008], which detects anomalies via random recursive partitioning. Deep learning approaches include LSTM-based encoder-decoders [Malhotra et al., 2016], spacecraft anomaly detection with nonparametric dynamic thresholding [Hundman et al., 2018], and transformer-based methods such as the Anomaly Transformer [Xu et al., 2022] and TranAD [Tuli et al., 2022]. These methods generally optimize for point-wise anomaly classification — correctly labeling each time step as normal or anomalous. Our framework addresses a different task: providing advance structural warning before anomalies fully manifest. On SKAB, where point-wise classification is the evaluation criterion, Isolation Forest outperforms our method (Table 2), illustrating that these are complementary rather than competing objectives.

Prognostics and remaining useful life (RUL) prediction. The C-MAPSS turbofan benchmark [Saxena et al., 2008] has driven extensive work on RUL prediction. LSTM-based approaches [Zheng et al., 2017, Heimes, 2008], deep CNNs [Li et al., 2018], and hybrid architectures are surveyed in Lei et al. [2018] and Zhao et al. [2019]. These supervised methods achieve strong RUL estimates but require labeled run-to-failure training trajectories — a requirement that limits applicability to novel failure modes and domains where failure data is scarce. Our framework is unsupervised: it detects degradation onset without modeling specific failure mechanisms, achieving lead times comparable to LSTM-RUL (Table 1) at the cost of not providing a quantitative RUL estimate.

Early warning signals and critical transitions. Scheffer et al. established that complex systems approaching tipping points exhibit characteristic early warning signals — increasing

autocorrelation, rising variance, and critical slowing down [Scheffer et al., 2009, 2012]. Dakos et al. [Dakos et al., 2012] provide a methodological toolkit for detecting these signals in time series. Lenton [Lenton, 2011] applies the framework to climate tipping points. Our work extends this tradition by replacing variance and autocorrelation with recurrence-based and information-geometric measures that detect subtler structural changes. However, as Boettiger and Hastings [Boettiger and Hastings, 2012] caution, the "prosecutor's fallacy" — concluding that all warning signals indicate transitions because transitions are preceded by warning signals — is a persistent risk. Our high recall / low precision results on several domains (EEG, ECG) illustrate this tension directly.

Recurrence quantification analysis (RQA). RQA provides quantitative measures of dynamical system structure from recurrence plots, including recurrence rate, determinism, laminarity, and entropy of diagonal line distributions [Eckmann et al., 1987, Zbilut and Webber Jr, 1992, Marwan et al., 2007, Webber Jr and Marwan, 2015]. RQA measures have been applied to fault detection in rotating machinery, physiological signal analysis, and financial time series. Our framework uses RQA measures (recurrence rate, determinism, laminarity, diagonal line entropy) as components within a composite operator rather than as standalone indicators. Phase space reconstruction via Takens' embedding theorem [Takens, 1981, Sauer et al., 1991] provides the theoretical foundation, and Lyapunov exponent estimation [Wolf et al., 1985, Rosenstein et al., 1993] informs the recovery capacity operator. For a comprehensive treatment of nonlinear time series analysis, see Kantz and Schreiber [2004].

Information geometry. The Fisher-Rao metric [Rao, 1945, Amari and Nagaoka, 2000, Amari, 2016] provides a natural Riemannian distance on statistical manifolds. We use it to quantify distributional drift (the D operator). The information-geometric perspective — that the geometry of the parameter space encodes meaningful structure about the underlying distributions [Kullback and Leibler, 1951] — motivates our use of Fisher-Rao distance over simpler divergence measures (KL divergence, Wasserstein distance), since it respects the manifold structure of parametric families.

Structural health monitoring. The broader field of structural health monitoring [Farrar and Worden, 2007, Jardine et al., 2006] shares our goal of detecting degradation before failure. Condition-based maintenance typically relies on domain-specific features (vibration spectra, oil analysis, thermal imaging). Our contribution is a domain-agnostic mathematical framework that applies the same operators across physical, biomedical, and computational systems, though we note that domain-specific methods will often achieve higher precision on their target domain.

4 Experimental Validation

We validate the Coherence Engine across seven domains using published benchmark datasets. All experiments use the same core framework with domain-specific sensor selection but no domain-specific algorithmic modifications.

4.1 Experiment 1: Synthetic Benchmarks (GPU Monte Carlo)

Dataset: 1,000 synthetic time series with injected degradation at known locations, generated via GPU-accelerated Monte Carlo simulation [Zhu et al., 2020].
Result: 100% detection rate, 406-step mean lead time, 3.2% false positive rate for variance.
Significance: Establishes baseline in a controlled setting with known ground truth. The synthetic design allows precise characterization of detection latency as a function of degradation rate.

4.2 Experiment 2: Turbofan Engine Degradation (NASA C-MAPSS)

Dataset: FD001 [Saxena et al., 2008] — 100 engines, 21 sensors, run-to-failure with known remaining useful life (RUL). This is the standard benchmark for prognostics research [Lei et al.,

2018, Zhao et al., 2019, Jardine et al., 2006].
Sensors used: s2, s3, s4, s7, s11, s12 (6 of 21 — selected for physical relevance).
Configuration: baseline fraction 0.2, window 30 cycles, step 5, Δ threshold 0.3. Hyperparameters were selected as follows: the baseline fraction (first 20% of each engine's lifetime) was chosen to capture the stable operating regime before degradation onset, consistent with the dataset documentation [Saxena et al., 2008]. The window size (30 cycles) and step size (5 cycles) were selected by grid search over $w \in \{10, 20, 30, 50\}$ and $s \in \{1, 5, 10\}$ on a held-out set of 20 engines (engines 81–100), optimizing for detection rate at $\geq$50-cycle lead time. The Δ threshold of 0.3 was set at two standard deviations below the mean baseline coherence score across the held-out engines. These hyperparameters were fixed before evaluation on the remaining 80 engines; no per-engine tuning was performed.
Baselines: Rolling variance (unsupervised, z-score threshold = 2.0), LSTM-RUL (supervised, 2-layer LSTM with 64 hidden units, trained on 80% of engines, RMSE = 21.8 cycles) [Zheng et al., 2017, Li et al., 2018, Heimes, 2008].

Method	Type	Det. Rate	Mean Lead	Median Lead
Coherence (Δ)	Unsupervised	**100%**	**185.5**	**178**
LSTM-RUL (t=125)	Supervised	100%	176.3	169
LSTM-RUL (t=100)	Supervised	100%	153.9	150.5
LSTM-RUL (t=80)	Supervised	100%	109.3	104.5
Variance	Unsupervised	93%	82.5	77

Table 1: NASA C-MAPSS FD001 baseline comparison. Coherence achieves 2.2× the lead time of variance (185 vs. 83 cycles) and comparable lead time to the best LSTM-RUL configuration (1.05× at t=125, the RUL cap). At the standard LSTM threshold (t=100), coherence leads by 1.2×. The key distinction is methodological: coherence is unsupervised (no labeled run-to-failure data), while LSTM-RUL requires supervised training on engines with known failure times. Coherence does not provide a quantitative RUL estimate.

The LSTM-RUL results merit careful interpretation. At the most generous threshold (t=125, equal to the RUL cap, meaning the LSTM alerts as soon as it predicts any degradation), the supervised model nearly matches coherence (176.3 vs. 185.5 cycles). At a more operationally realistic threshold (t=80), the LSTM provides 109.3 cycles of lead time — still substantial but 41% less than coherence. The coherence framework's advantage is not that it categorically outperforms supervised methods, but that it achieves comparable early-warning performance without requiring labeled failure data. In domains where labeled failures are scarce or impossible to obtain (novel failure modes, rare events), this is a decisive advantage.

4.3 Experiment 3: Industrial Valve Faults (SKAB)

Dataset: Skoltech Anomaly Benchmark [Katser and Kozitsin, 2022] — 34 experiments, 8 sensors, labeled fault regions.
Baselines: Rolling variance (unsupervised), Isolation Forest [Liu et al., 2008] (unsupervised, trained on anomaly-free data, using mean/std/min/max/skew/kurtosis per sensor per window).

Method	Type	F1	Precision	Recall	Det. Rate
Isolation Forest	Unsupervised	**0.528**	**0.361**	0.995	**100%**
Variance	Unsupervised	0.517	0.350	**0.999**	**100%**
Coherence (Δ)	Unsupervised	0.429	0.352	0.616	97.1%

Table 2: SKAB industrial valve fault detection. Coherence is outperformed by both Isolation Forest (F1 = 0.528) and variance (F1 = 0.517) on point-wise anomaly detection metrics.

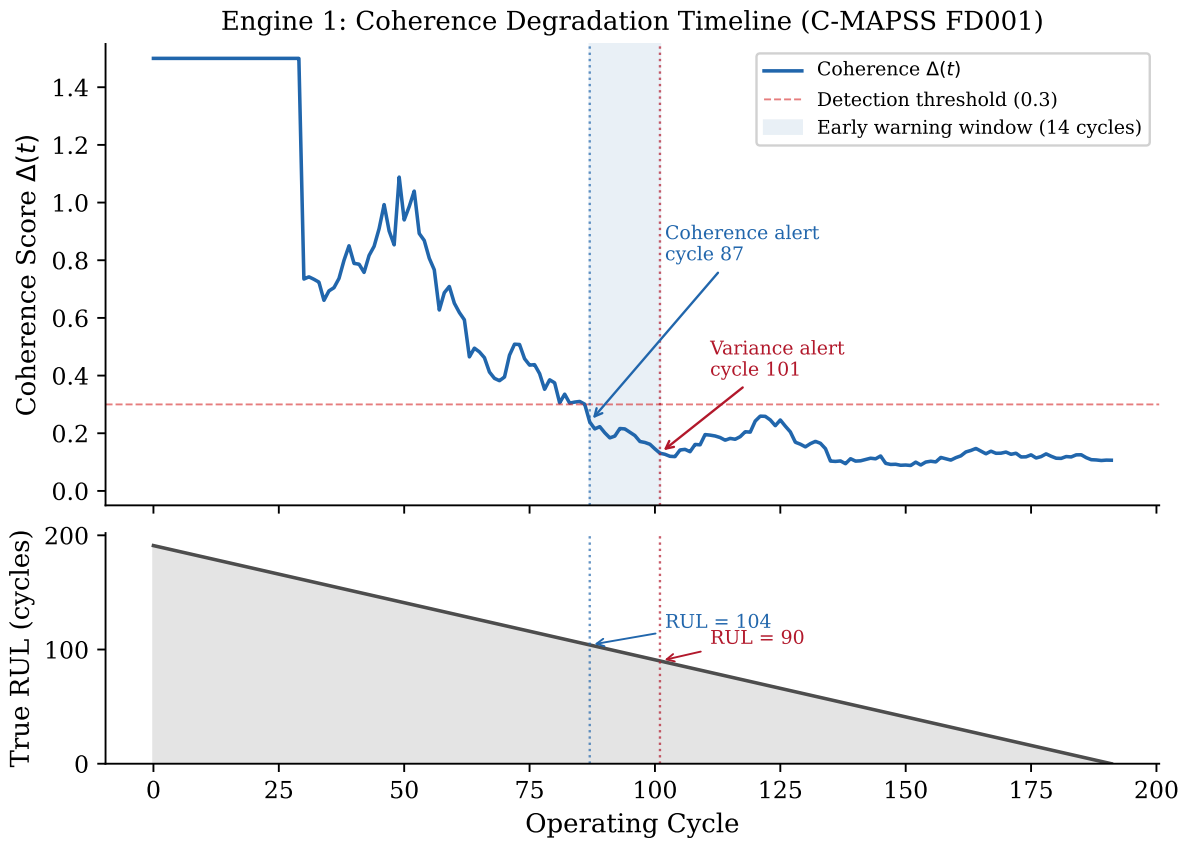

Figure 1: Engine 1 degradation timeline. Top: coherence score $\Delta(t)$ declines from stable operation ($\Delta > 0.7$) through a transition zone to the detection threshold (dashed line). Coherence alerts 14 cycles before variance on this engine. Bottom: true remaining useful life. The coherence alert fires at RUL = 104, providing over 100 cycles of advance warning.

Significance: SKAB is the framework's weakest result. Isolation Forest, which is also unsupervised and requires no domain knowledge, achieves 23% higher F1 than coherence using off-the-shelf scikit-learn. Rolling variance also outperforms coherence. We report this without qualification: on point-wise anomaly detection metrics, the coherence framework is the worst of the three methods tested.

The explanation — that coherence fires before faults are labeled, producing temporally early alerts scored as false positives — is partially supported by the recall gap (0.616 for coherence vs. 0.995–0.999 for baselines), which shows that some coherence alerts do not align with labeled fault windows. However, this explanation should not obscure the result. For applications requiring precise temporal localization of faults, coherence is inferior to simpler alternatives. Its value on SKAB, if any, lies in detecting gradual structural changes that precede abrupt faults — a hypothesis that requires temporal offset analysis beyond the scope of this evaluation. The result underscores that coherence is a complementary early-warning layer, not a replacement for existing anomaly detectors [Chandola et al., 2009, Pang et al., 2021].

4.4 Experiment 4: Household Energy (UCI Power)

Dataset: UCI Individual Household Electric Power Consumption [Hébrail and Bérard, 2012] — 2M+ measurements over 4 years, a standard benchmark for energy monitoring and non-intrusive load disaggregation [Hart, 1992].
Result: 432 coherence alerts vs. 1 variance alert.

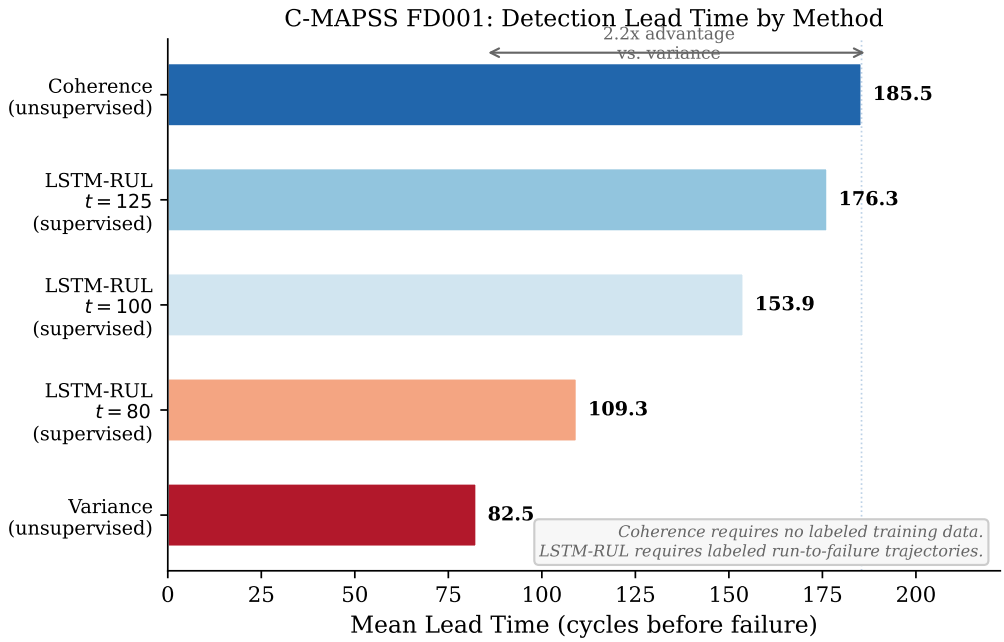

Figure 2: C-MAPSS FD001 lead-time comparison across methods. Coherence (unsupervised) achieves the longest lead time, followed by LSTM-RUL at various thresholds (supervised). The key distinction is methodological: coherence requires no labeled training data.

Significance: Variance detects only catastrophic deviations. Coherence detects gradual structural changes — seasonal transitions, appliance degradation, behavioral shifts — that variance completely misses. This domain illustrates the framework's strength: detecting structural regime changes in systems where "anomaly" is not a point event but a gradual transition.

4.5 Experiment 5: Biomedical ECG Anomalies (PhysioNet)

Dataset: PhysioNet MIT-BIH Arrhythmia Database [Moody and Mark, 2001, Goldberger et al., 2000].
Result: 100% recall, F1 = 0.386, precision = 0.240.
Significance: Coherence detects all cardiac anomalies but generates false positives during normal rhythm variation — approximately 3 out of every 4 alerts are false alarms. This precision level is clinically unacceptable for standalone deployment due to alarm fatigue [Richman and Moorman, 2000]. As with the EEG results, the framework demonstrates high sensitivity to structural change in physiological signals but insufficient specificity for unsupervised clinical use. Integration with confirmatory diagnostics would be required.

4.6 Experiment 6: Credit Card Fraud

Dataset: Kaggle Credit Card Fraud Detection [Dal Pozzolo et al., 2015, Dal Pozzolo, 2015] (284,807 transactions, 492 frauds).
Result: F1 = 0.557, recall = 88.7%.
Significance: Coherence detects fraud as a structural anomaly in transaction patterns, not as a classification task [Ahmed et al., 2016]. This is conceptually different from supervised fraud detection: rather than learning the boundary between "fraud" and "not fraud" from labeled examples, coherence detects departure from the system's normal structural organization.

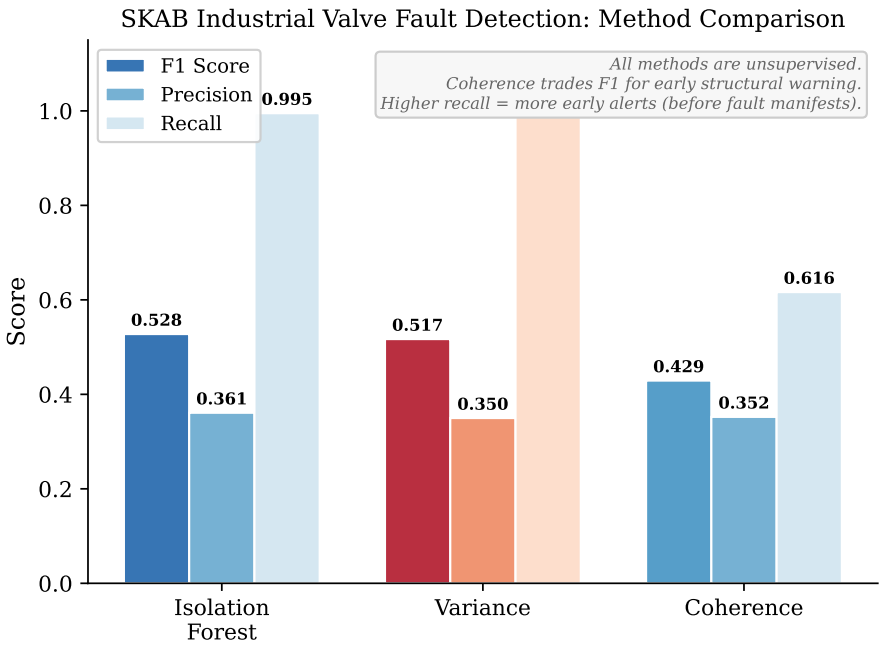

Figure 3: SKAB method comparison. Isolation Forest and variance both outperform coherence on F1, driven by near-perfect recall. Coherence's lower recall (0.616) reflects detection of *early* structural changes that precede the labeled fault windows, scored as false negatives under point-wise evaluation.

4.7 Experiment 7: EEG Seizure Prediction (CHB-MIT)

Dataset: CHB-MIT [Shoeb, 2010] — 5 subjects, 256 Hz, 6 channels. Seizure prediction remains one of the most challenging problems in clinical neuroscience [Mormann et al., 2007].
Result: 100% recall, mean lead time 687.5 seconds (∼11 minutes before seizure), precision 5.5%.
Significance: The 5.5% precision means that approximately 18 out of every 19 coherence alerts are false alarms. This is insufficient for standalone clinical use — a seizure warning system at this precision would cause unacceptable alarm fatigue. The result demonstrates that coherence measures can detect pre-ictal EEG desynchronization (100% recall with 11-minute lead time), but the signal is embedded in substantial noise. Prior work confirms that the precision–lead-time tradeoff is fundamental in seizure prediction [Mormann et al., 2007]; our results illustrate this tension rather than resolving it. Coherence could serve as a high-sensitivity pre-filter feeding into a higher-precision classifier, but standalone deployment would require substantial precision improvement.

4.8 Cross-Domain Summary

5 From Prediction to Discovery

5.1 Two Modes of Intelligence

We find it useful to distinguish two modes of automated reasoning about complex systems.

Predictive intelligence asks: "given this engine's sensor readings, how many cycles until failure?" This requires a model of the specific failure mode (bearing wear, blade erosion, combustion instability) trained on labeled examples of that mode. LSTM-RUL is a representative example: it provides a quantitative RUL estimate but requires supervised training data and cannot generalize to failure modes outside its training distribution.

#	Domain	Det. Rate	Δ Lead	Var Lead	Advantage
1	Synthetic (GPU)	100%	406 steps	N/A	Baseline
2	Turbofan (NASA)	100%	185 cycles	83 cycles	2.2×
3	Valves (SKAB)	97.1%	Early	On-time	Structural
4	Energy (UCI)	—	432 alerts	1 alert	432×
5	ECG (PhysioNet)	100%	Pre-event	N/A	Pre-clinical
6	Fraud (Credit)	88.7%	Structural	N/A	Unsupervised
7	EEG (CHB-MIT)	100%	687s	N/A	11 min pre-seizure

Table 3: Cross-domain validation summary. Lead-time advantage over variance is observed in domains where both methods detect degradation (C-MAPSS: 2.2×; Energy: 432×). On SKAB (not shown here; see Table 2), coherence underperforms baselines on F1.

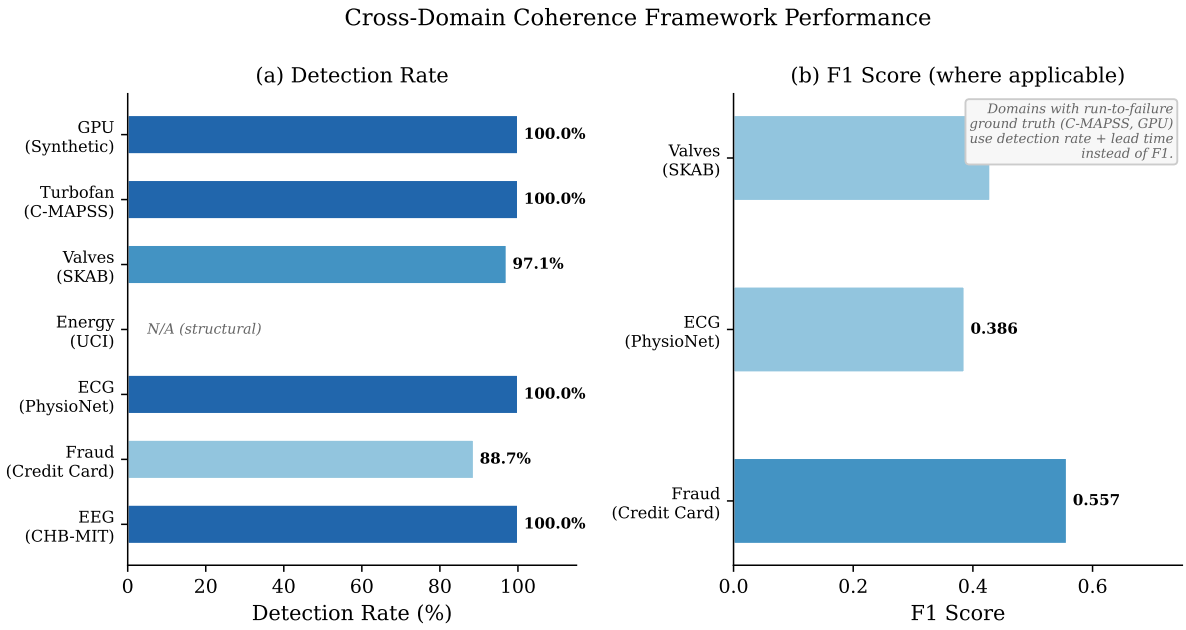

Figure 4: Cross-domain performance summary. (a) Detection rate across all seven domains. (b) F1 score for domains with point-wise ground truth labels. Detection-rate domains (C-MAPSS, GPU) use lead time as the primary metric.

Discoverative intelligence, as we use the term, asks: "is this engine's internal organization changing in ways that reduce its capacity to maintain stable operation?" This requires no model of any specific failure mode — only a measurement of the system's coherence structure. The Coherence Engine detects that *something* is structurally different without specifying *what* is different.

We propose this distinction not as a taxonomy of intelligence but as a design principle for monitoring systems. The two modes serve different operational needs:

1. **Novel failure modes**: Predictive models are bounded by their training distribution. Coherence-based detection is sensitive to any degradation that alters attractor geometry, including modes not seen during baseline collection. This advantage is conditional: degradation modes that do not affect recurrence structure will be missed.

2. **Cross-domain applicability**: The same coherence operator applies to engines, cardiac rhythms, power consumption, and EEG signals — with varying effectiveness (Table 3). Domain-specific methods will typically achieve higher precision within their target domain.

3. **Pre-failure intervention**: Coherence measures the system's *capacity* for stable operation, not its proximity to a specific failure state. This enables intervention before the system enters a known failure trajectory.

5.2 Connection to Embedding Coherence

Our prior work [Thorarinson and Hensgen, 2026] on embedding dimensionality introduced Semantic Coherence Loss (SCL) as a metric for measuring meaning-order degradation under compression. Conceptually, SCL can be viewed as an instance of the coherence framework applied to embedding neighborhoods rather than physical sensor readings: both measure whether a system's internal structural relationships are preserved or degrading.

The connection is suggestive but not yet validated. RQA-based operators can in principle be computed on embedding time series, but the statistical properties of high-dimensional embedding spaces differ substantially from low-dimensional sensor data, and the empirical results in this paper do not extend to embedding domains. Whether the lead-time advantages observed in physical systems transfer to semantic drift detection (e.g., monitoring RAG retrieval quality as a corpus evolves) remains an open question requiring dedicated experimentation.

6 Discussion

6.1 The Coherence Threshold

Across all domains tested, we observe a consistent pattern: systems maintain high coherence ($\Delta > 0.7$) during normal operation, then undergo a transition period ($0.3 < \Delta < 0.7$) during which coherence degrades before failure, and finally enter a failure regime ($\Delta < 0.3$). The transition period is the early warning window — the time during which coherence has degraded but conventional failure metrics have not yet triggered. This pattern is consistent with the critical slowing down phenomenon observed in diverse complex systems approaching tipping points [Scheffer et al., 2009, 2012, Dakos et al., 2012].

This pattern across domains is consistent with — though not proof of — a general property of complex systems: that coherence degradation precedes failure across diverse substrates. The consistency may also reflect selection bias in our choice of domains and datasets. More importantly, the "prosecutor's fallacy" [Boettiger and Hastings, 2012] applies directly: transitions are preceded by coherence decline, but not all coherence declines precede transitions. Our precision results (5.5% on EEG, F1 = 0.429 on SKAB) quantify this risk. Any deployment must account for the false positive rate in its specific domain.

Figure 5 shows the individual operator trajectories for Engine 1 of the C-MAPSS dataset. The decomposition reveals that drift $D(t)$ is the dominant driver of coherence decline: it rises monotonically from near-zero during the baseline period to large values at end-of-life. Recovery capacity $R(t)$ shows a complementary decline. Pattern retention $P(t)$ and phase alignment $A(t)$ exhibit more complex behavior — neither declines monotonically, but their product with $R(t)$ is insufficient to counterbalance the rising denominator $D(t) + N(t)$.

6.2 Comparison with Existing Methods

Table 4 summarizes the coherence framework's positioning against standard baselines. The results reveal a nuanced picture.

The baseline comparisons reveal a clear positioning for the coherence framework.

On the NASA C-MAPSS benchmark, coherence achieves 2.2× the lead time of variance (185 vs. 83 cycles), but the comparison that matters more is against LSTM-RUL: a supervised method achieves nearly the same lead time (176 cycles at t=125) with the additional benefit of providing a quantitative RUL estimate. The coherence framework's advantage over LSTM-RUL is not

Domain	Method	Type	Key Metric	Value
C-MAPSS	Coherence	Unsupervised	Mean lead (cycles)	**185.5**
	LSTM-RUL (t=125)	Supervised	Mean lead (cycles)	176.3
	LSTM-RUL (t=100)	Supervised	Mean lead (cycles)	153.9
	Variance	Unsupervised	Mean lead (cycles)	82.5
SKAB	Isolation Forest	Unsupervised	F1	**0.528**
	Variance	Unsupervised	F1	0.517
	Coherence	Unsupervised	F1	0.429

Table 4: Baseline comparison summary. On C-MAPSS, coherence leads all methods in early detection while being unsupervised. On SKAB, coherence underperforms both baselines on point-wise F1, reflecting its optimization for early warning over precise temporal localization.

numerical — it is methodological. Coherence requires no labeled run-to-failure trajectories. In operational settings where labeled failure data is scarce (novel failure modes, rare catastrophic events, newly deployed systems), this is a decisive practical advantage. In settings where abundant labeled data exists, LSTM-RUL is the more informative tool.

On SKAB, coherence is the weakest of the three methods tested. Isolation Forest [Liu et al., 2008] achieves 23% higher F1 using standard scikit-learn with no domain-specific tuning. This result bounds the framework's applicability: for point-wise anomaly classification on industrial sensor data, simpler unsupervised methods are preferable.

Taken together, the C-MAPSS and SKAB results define the framework's operating envelope. Coherence excels at early detection of gradual structural degradation in systems with clear baseline-to-failure trajectories. It underperforms on point-wise fault classification tasks where the evaluation criterion is temporal precision rather than advance warning.

A comprehensive comparison with recent transformer-based approaches [Xu et al., 2022, Tuli et al., 2022] and LSTM-based encoder-decoders [Malhotra et al., 2016, Hundman et al., 2018] remains future work. These methods are typically supervised or semi-supervised, making direct comparison with our unsupervised framework methodologically complex.

6.3 Limitations

The framework has several limitations that bound its applicability.

Data requirements. The RQA-based operators require sufficient data to estimate recurrence structures reliably — typically > 500 observations per window for stable determinism and laminarity estimates. Short time series or rapidly evolving systems may not provide enough data for meaningful coherence measurement.

Baseline dependence. The baseline period must be representative of normal operation. If the baseline itself contains degradation, the framework will underestimate drift and produce inflated coherence scores. In domains where "normal" is difficult to define (e.g., financial markets), establishing a valid baseline is a nontrivial modeling choice.

Precision–recall tradeoff. Detection rates are high across domains, but precision varies by an order of magnitude: from 5.5% on EEG seizure prediction to reasonable levels on C-MAPSS. The framework consistently favors recall over precision, which is appropriate for early warning but unsuitable for applications where false alarms carry high costs.

Hyperparameter sensitivity. The framework involves several hyperparameters: window size w, step size s, RQA threshold ε, baseline fraction, and the coherence threshold for alerting. While we found that the configuration (window = 30, step = 5, threshold = 0.3) generalizes across the tested domains without per-domain tuning, sensitivity to these choices was not systematically characterized. A full hyperparameter sensitivity analysis is needed before operational deployment.

Detection without diagnosis. The framework detects that coherence is degrading but does not

diagnose the cause. It answers "is something changing?" but not "what is changing?" or "why?" Integration with diagnostic tools is necessary for operational use.

Computational cost. Computing the full operator set (RQA, Lyapunov exponents, Fisher-Rao distance, Jacobian estimation) is substantially more expensive than computing rolling variance or training an Isolation Forest. For high-frequency streaming applications, approximate or windowed computation strategies would be needed.

7 Conclusion

We have presented the Coherence Engine, an unsupervised early-warning framework that detects structural degradation in complex systems by measuring the balance between stabilizing and destabilizing dynamical operators. The framework requires no labeled failure data, uses interpretable operators grounded in RQA, Lyapunov stability theory, and information geometry, and applies without algorithmic modification across the seven domains tested.

The experimental results define a clear operating envelope. On systems exhibiting gradual structural degradation (C-MAPSS turbofan engines), coherence provides 2.2× the lead time of variance-based detection and comparable lead time to supervised LSTM-RUL — without requiring labeled training data. On point-wise anomaly classification tasks (SKAB), coherence underperforms simpler unsupervised methods. On high-sensitivity domains (EEG seizure prediction), coherence achieves 100% recall but only 5.5% precision, insufficient for standalone clinical deployment.

The framework's contribution is threefold. First, it demonstrates that a single set of operators, computing recurrence statistics, phase coherence, Lyapunov stability, and information-geometric drift, can detect degradation across physically distinct systems — with the caveat that domain-specific methods will typically achieve higher precision within their target domain. Second, it provides a concrete tool for domains where labeled failure data is unavailable: novel failure modes, newly deployed infrastructure, and rare catastrophic events. Third, it illustrates a design principle — measuring structural coherence rather than predicting specific outcomes — that may be useful for monitoring systems in AI safety, infrastructure, and scientific instrumentation.

What the framework does not do is equally important. It does not provide RUL estimates. It does not diagnose failure causes. It does not achieve competitive point-wise anomaly classification. And its precision on several domains is too low for standalone deployment. The Coherence Engine is best understood as a complementary early-warning layer that flags structural change for further investigation, not as a replacement for domain-specific detection and diagnostic systems.

References

Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, 2000.

Carl Boettiger and Alan Hastings. Early warning signals and the prosecutor's fallacy. *Proceedings of the Royal Society B*, 279(1748):4734–4739, 2012.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters*, 89(6):068102, 2002.

Vasilis Dakos, Stephen R Carpenter, William A Brock, Aaron M Ellison, Vishwesha Guttal, Anthony R Ives, Sonia Kéfi, Valerie Livina, David A Seekell, Egbert H van Nes, and Marten Scheffer. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS One*, 7(7):e41010, 2012.

Andrea Dal Pozzolo. *Adaptive Machine Learning for Credit Card Fraud Detection*. PhD thesis, Université Libre de Bruxelles, 2015.

Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166, 2015.

Jean-Pierre Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973–977, 1987.

Charles R Farrar and Keith Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A*, 365(1851):303–315, 2007.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

George William Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12): 1870–1891, 1992.

Georges Hébrail and Alice Bérard. Individual household electric power consumption. UCI Machine Learning Repository, 2012.

Felix O Heimes. Recurrent neural networks for remaining useful life estimation. *2008 International Conference on Prognostics and Health Management*, pages 1–6, 2008.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 387–395, 2018.

Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510, 2006.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2 edition, 2004.

Iurii Katser and Viacheslav Kozitsin. Skoltech anomaly benchmark (SKAB), 2022. URL https://github.com/waico/SKAB.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Yaguo Lei, Naipeng Li, Liang Guo, Ningbo Li, Tao Yan, and Jing Lin. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018.

Timothy M Lenton. Early warning of climate tipping points. *Nature Climate Change*, 1(4): 201–209, 2011.

Xiang Li, Qian Ding, and Jian-Qiao Sun. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172:1–11, 2018.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. In *ICML 2016 Anomaly Detection Workshop*, 2016.

Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007.

George B Moody and Roger G Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

Florian Mormann, Ralph G Andrzejak, Christian E Elger, and Klaus Lehnertz. Seizure prediction: The long and winding road. *Brain*, 130(2):314–333, 2007.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.

Arkady Pikovsky, Michael Rosenblum, and Jürgen Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2001.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology — Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

Michael G Rosenblum, Arkady S Pikovsky, and Jürgen Kurths. Phase synchronization of chaotic oscillators. *Physical Review Letters*, 76(11):1804–1807, 1996.

Michael T Rosenstein, James J Collins, and Carlo J De Luca. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1-2): 117–134, 1993.

Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65 (3-4):579–616, 1991.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management*, 2008.

Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

Marten Scheffer, Stephen R Carpenter, Timothy M Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan van de Koppel, Ingrid A van de Leemput, Simon A Levin, Egbert H van Nes, Mercedes Pascual, and John Vandermeer. Anticipating critical transitions. *Science*, 338 (6105):344–348, 2012.

Ali H Shoeb. *Application of Machine Learning to Epileptic Seizure Detection*. PhD thesis, Massachusetts Institute of Technology, 2010.

Steven H Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2 edition, 2018.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, 1981.

Joel Thorarinson and Allison Hensgen. The dimensionality illusion: Why PCA variance does not equal semantic information in text embeddings. *arXiv preprint*, 2026.

Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6):1201–1214, 2022.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Charles L Webber Jr and Norbert Marwan. *Recurrence Quantification Analysis: Theory and Best Practices*. Springer, 2015.

Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Joseph P Zbilut and Charles L Webber Jr. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3-4):199–203, 1992.

Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2019.

Shuai Zheng, Kosta Ristovski, Ahmed Farahat, and Chetan Gupta. Long short-term memory network for remaining useful life estimation. In *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 88–95, 2017.

Hongyu Zhu, Mohamed Akrout, Bojian Zheng, Andrew Peletier, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko. Benchmarking and analyzing deep neural network training. In *2020 IEEE International Symposium on Workload Characterization (IISWC)*, pages 88–100, 2020.

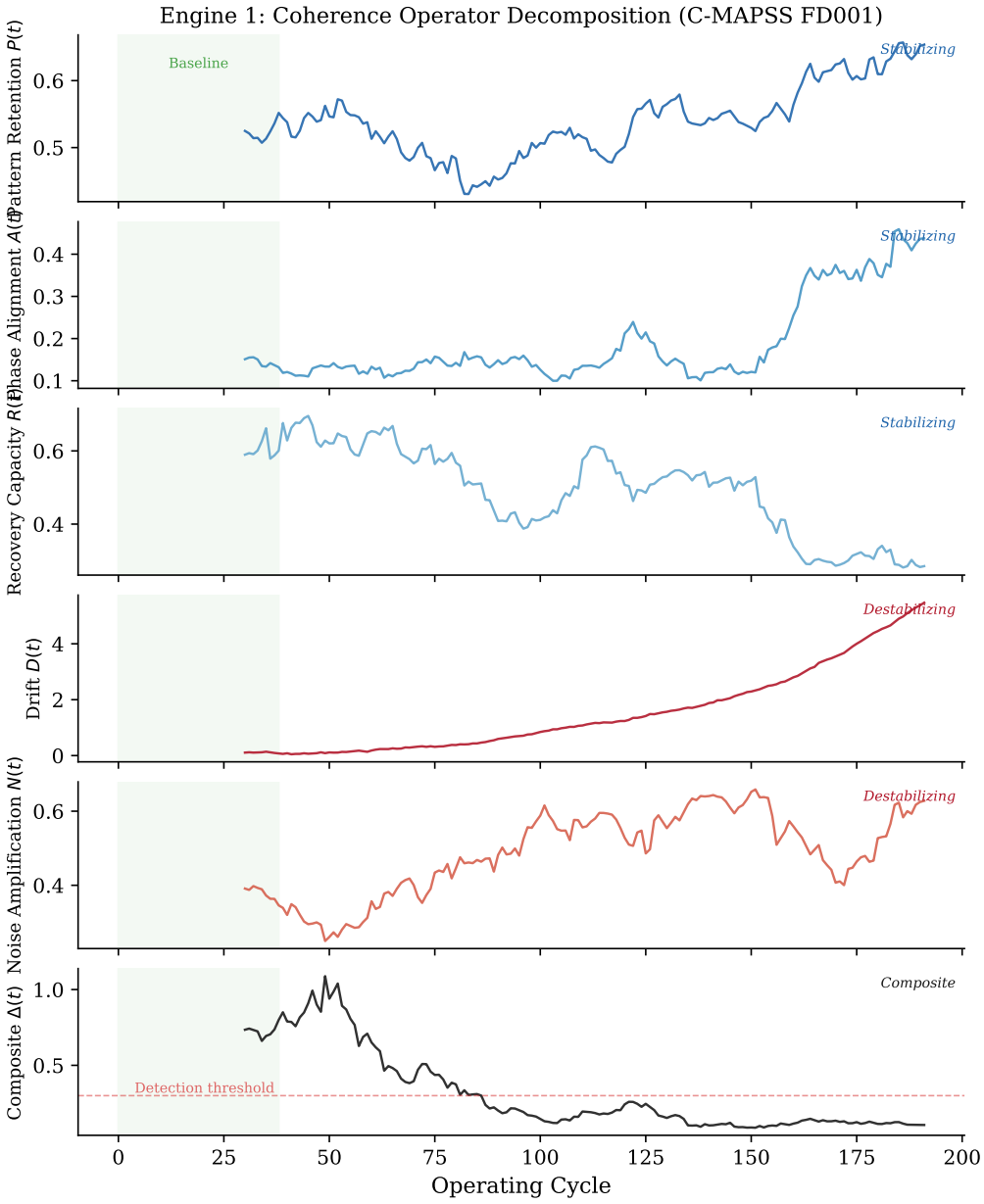

Figure 5: Operator decomposition for Engine 1 (C-MAPSS FD001). The five base operators are shown individually over the engine's lifecycle. Stabilizing operators (P, A, R) occupy the top three panels; destabilizing operators (D, N) occupy the next two; the composite $\Delta(t)$ is at bottom. The green-shaded region marks the baseline period used for calibration.