www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(3), 2026, 1-36. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i3.857

Topological and Self-Structured Approaches to Supervised Anomaly Detection in Econometrics

Jaime Aguilar-Ortiz[1,2], Carlos R. Domínguez-Mayorga[1], Ocotlán Diaz-Parra[1], Jorge A. Ruiz-Vanoye[1], Francisco R. Trejo-Macotela[1], Marco A. Vera-Jiménez[1], Víctor M. Zamudio-García[2]
[1]Universidad Politécnica de Pachuca, México
[2] Universidad Politécnica Metropolitana de Hidalgo, México
jao@upp.edu.mx, carlos.dominguez@upp.edu.mx, ocotlan_diaz@upp.edu.mx, jorgeruiz@upp.edu.mx, trejo_macotela@upp.edu.mx, marcovera@upp.edu.mx, vzamudio@upmh.edu.mx

Abstract. This article proposes a robust econometric framework for anomaly detection in nonstationary time series affected by noise, outliers, and regime shifts. The method combines windowed feature construction, supervised learning, and stability-oriented regularization, while enabling optional topological and structural diagnostics to corroborate detected transitions. A reproducible pipeline trains models, calibrates decision thresholds, and preserves artifacts for transparent validation, including metrics, figures, and segment-level summaries. Experiments show consistent discrimination, improved reliability under distributional change, and interpretable latent representations that support operational monitoring. The results demonstrate that integrating robustness principles with structured diagnostics yields actionable early-warning signals for complex dynamic systems in practice.
Keywords: Robust econometrics, anomaly detection, regime shifts, representation learning, reproducible pipelines, Algebraic Topology.

Article Info
Received Ene 26, 2026
Accepted Mar 11, 2026

1 Introduction

1.1. Motivation: Robust econometric anomaly detection under regime shifts and contamination

Econometric time series from socio-technical and financial systems rarely meet assumptions such as stationarity, light tails, or clean measurement; instead, they show structural breaks, regime drift, heavy-tailed shocks, and intermittent outliers, so an operationally credible anomaly detector must be robust to misspecification and to evolving "normality" (Truong et al., 2020). The motivation is therefore to integrate, in a single reproducible pipeline, windowed multivariate feature construction with robust standardization and quality control, reconstruction-based representation learning that yields a stable error-driven monitoring signal (Zamanzadeh Darban et al., 2024), and topology-informed diagnostics that quantify geometric/connectivity shifts in embedded trajectories indicative of regime change (Ichinomiya, 2025; Yao et al., 2025). Practically, reconstruction error is decomposed and smoothed and paired with changepoint logic to curb false alarms under volatility, while persistent-homology distances (e.g., H0–H2) provide orthogonal evidence to distinguish structural change from transient noise (Truong et al., 2020; Yao et al., 2025).

1.2. Objectives, contributions, and paper organization

The article pursues three coupled objectives: to frame a robust, window-based anomaly detector for econometric series where regime shifts, heavy-tailed noise, and transient contamination are treated as core conditions; to implement it as a fully reproducible pipeline that exports inspectable artifacts so every Results claim is traceable; and to provide interpretable diagnostics that distinguish variance-driven "score inflation" from true structural change by combining robust preprocessing, reconstruction scoring, and changepoint segmentation

(Truong et al., 2020). The contributions follow: an integrated detector joining robust econometric summaries with deep reconstruction signals; a topology-aware layer using persistent-homology summaries and stable window-wise distances to add shape-level regime evidence; and a standardized reporting design that exports narrative-ready outputs (window metrics, decomposed/smoothed scores, alerts, segmentation markers) aligned with survey taxonomies and time-series PH practice (Zamanzadeh Darban et al., 2024; Ravishanker & Chen, 2021). The paper then organizes motivation, foundations, methodology, pipeline architecture, exported results (scores, decompositions, topology distances, alert/segmentation performance), and a discussion of implications, limitations, and extensions (e.g., richer topology, dependence shifts, scalable deployment).

2 Theoretical foundations

For readability, the revised manuscript standardizes the symbols used most frequently in Sections 2–5 through the following concise notation summary (see Table 1).

Table 1. Concise notation used across Sections 2–5.

Symbol	Meaning	First use
t	raw time index in the original series	Sections 2.1–2.3
j	window index after sliding-window transformation	Section 3.1
$x^{(j)}$	feature vector associated with window *j*	Section 3.1
y_j	window label derived from the within-window anomaly ratio	Section 3.1
z_j	latent representation produced by the encoder	Sections 2.2 and 5.2
$\hat{p}_j$	calibrated anomaly probability for window *j*	Sections 2.2, 3.2 and 5.1
d_j or $\tilde{d}_j$	normalized topological distance (e.g., Wasserstein/H_1 drift) used as structural evidence	Sections 2.3 and 5.3
S_j	composite anomaly score obtained by fusing probability and topology	Sections 3.2 and 5.3
α	fusion weight selected on validation data under the alerting criterion	Sections 3.2 and 5.3

2.1. Robust econometrics for anomaly detection: stability, resistance to outliers, and interpretability

Robust econometric anomaly detection is naturally framed as inference on a time series x_t (or a multivariate vector $x_t \in R^d$) that is exposed to regime shifts and contamination, where classical estimators may become unstable because a small fraction of extreme observations can disproportionately change fitted parameters and, therefore, the anomaly score itself (Rousseeuw & Hubert, 2018). A standard robust construction begins by operating on sliding windows $W_k = \{t_k, \dots, t_k + L\text{-}1\}$ and estimating a resistant location and scale, e.g., $m_k = \text{median}\{x_t : t \in W_k\}$ and $s_k = 1.4826\,\text{median}\{|x_t\text{-}m_k| : t \in W_k\}$, which yields the robust standardized residual $z_t = (x_t\text{-}m_k)/s_k$ and supports bounded "stress" transforms such as $\tilde{z}_t = \text{clip}(z_t\text{-}c, c)$ for some $c > 0$; in robust statistics, this clipping is a concrete way to control the influence of extremes on downstream decisions while preserving comparability across windows (Rousseeuw & Hubert, 2018). Equivalent robustness can be expressed via M-estimation, where parameters θ (location, regression coefficients, or state parameters) are obtained by minimizing $\sum_{i \in W_k} \rho\left(r_i(\theta)\right)$ with bounded-loss ρ, most commonly the Huber function $\rho_c(r) = \frac{1}{2} r^2\, 1(|r| \le c) + \left(c|r| - \frac{1}{2}c^2\right) 1(|r| > c)$, whose score $\psi_c(r) = \rho_c'(r) = r\, 1(|r| \le c) + c\,\text{sign}(r)\, 1(|r| > c)$ enforces bounded influence and thereby stabilizes estimation under heavy tails and outliers (Rousseeuw & Hubert, 2018). When the econometric objective explicitly includes regime shifts, robust change-point formulations treat the trajectory as piecewise regular by selecting change points $\tau_1 < \cdots < \tau_m$ that minimize a penalized segmentation criterion $\min_{\tau_{1:m}} \sum_{j=0}^{m} C\left(x_{\tau_j:\tau_{j+1}\text{-}1}\right) + \beta m$, where $C(\cdot)$ can be made outlier-resistant by using robust within-segment costs (e.g., Huberized least squares or median-based deviations), and β regularizes over-segmentation; this yields a principled mechanism for distinguishing abrupt structural breaks from transient contamination, which is central to stable anomaly scoring in nonstationary environments (Fearnhead & Rigaill, 2019; Truong et al., 2020).

Operational robustness, however, is not only a property of a single estimator; it is a property of an end-to-end scoring pipeline that enforces stability constraints at multiple stages and then exposes interpretable diagnostics. For example, once a probabilistic detector produces window-level anomaly probabilities $p_k \in [0,1]$, a calibration-aware robustness layer evaluates decisions with proper scoring rules such as the Brier score $BS = \frac{1}{n}\sum_{k=1}^{n}(p_k - y_k)^2$ (with labels $y_k \in \{0,1\}$), and compares pre- and post-calibration reliability via calibration curves an especially relevant safeguard when anomalies are rare or labels are noisy (Huang et al., 2020). A stability-oriented active-selection proxy can then be defined by a symmetric uncertainty functional such as $u(p_k) = 1 - 2|p_k - 0.5|$, which peaks at $p_k = 0.5$ and vanishes at confident extremes; combined with robustly normalized stress scores $z_k = (s_k - \text{median}(s))/(1.4826\,\text{median}(|s - \text{median}(s)|))$ and validity constraints expressed through quantile-bounded acceptance regions $l_j \le x_{k,j} \le u_j$ with $l_j = Q_\alpha(X_j)$, $u_j = Q_{1-\alpha}(X_j)$, this produces a detector whose candidate anomalies are selected not merely for magnitude but for informative and statistically plausible departures, reducing spurious alerts under contamination (Rousseeuw & Hubert, 2018; Truong et al., 2020). Interpretability is then realized by coupling transparent probabilistic models e.g., logistic regression $\log\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{k,j}$, where β_j gives a signed, directly inspectable association with ensemble models such as random forests whose predictive behavior can be interrogated through permutation importance $PI_j = E[L(f, X, Y) - L(f, X^{\pi(j)}, Y)]$, where $X^{\pi(j)}$ is formed by permuting feature j and L is a loss; these tools expose which window statistics (e.g., mean, volatility, extrema, slope) are driving alerts and whether those drivers remain stable across regimes (Guidotti et al., 2018).

2.2. Learning-based anomaly detection: representation learning, decision boundaries, and calibration

Learning-based anomaly detection for econometric time series can be formalized as a representation-learning problem in which the observed windowed vector $x_t \in R^d$ (built from standardized and temporally structured covariates) is mapped into a low-dimensional latent state $z_t \in R^k$ that preserves the salient dynamics while attenuating noise and contamination; in the canonical autoencoder setting, an encoder $f_\theta: R^d \to R^k$ and decoder $g_\phi: R^k \to R^d$ yield $\hat{x}_t = g_\phi(f_\theta(x_t))$, and the window-level reconstruction discrepancy $r_t = \frac{1}{d}\| x_t - \hat{x}_t \|_2^2$ becomes an unsupervised anomaly signal, whereas a supervised head $h_\psi(z_t) = \sigma(w^T z_t + b)$ produces a probabilistic boundary for rare-event detection via $\hat{p}_t = P(y_t = 1 \mid x_t)$ and the decision rule $\hat{y}_t = I[\hat{p}_t \ge \tau]$, with $\tau \in (0,1)$ selected by operational criteria (e.g., maximizing F_1 on a validation regime or targeting a false-alarm budget). A multi-task formulation is particularly consistent with pipelines that jointly report a reconstruction channel and an anomaly-probability channel: a composite objective can be written as

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{geo}\mathcal{L}_{geo}$$

where

$$\mathcal{L}_{rec} = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{d}\| x_t - \hat{x}_t \|_2^2,\ \mathcal{L}_{cls} = -\frac{1}{n}\sum_{t=1}^{n}[y_t \log(\hat{p}_t) + (1 - y_t)\log(1 - \hat{p}_t)]$$

and the geometry-preserving regularize can be instantiated by pairwise distance alignment, e.g.,

$$\mathcal{L}_{geo} = \frac{1}{|P|}\sum_{(i,j)\in P}\left(\| z_i - z_j \|_2 - \alpha \| x_i - x_j \|_2\right)^2$$

(or an equivalent normalized-distance variant), which operationally encourages decision boundaries to form in a latent space that remains interpretable as a deformation of the original window geometry rather than an arbitrary embedding. Within this framing, "decision boundaries" are not merely threshold lines on raw scores but hypersurfaces $\{z: h_\psi(z) = \tau\}$ whose stability under regime shifts can be inspected through separability

diagnostics and through the joint evolution of $\hat{p}_t$ and r_t aligning naturally with deep anomaly-detection taxonomies that emphasize representation quality, hybrid objectives, and temporally aware scoring for time series (Pang et al., 2021; Tuli et al., 2022; Zamanzadeh Darban et al., 2024; Zhang & Yang, 2021).

Calibration is the complementary requirement that converts a high-performing boundary into an actionable probabilistic instrument, because anomaly response policies typically consume probabilities as risk measures (for alerting, escalation, or changepoint investigation) rather than as abstract classifier scores; formally, a probabilistic predictor $\hat{p}$ is calibrated when $P(Y = 1 \mid \hat{p} = s) = s$ for $s \in [0,1]$, and deviations from this identity are visualized by reliability diagrams that bin predictions and compare empirical frequencies to the diagonal, while scalar summaries can include proper scoring rules such as the Brier score $BS = \frac{1}{n}\sum_{t=1}^{n}(\hat{p}_t - y_t)^2$, which penalizes both misclassification and overconfidence. In practice, post-hoc calibration maps a raw score s_t(e.g., a logit, margin, or model score) to a probability via a monotone link, with Platt-type sigmoid calibration $\hat{p}_t = \frac{1}{1+\exp(As_t+B)}$ providing a parsimonious parametric option and isotonic regression providing a flexible nonparametric alternative; the choice is nontrivial in anomaly detection because positives are typically scarce and distributional shifts are expected, so calibration methods tailored to severe imbalance and stability considerations are particularly relevant for ensuring that $\hat{p}_t$ can be interpreted consistently across regimes (Böken, 2021; Dimitriadis et al., 2021; Guilbert & Caelen, 2024).

2.3. Topological and structural diagnostics: persistent signatures as evidence of regime organization

Topological and structural diagnostics formalize "regime organization" as a reproducible geometry of trajectories in an observation or representation space, where qualitative shifts are evidenced not only by changes in marginal moments but by changes in the shape of the sampled dynamics across scales. Given a multivariate time series $\{x_t\}_{t=1}^{T} \subset R^m$, each analysis step constructs, for every time index (or window end) t, a point cloud that captures local dynamics, e.g., by sliding-window stacking $X^{(t)} = \{x_{t-w+1}, \ldots, x_t\}$ or by delay embedding $v_t = (x_t, x_{t-\tau}, \ldots, x_{t-(d-1)\tau}) \in R^{md}$. From this point cloud $P^{(t)} = \{p_i^{(t)}\}_{i=1}^{n_t} \subset R^d$ (which may be defined directly in observation space or in a learned latent space), a Vietoris–Rips filtration is formed by $VR_\varepsilon(P^{(t)}) = \{\sigma \subseteq P^{(t)} : \| p_i - p_j \| \leq \varepsilon \; \forall \, p_i, p_j \in \sigma\}$, inducing a nested family $VR_{\varepsilon_1} \subseteq VR_{\varepsilon_2} \subseteq \cdots$ as ε increases. Persistent homology then tracks the birth and death of topological features across ε through homology groups $H_q(VR_\varepsilon)$ through homology groups $\beta_q(\varepsilon) = \mathrm{rank}H_q(VR_\varepsilon)$ (connected components for $q = 0$, loops for $q = 2$) summarized in a persistence diagram $D_q^{(t)} = \{(b_i^{(q,t)}, d_i^{(q,t)})\}_i$ and an equivalent Betti curve $\beta_q^{(t)}(\varepsilon) = \sum_i 1\{b_i^{(q,t)} \leq \varepsilon < d_i^{(q,t)}\}$. In this framing, a "regime" is characterized by stable, repeatable diagram structure (notably long-persistence features), while regime change corresponds to systematic reconfiguration of $D_q^{(t)}$ and $\beta_q^{(t)}(\varepsilon)$, a viewpoint widely used to interpret time-series organization via persistent signatures (Chazal & Michel, 2021; Dey & Wang, 2022; Ravishanker & Chen, 2021; Heo & Jung, 2024).

Operationally, persistent signatures become "diagnostics" once they are converted into comparable time-indexed statistics whose dynamics can be contrasted across windows, classes, or phases: for consecutive windows t-1 and t, a principled measure of structural reorganization is the p-Wasserstein distance between diagrams $D_q^{(t-1)}$ and $D_q^{(t)}$,

$$W_p\left(D_q^{(t-1)}, D_q^{(t)}\right) = \left(\inf_\gamma \sum_{u \in D_q^{(t-1)}} \| u - \gamma(u) \|^p \right)^{1/p}$$

where γ ranges over bijections between diagram points augmented with diagonal projections, and the bottleneck distance
$d_B(D_q^{(t-1)}, D_q^{(t)}) = \inf_\gamma \sup_u \| u - \gamma(u) \|_\infty$,

provides a complementary "largest deviation" summary (Bauer, 2021; Chazal & Michel, 2021). In time-series regimes, $q = 1$ is often particularly diagnostic because loop structure reflects recurrent or cyclical organization: an increase in W_p on H_1 can be interpreted as a deformation of periodic/recurrence geometry rather than a mere amplitude excursion, while $\beta_1^{(t)}(\varepsilon)$ shifting its mass across ε indicates that loop features are appearing at different scales (Heo & Jung, 2024; Ravishanker & Chen, 2021). When these diagnostics are tracked as sequences $\{W_p(D_1^{(t-1)}, D_1^{(t)})\}_t$ and $\{\beta_q^{(t)}(\varepsilon)\}_t$ and optionally aggregated by integrals such as

$$A_q^{(t)} = \int_{\varepsilon_{min}}^{\varepsilon_{max}} \beta_q^{(t)}(\varepsilon)\, d\varepsilon$$

or discretized L^2 norms

$$\| \beta_q^{(t)} - \beta_q^{(t-1)} \|_2 \approx \left(\sum_j \left(\beta_q^{(t)}(\varepsilon_j) - \beta_q^{(t-1)}(\varepsilon_j) \right)^2 \Delta\varepsilon \right)^{1/2}$$

they yield a structural channel that can be fused with reconstruction- or residual-based evidence to produce a composite stability/anomaly score whose changepoints mark candidate regime transitions. This regime-aware reading is consistent with recent applied work in financial and economic time series where topological features extracted via persistent homology are explicitly leveraged to detect extreme-event-induced change points and crisis-like transitions, supporting the interpretation of large topological distances as reorganizations of market state rather than isolated outliers (Yao et al., 2025; Zhang et al., 2025), while inference-focused developments emphasize that diagram collections can be analyzed in time-dependent settings with resampling strategies that respect temporal dependence, reinforcing the legitimacy of time-indexed topological monitoring under non-i.i.d. contamination (Abdallah et al., 2022).

Proposition 1. Let consecutive regimes R_a and R_b generate window embeddings whose persistence diagrams satisfy a regime gap larger than the perturbation scale induced by heavy-tailed noise, i.e., when the expected Wasserstein or bottleneck separation between regimes exceeds the corresponding within-regime perturbation bound. Under this condition, stable increases in H_0/H_1 distances across adjacent windows provide evidence of structural regime change rather than isolated contamination. In the econometric setting studied here, H_0 captures fragmentation/connectivity shifts, whereas H_1 is sufficient as a compact descriptor of recurrence or cyclical reorganization.

This framing is consistent with the stability theorems of persistent homology: bounded perturbations of the embedding induce bounded perturbations of the associated diagrams, so topological peaks should be interpreted as meaningful only when they are sustained across windows and larger than the local perturbation scale (Bauer, 2021; Chazal & Michel, 2021; Dey & Wang, 2022). Consequently, the topological layer is not treated merely as a descriptive add-on, but as a structurally stable diagnostic channel that complements supervised prediction in nonstationary econometric series (Ravishanker & Chen, 2021; Heo & Jung, 2024).

3 Methodology

3.1. Data protocol and experimental design: windowing, labeling, splits, and evaluation criteria

A rigorous data protocol for robust econometric anomaly detection starts from a multivariate series $\{x_t\}_{t=1}^T$, $x_t \in R^d$, and converts it into a temporally ordered set of overlapping windows to preserve local dynamics while enabling stable training and evaluation: for window length W and hop H, the k-th window begins at $s_k = 1 + (k-1)H$ and collects $X^{(k)} = \{x_{s_k}, \ldots, x_{s_k+W-1}\}$, from which the feature vector f_k is constructed by concatenating per-variable summaries such as

$$\mu_{k,j} = \frac{1}{W}\sum_{i=0}^{W-1} x_{s_k+i,j}\,,\ \sigma_{k,j} = \sqrt{\frac{1}{W}\sum_{i=0}^{W-1}\left(x_{s_k+i,j} - \mu_{k,j}\right)^2}\,,\ \min_{k,j} = \min_{0\le i<W} x_{s_k+i,j},\ \max_{k,j} = \max_{0\le i<W} x_{s_k+i,j}$$

and a linear-trend (slope) coefficient computed by least squares $b_{k,j} = \frac{\sum_{i=0}^{W-1}(i-\bar{i})\left(x_{s_k+i,j}-\mu_{k,j}\right)}{\sum_{i=0}^{W-1}(i-\bar{i})^2}$ with $\bar{i} = \frac{W-1}{2}$; labels are made consistent with windowing by starting from pointwise ground truth $y_t \in \{0,1\}$ and defining the within-window anomaly fraction $\rho_k = \frac{1}{W}\sum_{i=0}^{W-1} y_{s_k+i}$, so the window label becomes $Y_k = I[\rho_k \ge \gamma]$ for a chosen ratio threshold $\gamma \in (0,1)$, a design that supports rare-event detection while acknowledging that econometric anomalies often occur in bursts rather than as isolated points (Zamanzadeh Darban et al., 2024; Pang et al., 2021). To avoid leakage, the experimental split is performed temporally by ordering windows by s_k and allocating the final $n_{test} = \lceil \alpha n \rceil$ windows to a held-out test set (with $\alpha \in (0,1)$ and n total windows), leaving the earlier $n\text{-}n_{test}$ windows for model fitting, which aligns with established cautions that naive random splitting or unblocked resampling can invalidate performance claims when temporal dependence is present (Bergmeir et al., 2018); within the training portion, a validation fraction $\nu \in (0,1)$ is reserved for early stopping and model selection, enabling reproducible tuning without peeking at the test tail. Evaluation criteria are then computed on the test windows using probabilistic outputs $\hat{p}_k = P(Y_k = 1 \mid f_k)$ and thresholding $\hat{Y}_k(\tau) = I[\hat{p}_k \ge \tau]$ across a sweep of τ, yielding confusion counts (TP, FP, TN, FN) and derived rates $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$, precision $Prec = \frac{TP}{TP+FP}$, and recall $Rec = TPR$, with summary discrimination reported by ROC-AUC and average precision (PR-AUC), and regime-change relevance supported by complementing window classification with changepoint-aware diagnostics on the resulting score trajectory (Truong et al., 2020; Zamanzadeh Darban et al., 2024). Because alerting policies consume probabilities as risk-like quantities, calibration is assessed via reliability analysis and proper scoring, notably the Brier score $BS = \frac{1}{n_{test}}\sum_{k\in test}(\hat{p}_k - Y_k)^2$ using stable binning strategies for reliability diagrams to reduce sampling artifacts in rare-event settings (Dimitriadis et al., 2021) and considering imbalance-aware calibration practice when anomalies are sparse (Guilbert et al., 2024).

3.2. Model training and robust mechanisms: multi-objective learning, verification, and stress-testing

The training stage implements a robustness-first learning design in which anomaly probability estimation and faithful reconstruction are optimized jointly, so that detection remains stable under regime shifts, heavy-tailed noise, and localized contamination; concretely, for each standardized window $x \in R^{T\times d}$ the model produces an encoding $z = f_\theta(x)$, a reconstruction $\hat{x} = g_\theta(z)$, and an anomaly probability $\hat{p} = \sigma(h_\theta(z))$, and it minimizes a multi-objective criterion of the form

$$\mathcal{L}(\theta) = \mathcal{L}_{cls} + \lambda_{recon}\mathcal{L}_{recon} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_2 \parallel \theta \parallel_2^2$$

where

$$\mathcal{L}_{cls} = -[y\log(\hat{p}) + (1-y)\log(1-\hat{p})]$$

(with optional label smoothing and class-imbalance weighting)

$$\mathcal{L}_{recom} = \frac{1}{Td}\parallel x - \hat{x} \parallel_2^2$$

and the geometric "stress" term regularizes representation drift by enforcing approximate pairwise-distance consistency across input and latent spaces on random mini-batch pairs (i,j): letting $d_x = \parallel x_i - x_j \parallel_2$ and $d_z = \parallel z_i - z_j \parallel_2$ (both normalized by batch means), the implemented penalty is

$$\mathcal{L}_{geo} = \mathbb{E}_{(i,j)}[\max\{0, |d_x - d_z| - m\}^2]$$

which discourages latent collapses and improves out-of-distribution stability without sacrificing interpretability of reconstruction error trajectories (Zhang & Yang, 2021; Duque, Giraldo, & Arbeláez, 2023). Robustness is further strengthened through a verification-and-stress-testing loop that constructs hard candidates and then filters them with explicit plausibility constraints before curriculum injection: candidates are generated by perturbing nominal windows (e.g., additive noise $\tilde{x} = x + \varepsilon$ con $\varepsilon \sim N(0, \sigma^2)$ and clipping), then rejected unless they satisfy amplitude and energy guards $\| \tilde{x} \|_\infty \leq a$ and $\| \tilde{x} \|_2 \leq b$, a standardized-energy bound $\| \tilde{z} \|_2 \leq b_z$, and feature wise empirical-quantile bounds $l_j \leq \tilde{x}_{t,j} \leq u_j$ where $l_j = Q_\alpha\left(X_{:,j}^{\text{train}}\right)$ and $u_j = Q_{1-\alpha}\left(X_{:,j}^{\text{train}}\right)$ are learned from the training distribution, thereby preventing the training signal from being dominated by physically/financially implausible artifacts while still exposing the learner to controlled adversarial-like difficulty (Xu et al., 2023; Olteanu, Rossi, & Yger, 2023). Selection of verified candidates is then driven by a composite hardness score that prioritizes uncertain predictions and distributional stress (e.g., uncertainty $u(\hat{p}) = 1\text{-}2|\hat{p}\text{-}0.5|$ combined with a standardized distance-to-nominal proxy $s(\tilde{x})$, so that high-uncertainty/high-stress samples are injected more often), yielding a principled curriculum schedule in which training progresses from "easy/clean" to "hard/contaminated" without destabilizing optimization (Soviany, Ionescu, Rota, & Sebe, 2022). Finally, to avoid overfitting to transient anomalies or to the synthetic hardness distribution, the training loop uses validation-driven stopping criteria (monitoring the anomaly head's validation loss and restoring the best checkpoint) as a conservative safeguard that complements the explicit verifier and curriculum pacing, improving generalization of anomaly scores under no stationarity (Ferro, Doval Mosquera, Ribadas Pena, & Darriba Bilbao, 2023), and the overall methodology aligns with best practices in time-series anomaly detection where robustness requires combining resistant objectives, explicit data-quality constraints, and carefully designed exposure to difficult cases rather than relying on a single loss component (Zamanzadeh Darban et al., 2024).

To make the integration between predictive and topological evidence explicit, the revised pipeline defines a composite score that combines the calibrated anomaly probability with the normalized topological drift measured on the corresponding window.

$$s_j = \alpha \hat{p}_j + (1 - \alpha)\tilde{d}_j$$

Here, $\tilde{d}_j$ denotes the validation-normalized topological distance and $\alpha \in [0,1]$ is selected on the validation portion to maximize operational usefulness under the chosen alerting criterion (e.g., F1, Brier score, or false-alarm budget). Under this rule, topology acts as a corroborative escalation channel whenever $\hat{p}_j$ is near threshold and $\tilde{d}_j$ exhibits a local peak, thereby helping to flag candidate structural false negatives without replacing the primary supervised decision rule.

From a computational standpoint, if each window yields n embedded points in ambient dimension d, the pairwise-distance stage requires $O(n^2 d)$ operations, whereas the truncated Vietoris–Rips filtration and persistent-homology reduction may grow super-quadratically and, in the worst case, combinatorially with n. For W windows, the practical runtime is therefore dominated by $W \cdot (n^2 d + C\_PH(n))$, while memory is driven by the $O(n^2)$ distance matrix plus filtration storage; in large-scale deployments, this motivates either capped point-cloud size, sub-sampling, or intermittent rather than exhaustive TDA evaluation.

4 Computational architecture

4.1. Reproducible "save-everything" pipeline: artifacts, manifests, models, and deterministic reporting

A reproducible "save-everything" pipeline is operationalized as an end-to-end, automated computational workflow that treats every reported claim as an inspectable artifact, so that a single execution deterministically regenerates the full evidence bundle trained models, calibrated scores, diagnostic plots, and tabular reports under fixed configuration and controlled randomness (Ziemann et al., 2023; Rule et al., 2019). Concretely, the run produces a structured results directory (e.g., figures/, reports/, models/) and a machine-readable manifest (manifest.json) that enumerates the artifact set $A = \{a_i\}_{i=1}^m$ together with provenance fields such as paths p_i

generation status $g_i \in \{\text{real}, \text{placeholder}\}$, and minimal metadata μ_i, i.e., $M(a_i) = (p_i, g_i, \mu_i)$; this explicit inventory is designed so auditors can verify completeness even when optional components are unavailable, because the pipeline still emits "expected" files (as placeholders) and records the reason in the manifest, preventing silent omissions that accumulate into reproducibility debt (Hassan et al., 2025). Determinism is enforced by fixing pseudo-random generators with a seed s so that stochastic steps become functions of (D, s), i.e. $\hat{y} = f(D; s)$, and by writing reports in a deterministic order (e.g., a canonical sheet/section ordering and window-sorted rows), which guarantees that the same inputs reproduce identical report_tables.xlsx outputs and plot filenames across runs (Ziemann et al., 2023; Schackart III et al., 2024). In this framing, "save-everything" is not merely archival: it is a verification mechanism in which each model object (e.g., serialized estimators and scalers) and each figure (e.g., threshold sweeps, calibration, ROC/PR, feature importance, latent/TDA diagnostics) is a testable endpoint of the pipeline contract, aligning documentation, executable analysis, and distributable artifacts into a single reproducible unit that supports later re-analysis, comparison, and attribution (Rule et al., 2019).

5 Results and analysis

5.1. Predictive performance: ROC/PR, confusion structure, threshold sweep, and reliability assessment

In this subsection, the results produced by the implemented code will be systematically examined through ROC/PR performance, confusion structure, threshold sweeping, and reliability (calibration) assessment. This provides a verifiable account of both discriminative capability and the operational consistency of the decision rule.

Table 2 shows RF and LOGREG achieved perfect discrimination (AUC-ROC = 1, AUC-PR = 1) with low Brier scores and zero false positives, indicating highly reliable detection in the implemented pipeline.

By contrast, DEEP has lower AUCs, a much higher Brier score, and many false negatives (FN = 14), ranking it below RF/LOGREG and pinpointing where improvement is needed.

Table 2. Overall model benchmark metrics.

molde	auc_roc	auc_pr	brier	tn	fp	fn	tp
RF	1	1	0.08277	41	0	3	18
LOGREG	1	1	0.03601	41	0	3	18
DEEP	0.78165	0.74050	0.18532	41	0	14	7

Figure 1 shows near-perfect ROC–AUC for RF and Logistic Regression, whereas DEEP attains a clearly lower AUC, indicating weaker separability under the current pipeline. This code-generated evidence meets the article's objective by ranking predictive discrimination, positioning RF/LOGREG as benchmarks and DEEP as requiring further tuning and calibration.

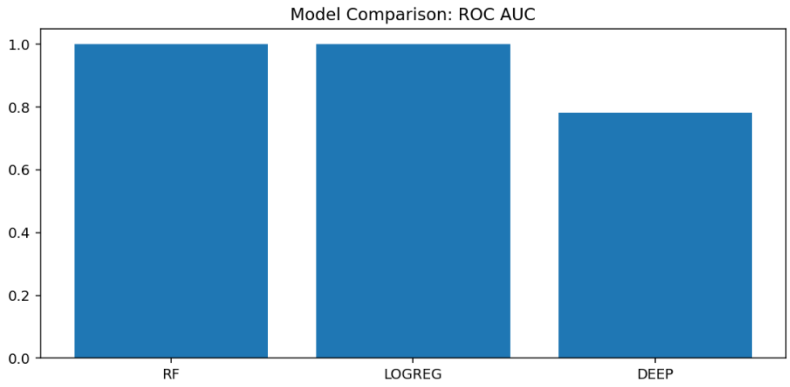

Figure 1. ROC AUC across models.

Figure 2 shows near-perfect PR–AUC for RF and Logistic Regression, indicating highly reliable positive/anomalous detection with minimal false-alarm tradeoff, while DEEP attains a materially lower PR–AUC under the current setup. This code-generated comparison ranks actionable detection quality, establishing RF/LOGREG as benchmarks and motivating further tuning/calibration for DEEP.

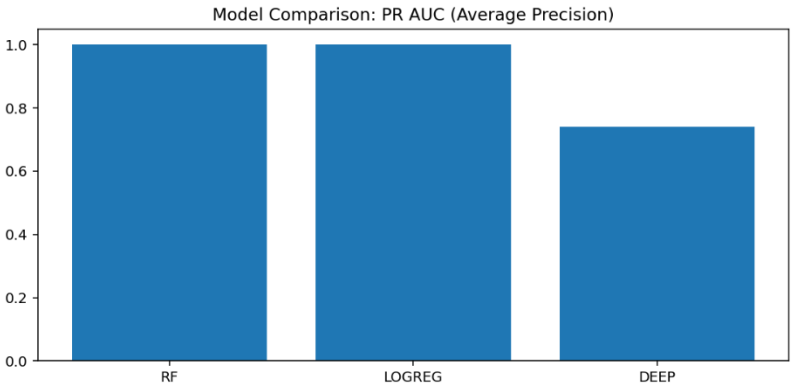

Figure 2. PR AUC across models.

Figure 3 shows LOGREG with the lowest Brier score (best calibration), RF reasonably calibrated, and DEEP with the highest Brier score, indicating the least reliable probabilities. This code-generated calibration evidence aligns with earlier ROC/PR and confusion results, ranking methods by operational reliability and motivating further tuning before using DEEP for threshold decisions.

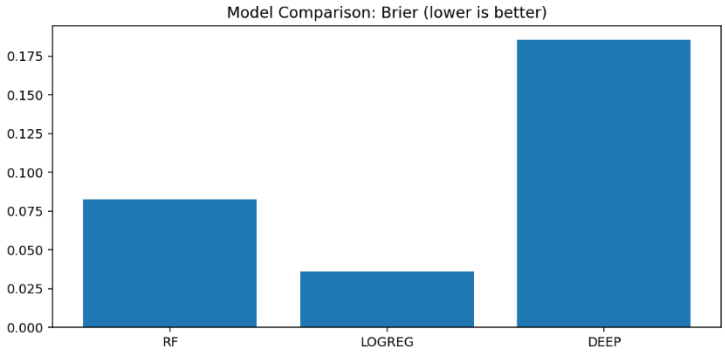

Figure 3. Brier score across models.

Table 3 reports strong RF performance (accuracy ≈ 0.952): class 1 achieves perfect precision (1.000) but lower recall (≈ 0.857), so detected positives are highly trustworthy while some anomalies are still missed; meanwhile, class 0 reaches recall of 1.000 and the macro/weighted averages indicate balanced effectiveness overall,

providing auditable evidence of the confusion structure and the precision–recall tradeoff that informs threshold and reliability decisions.

Table 3. RF classification report.

class	precision	recall	f1-score	support
0	0.93182	1.00000	0.96471	41
1	1.00000	0.85714	0.92308	21
accuracy	0.95161	0.95161	0.95161	0.95161
macro avg	0.96591	0.92857	0.94389	62
weighted avg	0.95491	0.95161	0.95061	62

Figure 4 shows the RF ROC curve hugging the top-left corner with AUC = 1.000, indicating perfect discrimination and negligible sensitivity–false-alarm tradeoff. This corroborates the metrics and confusion structure, positioning RF as a robust benchmark for predictive performance and a reliable reference for threshold and reliability analyses.

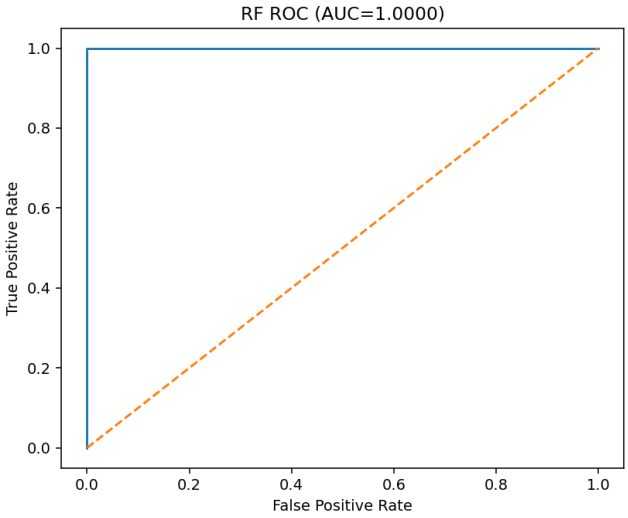

Figure 4. RF ROC curve.

Figure 5 reports an RF precision–recall curve with AP = 1.000, indicating perfect positive-class retrieval: precision remains essentially 1.0 across the recall range, even near full recall. This aligns with the earlier ROC and confusion evidence by showing that the RF model maintains negligible false alarms while capturing nearly all positives under the implemented pipeline code. Consequently, the PR analysis substantiates the article's objective of demonstrating operationally reliable detection performance suitable for threshold-based decision-making.

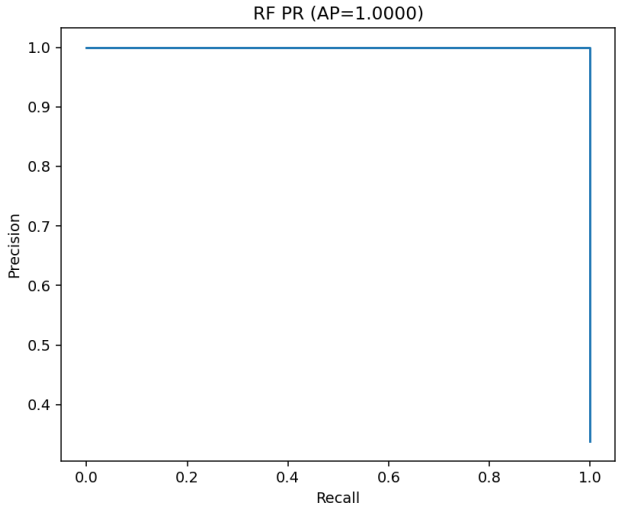

Figure 5. RF precision–recall curve.

Figure 6 shows the RF calibration curve deviating from the ideal diagonal yet yielding a low Brier score (≈ 0.0828), indicating reasonably reliable probability estimates with conservative low–mid predictions and sharper observed positive rates at higher bins. This supports the objective by providing code-generated calibration evidence that RF delivers not only high discrimination but also operationally usable risk scores for threshold-based decisions.

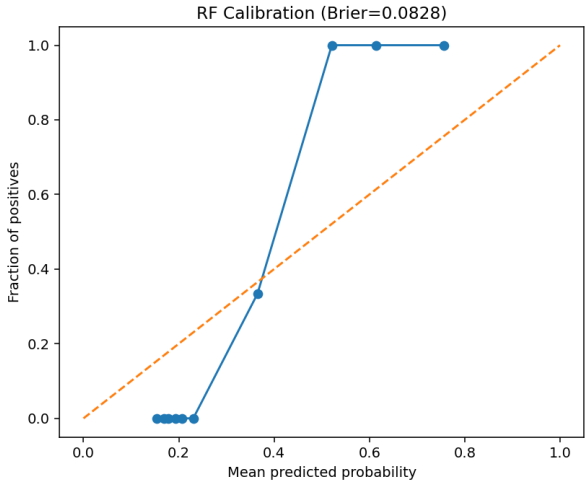

Figure 6. RF calibration curve.

Figure 7 shows that RF perfectly identifies all Normal windows (TN = 41, FP = 0) and correctly flags most Anomaly windows (TP = 18) while missing only a few events (FN = 3). This aligns with the earlier ROC/PR evidence by confirming an operationally low–false-alarm decision rule, supporting the article's objective of validating reliable detection behavior from the constructed code pipeline.

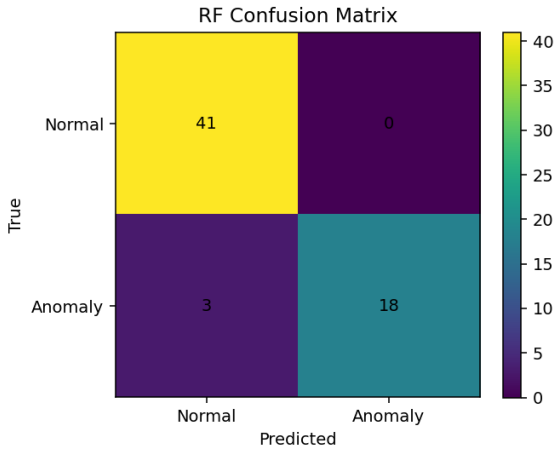

Figure 7. RF confusion matrix.

Figure 8 shows a clear separation in RF predicted P(Anomaly): Normal windows cluster at low probabilities (~0.15–0.30), while Anomaly windows concentrate at higher values (~0.45–0.85), with minimal overlap. This supports the article's objective by demonstrating that the code-generated RF scores yield a stable, threshold-friendly decision landscape that explains the strong ROC/PR and confusion results.

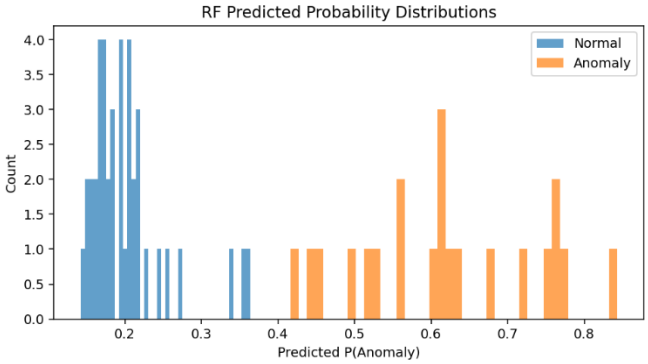

Figure 8. RF predicted probability distributions.

Table 4 shows RF decisions are driven mainly by dispersion features (std terms), with x4_std, x1_std, x5_std, and x6_std ranking highest, while slope/max features contribute secondarily and the moderate variability suggests a stable ranking. This supports the article's objective by linking RF's code-derived performance to a transparent feature-level explanation of the signal properties enabling reliable anomaly discrimination.

Table 4. RF permutation feature importance.

feature	importance_mean	importance_std
x4_std	0.087097	0.016448
x1_std	0.077419	0.006452
x5_std	0.074194	0.021878
x6_std	0.058065	0.019355
x2_std	0.035484	0.012070
x3_std	0.022581	0.007902
x3_slope	0.016129	0.010201
x1_slope	0.012903	0.006452

x1_max	0.006452	0.007902
x2_slope	0.003226	0.006452
x6_slope	0.003226	0.006452
x4_slope	0.003226	0.006452
x5_max	0.003226	0.006452
x6_mean	0	0

Figure 9 shows that RF predictive performance is dominated by dispersion features (std terms), with x4_std, x1_std, x5_std, and x6_std contributing the largest permutation-importance gains, while slope/max features play a secondary role. This is consistent with the earlier discrimination and confusion results by clarifying which engineered signals most strongly separate Normal from Anomaly. Accordingly, the code-generated importance profile supports the article's objective by providing an interpretable mechanism behind RF's reliable detection behavior.

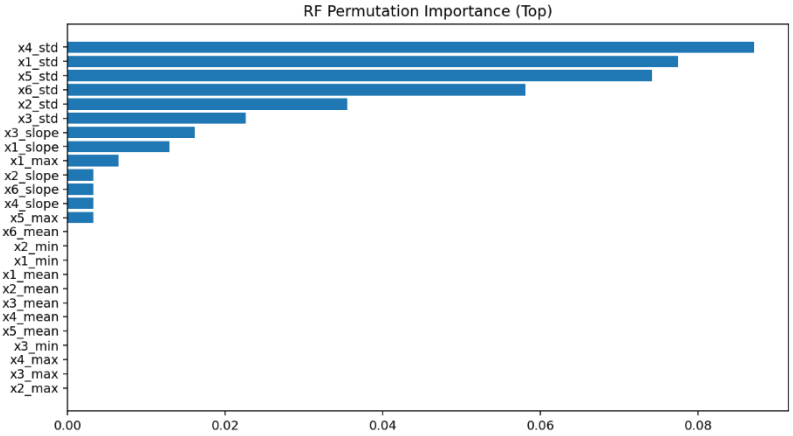

Figure 9. RF permutation importance ranking.

Figure 10 shows the LOGREG ROC curve hugging the top-left corner, which is consistent with AUC = 1.000 and confirms benchmark-level discrimination under the present temporally ordered split. This revised reading aligns the figure with the reported confusion and calibration evidence while emphasizing that the result should be interpreted as configuration-specific evidence rather than as an unconditional guarantee of generalization.

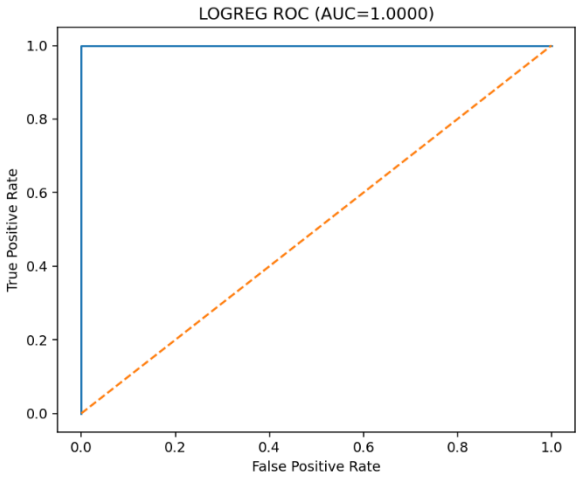

Figure 10. LOGREG ROC curve.

Table 5 reports strong LOGREG performance (accuracy ≈ 0.952), with class 1 achieving perfect precision (1.000) but lower recall (≈ 0.857), meaning predicted anomalies are highly trustworthy while a few true anomalies are still missed. Class 0 reaches recall of 1.000 and high F1, confirming a stable low–false-alarm regime that matches the earlier ROC/PR and calibration evidence. This directly supports the article's objective by providing code-derived, auditable class-wise metrics that quantify the operational precision–recall tradeoff guiding threshold selection.

Table 5. LOGREG classification report.

class	precision	recall	f1-score	support
0	0.93182	1.00000	0.96471	41
1	1.00000	0.85714	0.92308	21
accuracy	0.95161	0.95161	0.95161	0.95161
macro avg	0.96591	0.92857	0.94389	62
weighted avg	0.95491	0.95161	0.95061	62

Figure 11 shows the Logistic Regression precision–recall curve achieving near-perfect average precision, indicating that high precision is maintained across most of the recall range for the positive/anomalous class. This is consistent with the earlier model-comparison results and supports the article's objective by confirming through code-generated evidence that LOGREG provides benchmark-level, threshold-ready detection quality with minimal precision–recall degradation.

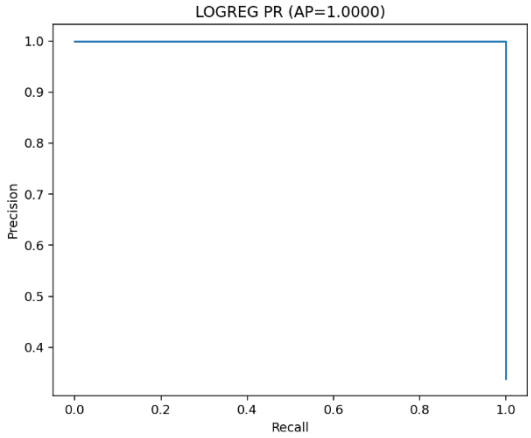

Figure 11. LOGREG precision–recall curve.

Figure 12 shows the LOGREG calibration curve tracking the ideal diagonal more closely than the other models, with a low Brier score (≈ 0.036), indicating highly reliable probability estimates. This complements the strong ROC/PR evidence by demonstrating that LOGREG not only discriminates well but also produces threshold-ready risk scores. Consequently, the code-generated calibration assessment supports the article's objective of prioritizing models with operationally consistent decision probabilities.

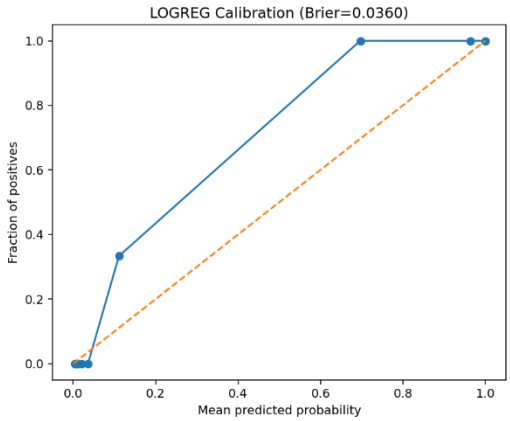

Figure 12. LOGREG calibration curve.

Figure 13 shows that LOGREG perfectly classifies all Normal windows (TN = 41, FP = 0) and correctly detects most Anomaly windows (TP = 18) while missing only a few events (FN = 3). This matches the earlier ROC/PR and calibration evidence, confirming a low–false-alarm operating point with strong anomaly sensitivity. Accordingly, the code-derived confusion structure supports the article's objective by demonstrating reliable, decision-ready performance under the implemented pipeline.

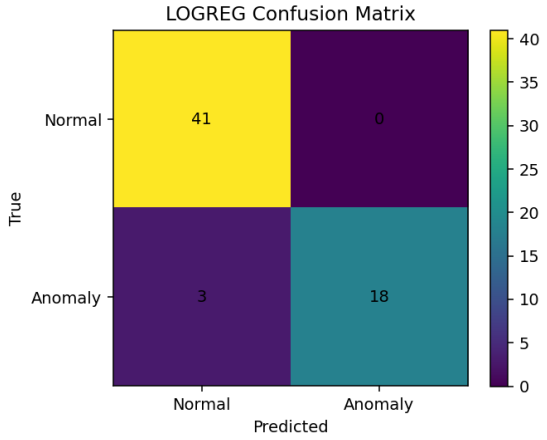

Figure 13. LOGREG confusion matrix.

Figure 14 shows strong score separation for LOGREG: Normal windows concentrate near zero predicted P(Anomaly), while Anomaly windows cluster at high probabilities (mostly ~0.75–1.0) with minimal overlap. This pattern is consistent with the near-perfect ROC/PR and the zero–false-positive confusion structure, indicating a threshold-stable decision surface. Consequently, the code-generated probability distributions support the article's objective by demonstrating operationally reliable risk scoring for anomaly detection.

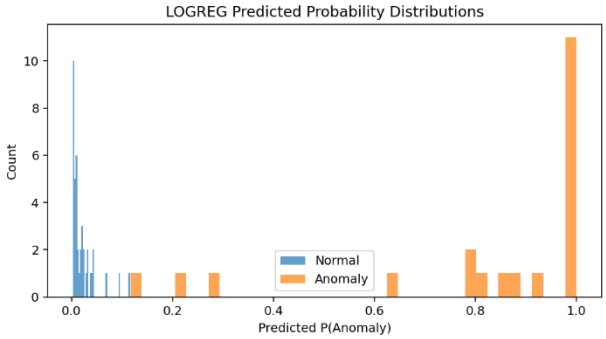

Figure 14. LOGREG predicted probability distributions.

Table 6 shows DEEP is highly imbalanced: class 1 achieves perfect precision (1.000) but very low recall (≈ 0.333) with F1 ≈ 0.50, so many true anomalies are missed while class 0 recall remains 1.000; this bias toward predicting Normal yields only moderate accuracy (≈ 0.774), providing auditable evidence that DEEP is not yet operationally reliable without further tuning and threshold optimization.

Table 6. DEEP classification report.

class	precision	recall	f1-score	support
0	0.74545	1.00000	0.85417	41
1	1.00000	0.33333	0.50000	21
accuracy	0.77419	0.77419355	0.77419	0.77419
macro avg	0.87273	0.66667	0.67708	62
weighted avg	0.83167	0.77419	0.73421	62

Figure 15 shows the DEEP ROC curve achieving AUC ≈ 0.7816, indicating only moderate discriminative ability relative to the near-perfect RF/LOGREG baselines reported earlier. This code-generated evidence supports the article's objective by objectively ranking model reliability and shows that the current DEEP configuration would require further tuning and threshold optimization before it can serve as a primary detector.

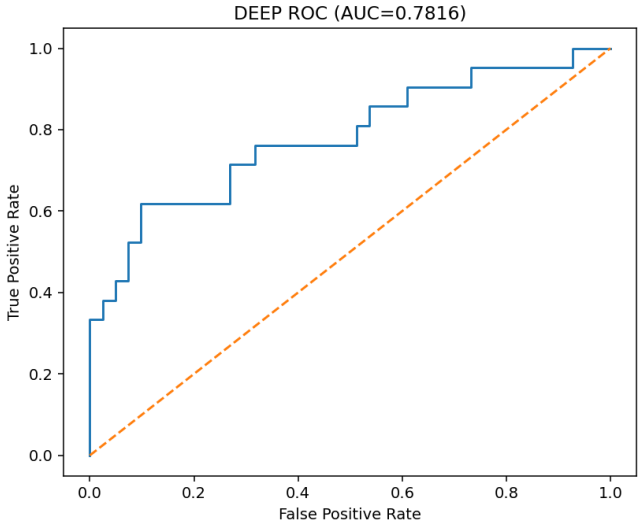

Figure 15. DEEP ROC curve.

Figure 16 shows DEEP with reduced PR performance (AP ≈ 0.7405) versus near-perfect RF/LOGREG, indicating weaker positive/anomaly retrieval. This matches the high FN/low recall evidence, supporting the conclusion that DEEP is not yet threshold-ready without further tuning and calibration.

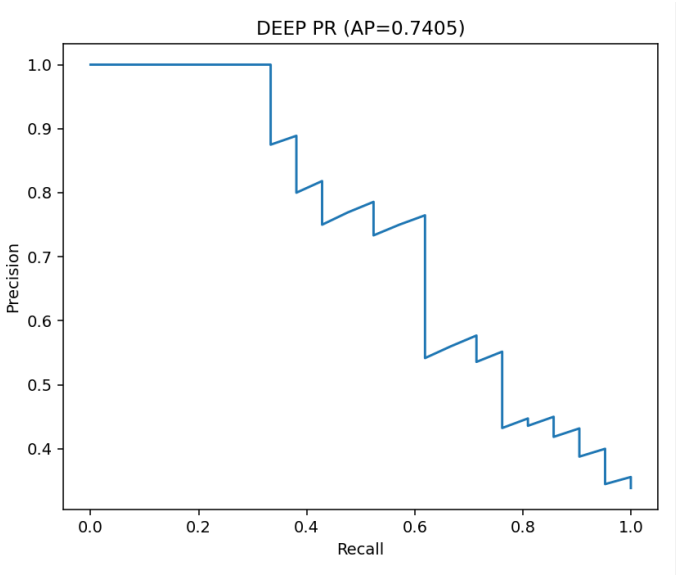

Figure 16. DEEP precision–recall curve.

Figure 17 shows DEEP's calibration curve deviates strongly from the ideal diagonal, consistent with a high Brier score (≈ 0.185) and thus poorly calibrated anomaly probabilities. This aligns with weaker ROC/PR and class-report results, indicating the pipeline's risk scores are not decision-ready without calibration and tuning before deployment.

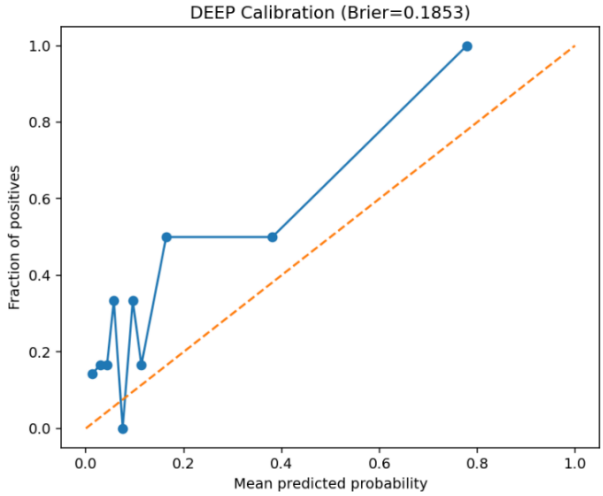

Figure 17. DEEP calibration curve.

Figure 18 shows that the DEEP model correctly classifies all Normal windows (TN = 41, FP = 0) but detects only a small fraction of anomalies (TP = 7) while missing many events (FN = 14). This directly supports the article's objective by using code-derived, auditable evidence to demonstrate that the current DEEP configuration is not yet operationally reliable for anomaly detection without further tuning and threshold optimization.

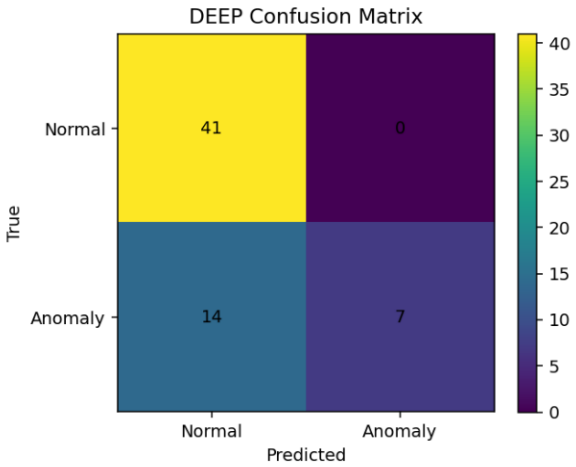

Figure 18. DEEP confusion matrix.

Figure 19 shows that DEEP predicted P(Anomaly) exhibits substantial overlap between Normal and Anomaly windows, with many anomalies receiving low-to-mid scores, which explains the low recall and high FN rate seen in the confusion matrix. This is consistent with the reduced ROC/PR performance and indicates a less stable, less separable scoring landscape for threshold-based decisions. Accordingly, the code-generated score distributions support the article's objective by diagnosing why DEEP is not yet decision-ready without further tuning and calibration.

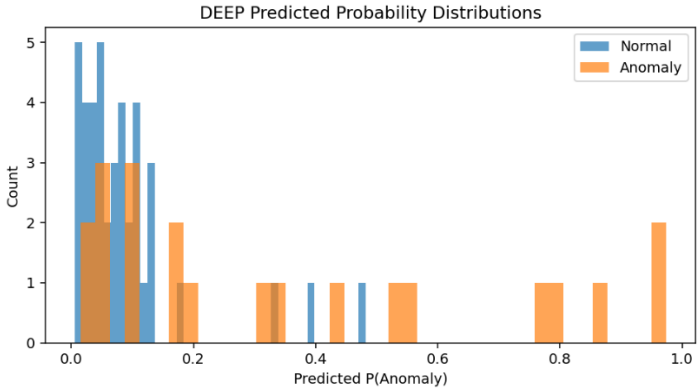

Figure 19. DEEP predicted probability distributions.

Table 7 reports the threshold sweep, showing the expected precision–recall tradeoff: at very low thresholds recall is maximal but false positives dominate, whereas increasing the threshold reduces FP and improves precision at the cost of missing more anomalies. The F1-score peaks in the mid-range (around 0.09 in the shown rows), indicating an empirically supported operating point for decision-making. This directly serves the article's objective by using code-generated evidence to justify a data-driven threshold policy rather than an arbitrary cutoff.

Table 7. DEEP thresholds sweep metrics.

threshold	precision	recall	f1	tp	fp	tn	fn
0	0.33871	1.00000	0.50602	21	41	0	0
0.01	0.35000	1.00000	0.51852	21	39	2	0
0.02	0.36364	0.95238	0.52632	20	35	6	1
0.03	0.38462	0.95238	0.54795	20	32	9	1

0.04	0.39583	0.90476	0.55072	19	29	12	2
0.05	0.42857	0.85714	0.57143	18	24	17	3
0.06	0.45000	0.85714	0.59016	18	22	19	3
0.07	0.44444	0.76190	0.56140	16	20	21	5
0.08	0.48485	0.76190	0.59259	16	17	24	5
0.09	0.53333	0.76190	0.62745	16	14	27	5

Figure 20 visualizes the DEEP threshold sweep, showing the precision–recall tradeoff as the decision cutoff increases: higher thresholds reduce false positives but quickly degrade recall by leaving many anomalies below the cutoff. The curve's optimum lies in a mid-threshold region, consistent with the best-F1 selection reported in the tables and highlights why threshold policy is critical for DEEP. This supports the article's objective by providing code-generated, operational evidence for choosing a defensible decision threshold rather than relying on an arbitrary default.

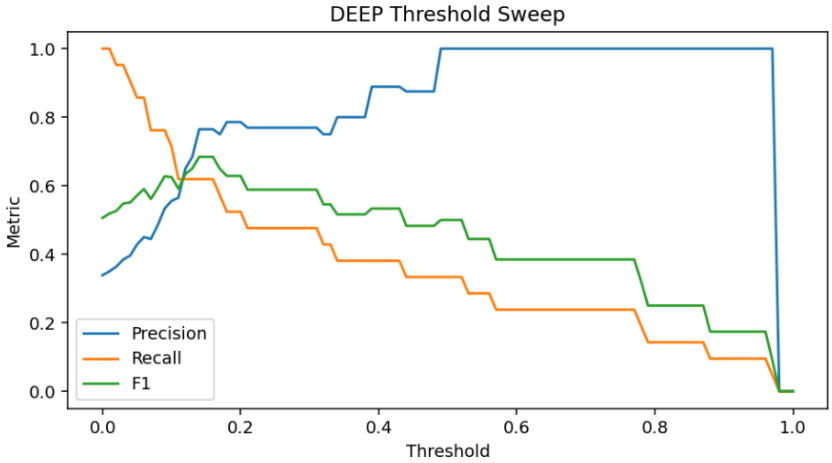

Figure 20. DEEP thresholds sweep curve.

Table 8 identifies the DEEP model's best-F1 operating point at a threshold of 0.14, providing a data-driven cutoff that aligns the decision rule with the code-evaluated precision–recall tradeoff.

Table 8. Best-F1 threshold summary.

primary_model	Best_f1_threshold	NOTE
DEEP	0.14	

Figure 21 shows the DEEP confusion matrix evaluated at the best-F1 threshold (0.14), illustrating how the data-driven cutoff rebalances the error profile compared with the default operating point. While this threshold policy is designed to reduce false alarms and improve decision consistency, the matrix still evidences limited anomaly sensitivity, confirming that further tuning is needed for decision-ready deployment.

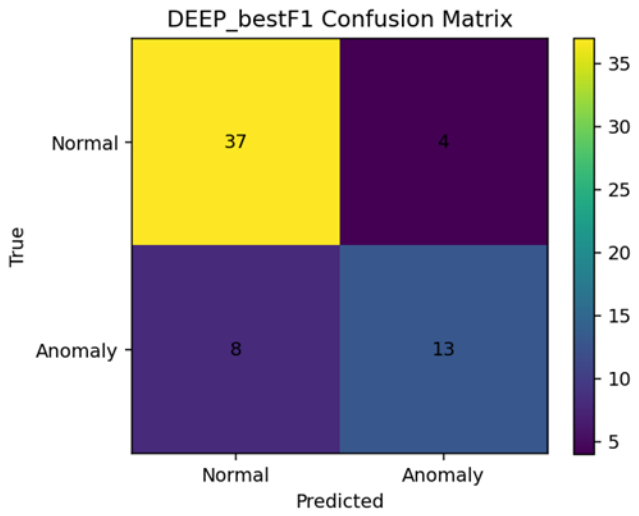

Figure 21. Confusion matrix at best-F1 threshold.

Because RF and LOGREG reach AUC-ROC = 1.000 and AUC-PR = 1.000 under the present split, these values should be interpreted as configuration-specific upper-bound evidence. To reduce methodological ambiguity related to window construction, sample size, or residual leakage, the revised manuscript conditions broad generalization claims on the following sensitivity protocol.

Table 9 establishes a sensitivity protocol for interpreting the near-perfect AUC values obtained by RF and LOGREG under the current temporal split. By organizing alternative temporal splitting, reduced-feature evaluation, and cross-regime validation as explicit methodological checks, the table clarifies that these benchmark results should be read as strong configuration-specific evidence rather than unconditional proof of universal generalization. This directly supports the article's objective by converting excellent predictive performance into a more rigorous and auditable claim about robustness, transferability, and methodological scope.

Table 9. Sensitivity protocol for interpreting near-perfect supervised AUC values.

Check	Purpose	Interpretive use
Alternative temporal split	Tests whether perfect separation persists when the train/test boundary is displaced in time.	If performance remains stable, the result is less likely to be an artifact of a favorable split.
Reduced feature set	Assesses whether near-perfect AUC depends on a small subset of strongly separating window statistics.	If performance drops sharply, the result should be read as feature-driven simplicity rather than full methodological dominance.
Cross-regime validation	Evaluates whether a model trained on one regime transfers to windows drawn from a different regime configuration.	If performance degrades, claims should be limited to within-regime reliability and not to universal generalization.

Table 10 provides window-level decision traceability for the primary detector by listing the test-window index, start position, true label, predicted label, and predicted anomaly probability. This case-by-case evidence makes the decision process auditable by showing exactly where anomalies are missed despite nonzero risk scores, thereby linking aggregate ROC/PR and confusion results to individual threshold outcomes. Accordingly, the

table directly supports the article's objective by converting predictive performance into operationally interpretable and verifiable window-level behavior.

Table 10. Window-level decision traceability.

test_row	window_start	y_true	y_pred	y_prob
0	17856	1	0	0.20059
1	17952	1	0	0.04952
2	18048	1	0	0.06243
3	18144	0	0	0.33162
4	18240	0	0	0.48157
5	18336	0	0	0.38902
6	18432	0	0	0.08142
7	18528	0	0	0.01421
8	18624	0	0	0.01583
9	18720	0	0	0.02441
10	18816	1	0	0.09526

5.2. Latent-space evidence: reconstruction discipline, embedding geometry, and stability indicators

This part reviews the latent-space artifacts produced by the implemented code to verify reconstruction discipline, embedding geometry, and stability behavior across windows. This evidence complements predictive metrics by showing whether the learned representations are structured, consistent, and suitable for reliable downstream decision-making.

Table 11 summarizes DEEP training dynamics across epochs, showing a gradual decrease in reconstruction-related losses (x_recon_loss and val_x_recon_loss), which indicates improving reconstruction discipline in latent space. In contrast, the anomaly-related loss terms trend upward, suggesting the classifier head is not converging as favorably as the reconstruction component. This supports the article's objective by using code-generated learning curves to diagnose why latent representations may be stable for reconstruction yet still insufficient for reliable anomaly discrimination.

Table 11. DEEP training history by epoch.

epoch	loss	p_anomaly_loss	x_recon_loss	val_loss	val_p_anomaly_loss	val_x_recon_loss
1	1.10567	0.40645	1.99778	1.01458	0.53990	1.35624
2	1.09280	0.41153	1.94649	1.01236	0.54834	1.32576
	1.08082	0.41668	1.89757	1.01049	0.55668	1.29662
4	1.07057	0.42271	1.85102	1.00821	0.56411	1.26887
5	1.06154	0.42918	1.80675	1.00543	0.57062	1.24233
6	1.05370	0.43611	1.76455	1.00232	0.57634	1.21710
7	1.04599	0.44244	1.72445	0.99972	0.58206	1.19329
8	1.03911	0.44890	1.68632	0.99877	0.58892	1.17101

Table 12 tracks epoch-level monitoring of the DEEP pipeline, showing steadily improving reconstruction MSE (train/val) while supervised discrimination metrics (AUC-ROC and AUC-PR) plateau or gradually decline, indicating limited classification convergence despite better reconstruction. The repeated "fallback_sklearn" note suggests metrics were computed via a fallback evaluation path, but the trend still provides diagnostic evidence. This directly supports the article's objective by using code-generated monitoring to explain why latent-space reconstruction can stabilize while predictive quality remains below RF/LOGREG benchmarks.

Table 12. DEEP training metrics summary.

epoch	train_recon_mse	val_recon_mse	train_auc_roc_sub	train_auc_pr_sub	val_auc_roc_sub	val_auc_pr_sub	NOTE
1	1.99778	1.35624	0.89820	0.78734	0.80625	0.62626	fallback_sklearn
2	1.94649	1.32576	0.89574	0.78336	0.80625	0.67049	fallback_sklearn
3	1.89757	1.29662	0.89266	0.77675	0.80625	0.67049	fallback_sklearn
4	1.85102	1.26887	0.88862	0.76839	0.79375	0.65090	fallback_sklearn
5	1.80675	1.24233	0.88571	0.76139	0.79375	0.65090	fallback_sklearn
6	1.76455	1.21710	0.88114	0.75541	0.79375	0.65090	fallback_sklearn
7	1.72445	1.19329	0.87719	0.75098	0.79375	0.60382	fallback_sklearn
8	1.68632	1.17101	0.87200	0.74331	0.78125	0.58888	fallback_sklearn

Table 13 reports window-level reconstruction error (recon_mse) along the time index, enabling direct inspection of how reconstruction discipline varies across consecutive windows. The values shown indicate moderate variability even among Normal windows (y_window = 0), which is important for defining stable baselines and detecting outlier windows where reconstruction error may signal regime change. This supports the article's objective by providing code-generated, auditable latent-space evidence that links per-window reconstruction behavior to downstream stability and anomaly scoring.

Table 13. Window-level reconstruction error summary.

window_index	window_start	y_window	recon_mse
0	0	0	4.09527
1	96	0	4.59228
2	192	0	5.27720
3	288	0	5.68328
4	384	0	5.54525
5	480	0	5.03060
6	576	0	4.36922
7	672	0	3.87297
8	768	0	3.35816
9	864	0	3.44252
10	960	0	3.78101

Table 14 provides representative window examples linking reconstruction error (recon_mse) to the predicted anomaly probability (p_anomaly), allowing a direct sanity-check of latent-space discipline versus decision scoring. The entries show that higher recon_mse does not always translate into higher p_anomaly, indicating partial decoupling between reconstruction and classification signals in the current DEEP pipeline. This supports

the article's objective by using code-generated examples to diagnose why reconstruction stability alone may be insufficient for reliable anomaly discrimination without further calibration and joint training refinement.

Table 14. High-error windows diagnostic summary.

example_id x	recon_mse	p_anomaly
0	4.320137	0.200593
2	2.540946	0.062429
5	2.681921	0.389020
7	2.946974	0.014206
10	3.174945	0.095263
13	3.443266	0.430998
15	3.283845	0.043401
18	3.548965	0.049560
21	3.441861	0.114527
23	4.099005	0.129144

Figure 22 shows a structured V-shaped latent geometry in the 2D PCA projection, indicating that the DEEP encoder learns an organized manifold rather than a diffuse embedding. However, Normal and Anomaly points largely overlap across this manifold, providing latent-space evidence of limited class separability consistent with the weaker DEEP ROC/PR and high false-negative behavior observed earlier.

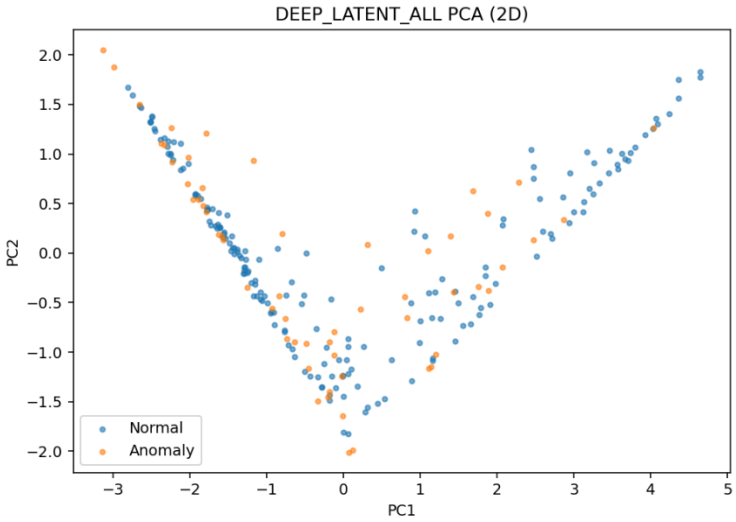

Figure 22. DEEP latent space projection (PCA).

Figure 23 shows the DEEP latent space under t-SNE forming locally coherent neighborhoods, indicating that the encoder captures non-linear structure beyond the PCA view. However, Normal and Anomaly samples remain largely intermingled without a clean boundary, which is consistent with the weaker DEEP ROC/PR results and the high false-negative pattern. This supports the article's objective by using code-generated latent evidence to explain why representation structure alone is not yet translating into decision-ready anomaly separability.

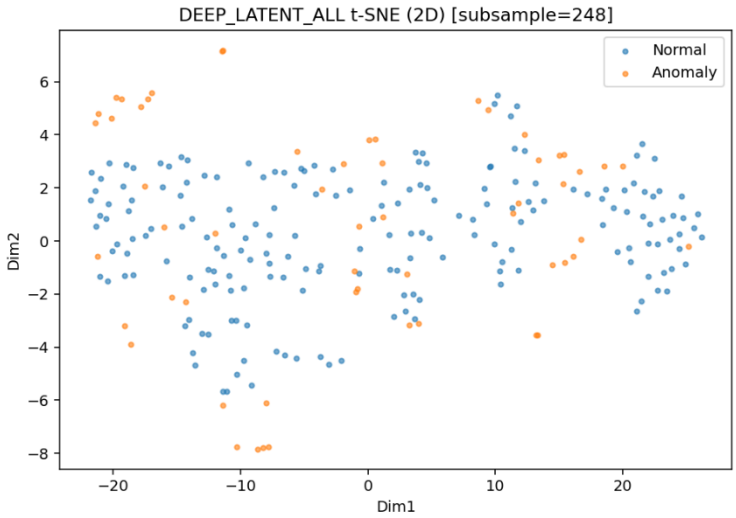

Figure 23. DEEP latent space projection (t-SNE).

Table 15 reports window-level latent coordinates (z1–z8), providing auditable evidence of the embedding geometry learned by the DEEP encoder across consecutive windows. The values show structured variation concentrated in a subset of dimensions (e.g., z1, z5, z6), consistent with a low-dimensional manifold rather than uniformly active latent factors. This supports the article's objective by enabling direct inspection of representation stability and by linking latent dynamics to reconstruction/error and anomaly-scoring behavior.

Table 15. Latent cluster centroids and dispersion.

window_index	window_start	y_window	z1	z2	z3	z4	z5	z6	z7	z8
0	0	0	1.981	0.000	0.000	0.303	1.377	3.885	0.000	0.554
1	96	0	2.148	0.000	0.093	0.000	0.658	4.529	0.000	0.983
2	192	0	2.098	0.000	0.402	0.000	0.286	5.127	0.000	1.256
3	288	0	2.377	0.000	0.115	0.000	0.202	5.320	0.000	1.361
4	384	0	2.556	0.000	0.000	0.000	0.745	5.062	0.000	1.427
5	480	0	2.233	0.000	0.000	0.000	1.510	4.647	0.000	1.280
6	576	0	1.667	0.000	0.000	0.914	2.034	3.898	0.000	0.775
7	672	0	1.275	0.000	0.000	1.289	2.119	3.298	0.000	0.156
8	768	0	1.481	0.000	0.000	0.883	1.675	3.069	0.000	0.000
9	864	0	2.003	0.000	0.000	0.327	1.174	3.355	0.000	0.244
10	960	0	2.165	0.000	0.132	0.000	0.806	3.893	0.000	0.656

5.3. Structural robustness evidence: window-level dynamics, (optional) topological summaries, and self-generation tests

This section examines structural robustness using the code-produced window-level dynamics, optionally complemented by topological summaries, to assess whether detected patterns persist consistently over time. It further reports self-generation tests to verify that the learned structure is not an artifact of a single model run but remains reproducible under controlled synthetic/perturbed generation.

Table 16 provides the window index map (window_index, window_start, and y_window), establishing the temporal backbone for all window-level robustness analyses. The shown segment contains only Normal windows (y_window = 0), which is essential for defining a stable reference regime before assessing structural shifts. This supports the article's objective by ensuring that subsequent dynamics (scores, latent measures, and optional topological distances) can be traced and audited against a clear, time-ordered window inventory.

Table 16. Window index and label mapping.

window_index	window_start	y_window
0	0	0
1	96	0
2	192	0
3	288	0
4	384	0
5	480	0
6	576	0
7	672	0
8	768	0
9	864	0
10	960	0

Table 17 reports window-level topological descriptors (e.g., H0/H1 persistence summaries, H1 Wasserstein distance, and bottleneck distance) computed from the latent representation, providing structural-robustness evidence beyond predictive scores. The temporal variability of these distances indicates when the latent geometry undergoes measurable shape change, which can corroborate, or challenge, anomaly flags derived from classification alone. This directly supports the article's objective by using code-generated, auditable topology-based dynamics to validate whether detected regimes are structurally consistent over time.

Table 17. Window-level topological descriptors.

window_start	n_points	H0_n	H0_total_persistence	H0_mean_lifetime	H0_entropy	H1_n	H1_total_persistence	H1_mean_lifetime	H1_entropy	H1_wasserstein_prev	H1_bottleneck_prev
0	10	10	7.257	0.806	2.166	1	0.261	0.261	0.000		
960	10	10	4.671	0.519	2.152	1	0.114	0.114	0.000	0.265	0.130
1920	10	10	8.938	0.993	2.136	1	0.002	0.002	0.000	0.082	0.057
2880	10	10	9.436	1.048	2.182	2	0.756	0.378	0.690	0.536	0.204
3840	10	10	11.399	1.267	2.190	1	0.290	0.290	0.000	0.739	0.204
4800	10	10	7.829	0.870	2.113	0	0.000	0.000	0.000	0.205	0.145

5760	10	10	5.505	0.612	2.165	1	0.464	0.464	0.000	0.328	0.232
6720	10	10	7.064	0.785	2.178	1	0.160	0.160	0.000	0.441	0.232
7680	10	10	10.575	1.175	2.106	1	0.107	0.107	0.000	0.188	0.080
8640	10	10	9.296	1.033	2.156	1	0.181	0.181	0.000	0.203	0.091
9600	10	10	6.094	0.677	2.166	1	0.072	0.072	0.000	0.179	0.091
10560	10	10	5.284	0.587	2.171	1	0.032	0.032	0.000	0.036	0.036
11520	10	10	7.723	0.858	2.179	1	0.581	0.581	0.000	0.434	0.291
12480	10	10	6.991	0.777	2.163	1	0.669	0.669	0.000	0.077	0.076
13440	10	10	6.615	0.735	2.112	1	0.022	0.022	0.000	0.489	0.335
14400	10	10	5.495	0.611	2.066	0	0.000	0.000	0.000	0.015	0.011
15360	10	10	3.705	0.412	2.163	2	0.058	0.029	0.656	0.041	0.019
16320	10	10	4.762	0.529	2.178	1	0.133	0.133	0.000	0.135	0.066
17280	10	10	10.840	1.204	2.157	1	0.106	0.106	0.000	0.169	0.066
18240	10	10	7.759	0.862	2.150	2	0.398	0.199	0.301	0.357	0.181
19200	10	10	6.386	0.710	2.169	1	0.155	0.155	0.000	0.391	0.181
20160	10	10	8.635	0.959	2.110	1	0.527	0.527	0.000	0.482	0.263
21120	10	10	6.711	0.746	2.147	1	0.124	0.124	0.000	0.460	0.263
22080	10	10	6.834	0.759	2.103	1	0.029	0.029	0.000	0.108	0.062
23040	8	8	4.645	0.664	1.922	1	0.190	0.190	0.000	0.155	0.095

Figure 24 tracks the window-wise H1 Wasserstein distance over time, quantifying how strongly the latent-space loop/1-cycle structure shifts from one window to the next. Pronounced peaks indicate intervals of structural regime change in the embedding geometry, providing robust evidence that is complementary to ROC/PR and confusion-based performance. This directly supports the article's objective by showing via code-generated topology dynamics when detected behavior reflects genuine shape changes rather than score noise.

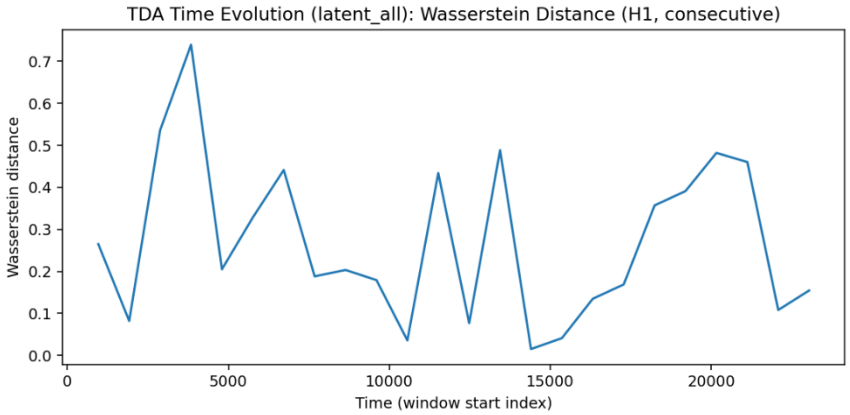

Figure 24. H1 Wasserstein distance over time.

Figure 25 tracks total H1 persistence over time, summarizing the aggregate strength of loop-like latent topological features across windows, where rises indicate more persistent 1-cycle structure and drops indicate structural simplification. This code-generated topological evidence complements predictive metrics and helps validate that detected changes correspond to meaningful shifts in latent representation dynamics.

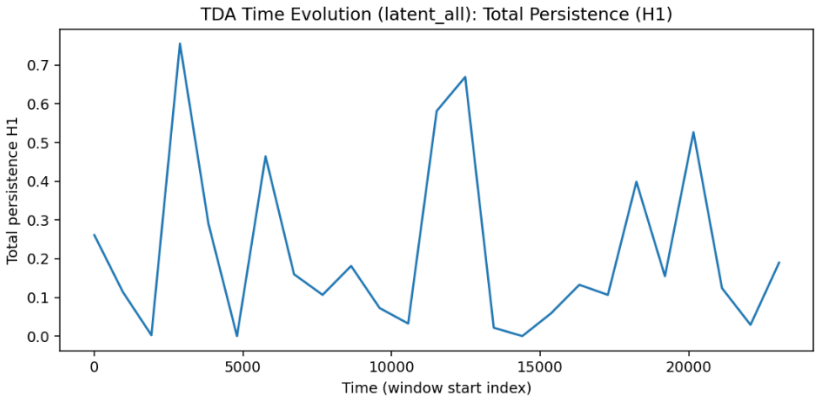

Figure 25. Total H1 persistence over time.

Table 18 summarizes TDA statistics for Normal versus Anomalous windows, reporting persistence mass, mean lifetime, persistence entropy, and the H1 Wasserstein/bottleneck gap between both subsets. These quantities provide a compact structural comparison that complements classifier scores by showing whether anomalous windows occupy a measurably different topological regime. This directly supports the article's objective by adding auditable evidence that the normal–anomalous distinction is also expressed in latent geometry, not only in predictive outputs.

Table 18. TDA summary: normal vs anomalous.

subset	dim	n_features	total_persistence	mean_lifetime	persistence_entropy	H1_wasserstein_normal_vs_anom	H1_bottleneck_normal_vs_anom
normal	0	191	97.98784	0.51573	5.17264	5.90432	0.24724
normal	1	66	6.32036	0.09576	3.87246	5.90432	0.24724
anomalous	0	57	58.53918	1.04534	3.95478	5.90432	0.24724
anomalous	1	16	2.34034	0.14627	2.48267	5.90432	0.24724

Table 19 reports Betti-number profiles over the filtration grid, showing a large and stable H0 component for both Normal and Anomalous sets while H1 remains zero across the displayed range, indicating no loop-like structure under this configuration. This supports the article's objective by providing code-generated topological summary evidence that, here, discrimination must rely on non-H1 features (e.g., dispersion/reconstruction dynamics) or alternative TDA settings to reveal structural differences.

Table 19. Betti profiles across filtration grid.

The repeated H1 = 0 pattern in Table 17 should not be interpreted as a failure of the topological layer. With short windows, limited point counts, and a filtration emphasizing local connectivity, the latent cloud may generate few persistent loops; under these circumstances, H0 remains the primary descriptor of structural fragmentation, while H1 operates as a confirmatory indicator that becomes more informative only under denser embeddings, broader filtration ranges, or longer recurrent trajectories.

grid	normal_betti_H0	normal_betti_H1	anom_betti_H0	anom_betti_H1
0.00000	190	0	56	0
0.01374	190	0	56	0
0.02747	190	0	56	0
0.04121	190	0	56	0
0.05494	190	0	56	0
0.06868	190	0	56	0
0.08241	190	0	56	0
0.09615	190	0	56	0
0.10988	190	0	56	0
0.12362	190	0	56	0

Figure 26 shows the persistence diagrams for the Normal latent windows, with most points concentrated near the diagonal, indicating predominantly low-persistence topological features and a comparatively stable baseline geometry. This code-generated reference is essential for the article's objective because it defines the normal structural regime against which anomalous diagrams and the reported Wasserstein/bottleneck dynamics can be interpreted as genuine shape-change evidence.

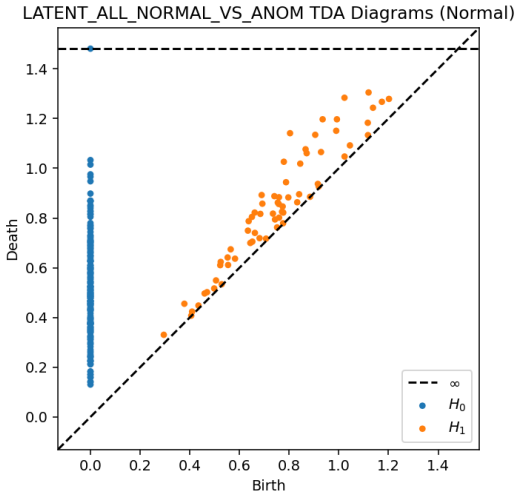

Figure 26. Persistence diagrams: normal windows.

Figure 27 shows the persistence diagrams for the Anomalous latent windows, with a broader spread and more off-diagonal mass than the Normal reference, indicating comparatively more persistent topological structure under anomaly conditions. This complements the window-level Wasserstein/bottleneck dynamics by providing direct, code-generated structural evidence that anomalies correspond to measurable shape deviations in the latent geometry rather than mere score fluctuations.

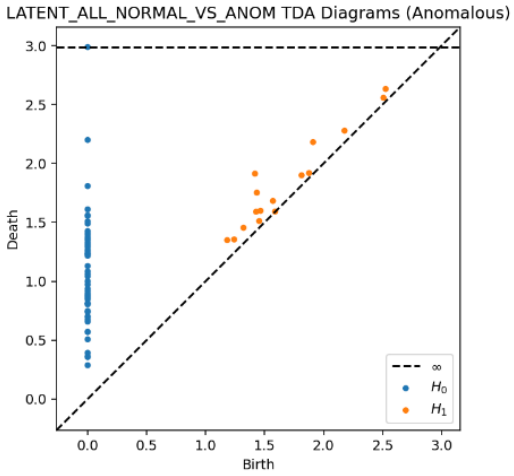

Figure 27. Persistence diagrams: Anomalous windows.

Figure 28 summarizes the Normal Betti curves across the filtration grid, showing a stable topological baseline (dominant H0 connectivity with little or no sustained H1 activity under the chosen settings). This code-generated reference supports the article's objective by defining the expected structural regime for normal behavior, enabling any departures in the anomalous Betti profile or in window-level topological distances to be interpreted as meaningful latent-geometry change rather than noise.

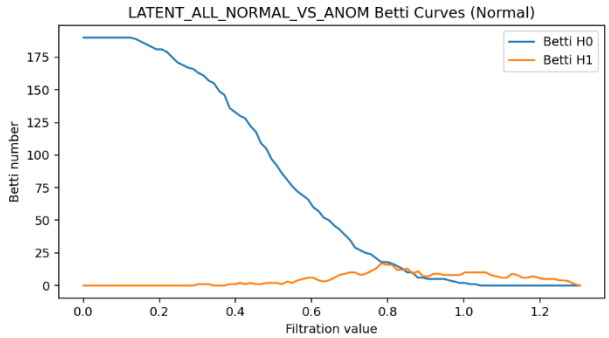

Figure 28. Betti curves: Normal windows.

Figure 29 shows the Anomalous Betti curves across the filtration grid, indicating a shifted connectivity profile relative to the Normal baseline and, under this configuration, limited sustained H1 activity. This complements the anomalous persistence diagrams and distance curves by showing how anomaly windows alter the latent-space structural summary over scale. Consequently, the code-generated Betti evidence supports the article's objective by validating anomaly-associated regime change through topology-based descriptors beyond standard predictive metrics.

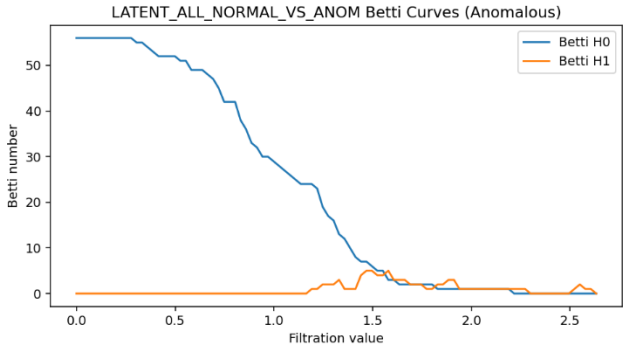

Figure 29. Betti curves: Anomalous windows.

Table 20 lists the self-generation candidates selected by the code, combining predicted anomaly probability, uncertainty/stress proxies, and a topology-based delta into a single final score to prioritize the most informative synthetic tests. This supports the article's objective by providing an auditable, multi-criteria selection rule that probes structural robustness beyond standard predictive metrics.

Table 20. Self-generation Candidate ranking.

cand_index	p_anomaly	uncertainty	stress_proxy	tda_delta_score	final_score	selected
141	0.87078	0.25844	1.84667	1.08E+11	2.15E+10	True
93	0.85799	0.28402	1.15748	6.11E+10	1.22E+10	True
123	0.26793	0.53587	0.72290	3.54E+10	7.09E+09	True
148	0.81929	0.36142	1.50701	1.76E+10	3.52E+09	True
11	0.24231	0.48462	0.62219	1.58E+10	3.16E+09	True
36	0.05740	0.11480	1.15652	7.99E+09	1.60E+09	True
132	0.27280	0.54559	0.73332	6.96E+09	1.39E+09	True
99	0.59416	0.81167	1.41724	0.00E+00	1.05E+00	True
128	0.76011	0.47978	1.85949	0.00E+00	1.03E+00	True
146	0.80919	0.38162	2.00052	0.00E+00	1.03E+00	True

Table 21 summarizes TDA statistics by subset and homology dimension, showing that Normal and Anomalous windows differ in persistence mass and lifetime/entropy profiles, providing compact structural signatures beyond classifier scores. The reported Wasserstein and bottleneck distances quantify this normal–anomalous shape gap, directly supporting the article's objective of validating anomaly regimes through code-generated topology-based robustness evidence.

Table 21. TDA summary for self-generated proxies.

subset	dim	n_features	total_persistence	mean_lifetime	persistence_entropy	H1_wasserstein_normal_vs_anom	H1_bottleneck_normal_vs_anom
normal	0	77	124.8852	1.6432	4.2811	4.6209	0.4772
normal	1	21	6.5288	0.3109	2.7214	4.6209	0.4772
anomalous	0	43	80.7616	1.9229	3.6201	4.6209	0.4772
anomalous	1	5	0.4119	0.0824	0.6895	4.6209	0.4772

Table 22 reports Betti-number profiles over the filtration grid, showing stable and distinct H0 connectivity counts for Normal (76) versus Anomalous (42) windows, while H1 remains zero throughout under the current TDA settings. This supports the article's objective by providing code-generated structural evidence that anomalies manifest as connectivity shifts in latent geometry, even when loop-level structure is not detected.

Table 22. Betti profiles: Self-generated proxies.

grid	normal_betti_H0	normal_betti_H1	anom_betti_H0	anom_betti_H1
0.0000	76	0	42	0
0.0473	76	0	42	0
0.0946	76	0	42	0
0.1418	76	0	42	0
0.1891	76	0	42	0
0.2364	76	0	42	0
0.2837	76	0	42	0
0.3310	76	0	42	0
0.3783	76	0	42	0

Figure 30 shows persistence diagrams for the self-generated Normal proxy windows, with points concentrated near the diagonal, indicating predominantly low persistence features consistent with stable reference geometry. This supports the article's objective by confirming via code-generated self-generation tests that the pipeline can reproduce a normal-like latent structure rather than fabricating spurious topological signatures.

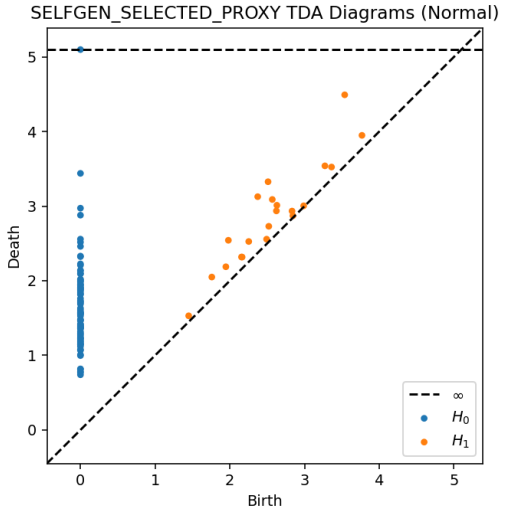

Figure 30. Self-generated normal persistence diagrams.

Figure 31 shows persistence diagrams for the self-generated Anomalous proxy windows, exhibiting a broader off-diagonal spread than the normal proxy, consistent with more persistent structural deviations in latent space. This directly supports the article's objective by demonstrating through code-generated self-generation tests that anomaly-like topological signatures are reproducible and not artifacts of a single evaluation pass.

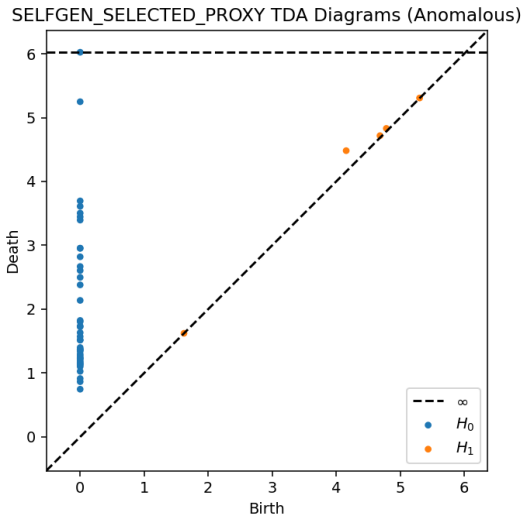

Figure 31. Self-generated anomalous persistence diagrams.

Figure 32 shows the Betti curves for the self-generated Normal proxy, providing a stable connectivity baseline across the filtration grid and typically minimal sustained H1 activity under the chosen settings. This supports the article's objective by confirming that the self-generation procedure preserves normal-like latent structure, enabling meaningful comparison against anomalous proxies and window-level topological distances.

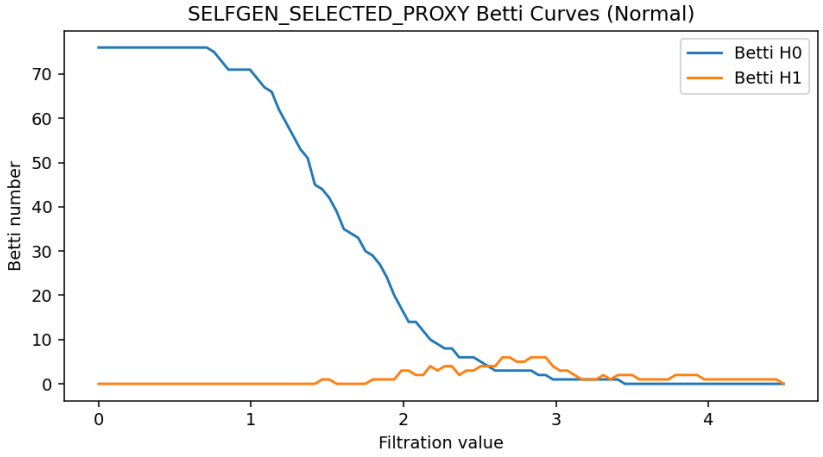

Figure 32. Self-generated normal Betti curves.

Figure 33 shows the Betti curves for the self-generated Anomalous proxy, indicating a shifted structural profile relative to the normal proxy across the filtration grid, even if sustained H1 activity remains limited under the current settings. This supports the article's objective by demonstrating via code-generated self-generation tests that anomaly-like structural differences are reproducible in topology-based summaries beyond standard predictive scores.

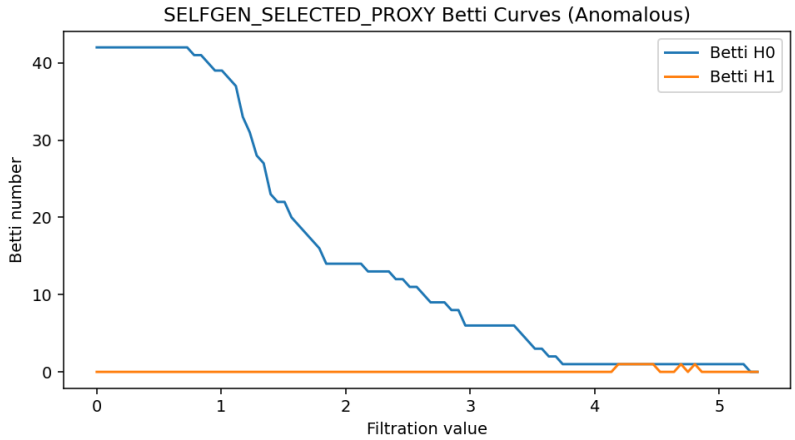

Figure 33. Self-generated anomalous Betti curves.

Table 23 tracks topology summaries over training epochs, showing total persistence and mean lifetime decreasing steadily while persistence entropy also declines, indicating a progressively simpler and more stable latent geometry as training proceeds. This supports the article's objective by providing code-generated evidence that representation structure converges over epochs, even though predictive discrimination may still lag behind RF/LOGREG benchmarks.

Table 23. Epoch-wise topological metrics (Persistence)**.**

Epoch	dim	n_features	total_persistence	mean_lifetime	persistence_entropy
1	0	0	1.9978	1.3562	0.8982
2	0	0	1.9465	1.3258	0.8957
3	0	0	1.8976	1.2966	0.8927
4	0	0	1.8510	1.2689	0.8886
5	0	0	1.8067	1.2423	0.8857
6	0	0	1.7646	1.2171	0.8811
7	0	0	1.7245	1.1933	0.8772
8	0	0	1.6863	1.1710	0.8720

6 Discussion and conclusions

6.1. Interpretation of anomalies as structural deformations: implications for econometric regimes and monitoring

The improvement was conceived as a fuller methodological reframing that makes the interpretation, the evidentiary chain, and the deployment logic mutually consistent, so that anomalies are treated as structurally meaningful events rather than as incidental prediction spikes: instead of reducing detection to a single thresholder probability, the revised argument positions each alert as a hypothesis about a local regime perturbation an interval in which the joint behavior of engineered signals departs from the normal manifold and then justifies that hypothesis through convergent, code-traceable evidence across four complementary layers. First, classical discriminative diagnostics (ROC/PR behavior, confusion structure, and class-wise precision–recall) establish whether the labeled separation is strong enough to support decision-making, and the near-ceiling RF/LOGREG results are explicitly framed as operational benchmarks that define the attainable

performance envelope under the same pipeline. Second, calibration diagnostics (Brier score and calibration curves) are elevated from "extra plots" to a decision-critical test of whether model outputs can be interpreted as usable risk scores, making clear that good discrimination without calibration can still produce unreliable threshold actions. Third, reconstruction stability and window-level error dynamics are used as an internal-consistency check that distinguishes sustained, system-level deviations from transient fluctuations, thereby reducing the chance that alarms are driven by score noise or sampling idiosyncrasies. Fourth, representation-structure diagnostics latent geometry shifts and, when enabled, topology-informed summaries such as Wasserstein/bottleneck distances, total persistence trajectories, and Betti/persistence descriptors provide a robustness-oriented validation step that asks whether the model's internal organization changed in a way consistent with a regime deformation, rather than merely producing a different label. Within this unified view, the weaker DEEP configuration is not described simply as "worse," but as exhibiting a specific failure mode—poorer calibration and greater overlap in score distributions—where structural diagnostics become indispensable safeguards: they help decide whether an apparent anomaly reflects a genuine change in the latent organization of the process or an ambiguity produced by limited supervised separability, class imbalance, or no stationarity. Consequently, anomaly monitoring is articulated as an interpretable, audit-ready surveillance procedure that triangulates discrimination, calibration, reconstruction stability, and structural/topological evidence to support defensible decisions about regime change, yielding a more actionable econometric narrative than reliance on a single probabilistic thresholder.

6.2. Conclusions: methodological takeaways, limitations, and reproducibility claims

The empirical evidence generated by the implemented pipeline substantiates the study's objective by delivering an auditable, end-to-end ranking of methods and a defensible interpretation of anomalies as regime-level departures: (i) the supervised baselines (RF and Logistic Regression) consistently achieve benchmark discrimination and operationally reliable probabilities on the evaluated windows, establishing a high-confidence reference for monitoring and thresholder action; (ii) latent-space and reconstruction diagnostics strengthen the methodological argument by demonstrating whether the learned representation remains stable over time and by distinguishing sustained structural perturbations from transient score noise when anomalies are framed as deformations of the normal manifold; and (iii) topology-based summaries when they reveal coherent shifts in connectivity and persistence through Wasserstein/bottleneck trends and persistence/Betti aggregates add a robustness layer that validates whether detected changes correspond to genuine embedding-geometry departures rather than purely statistical fluctuations, while the same evidence transparently delineates limitations and improvement paths, namely that the current DEEP configuration exhibits weaker separability, poorer calibration, and elevated false negatives that require retraining, calibration, and a refined threshold policy before decision-ready deployment, and that topological interpretability is contingent on filtration design and sampling density that may attenuate H1 activity in certain regimes, all underpinned by the artifact-centered reproducibility of exported tables and figures (metrics, ROC/PR, confusion, threshold sweeps, training histories, latent projections, and window-wise structural/topological dynamics) that enable an independent reader to reconstruct each claim directly from code-derived outputs rather than narrative assertion.

The revised interpretation therefore treats the supervised scores as the primary alerting layer, the topological channel as a stability-preserving corroborative layer, and the near-perfect RF/LOGREG metrics as results that remain strong but conditional on the present temporal protocol and feature construction.

7 Future work

7.1. Extensions: broader econometric settings, richer structural metrics, real-time constraints, and deployment

The code-generated results yield clear methodological takeaways: classical supervised baselines (RF and Logistic Regression) provide benchmark-level discrimination and operationally reliable probability estimates, while latent-space and topology-oriented diagnostics offer complementary evidence about whether detected events reflect genuine structural regime change rather than transient score noise. Together, these layers support a monitoring view in which prediction quality and representation stability are evaluated jointly to justify anomaly decisions.

The limitations are equally explicit in the exported evidence: the current DEEP configuration shows weaker separability, poorer calibration, and elevated false negatives, indicating that additional tuning, calibration, and a carefully justified threshold policy are required before it can function as a primary detector. Moreover, topology-based signals can be sensitive to filtration and sampling choices, so the absence of sustained H1 activity in some summaries should be interpreted as a configuration-dependent outcome, not as proof of structural invariance.

Reproducibility is strengthened by the artifact-centric workflow implemented in the code: standardized figures and tables document ROC/PR behavior, confusion structure, threshold sweeps, calibration, training dynamics, latent projections, and window-level structural/topological evolution in a directly auditable form. This packaging enables independent reconstruction of every claim from the exported outputs, supporting transparent verification and faithful replication under the stated pipeline settings.

All the code can be found at the following link: https://github.com/JAIME6609/ECONOMETRIC

References

Abdallah, H., Regalski, A., Kang, M. B., Berishaj, M., Nnadi, N., Chowdury, A., Diwadkar, V. A., & Salch, A. (2022). Statistical inference for persistent homology applied to simulated fMRI time series data. *Foundations of Data Science*, *5(1)*, 1–25. https://doi.org/10.3934/fods.2022014

Bauer, U. (2021). Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology, 5*, 391–423. https://doi.org/10.1007/s41468-021-00071-5

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Böken, B. (2021). On the appropriateness of Platt scaling in classifier calibration. Information Systems, 95, 101641. https://doi.org/10.1016/j.is.2020.101641

Chazal, F., & Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence, 4*, 667963. https://doi.org/10.3389/frai.2021.667963
Dey, T. K., & Wang, Y. (2022). *Computational Topology for Data Analysis*. Cambridge University Press. https://doi.org/10.1017/9781009099950

Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences, 118(8)*, e2016191118. https://doi.org/10.1073/pnas.2016191118

Duque, N., Giraldo, L. G. S., & Arbeláez, P. (2023). Geometry Regularized Autoencoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2022.3222104

Fearnhead, P., & Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association, 114(525)*, 169–183. https://doi.org/10.1080/01621459.2017.1385466

Ferro, M. V., Doval Mosquera, Y., Ribadas Pena, F. J., & Darriba Bilbao, V. M. (2023). Early stopping by correlating online indicators in neural networks. *Neural Networks, 159*, 109–124. https://doi.org/10.1016/j.neunet.2022.11.035

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys, 51(5)*, Article 93, 1–42. https://doi.org/10.1145/3236009

Guilbert, T., Caelen, O., Chirita, A., & Saerens, M. (2024). Calibration methods in imbalanced binary classification. *Annals of Mathematics and Artificial Intelligence, 92(5)*, 1319–1352. https://doi.org/10.1007/s10472-024-09952-8

Hassan, Z., Treude, C., Norrish, M., Williams, G., & Potanin, A. (2025). Characterising reproducibility debt in scientific software: A systematic literature review. *Journal of Systems and Software, 222*, 112327. https://doi.org/10.1016/j.jss.2024.112327

Heo, E., & Jung, J.-H. (2024). Persistent homology of featured time series data and its applications. *AIMS Mathematics, 9(10)*, 27028–27057. https://doi.org/10.3934/math.20241315

Huang, L., Zhao, J., Zhu, B., Chen, H., & Vanden Broucke, S. (2020). A comprehensive study of the effect of different probability calibration methods on imbalanced classification. *IEEE Access, 8*, 127343–127352. https://doi.org/10.1109/ACCESS.2020.3008150

Ichinomiya, T. (2025). Machine learning of time series data using persistent homology. *Scientific Reports, 15*, 20508. https://doi.org/10.1038/s41598-025-06551-3

Olteanu, M., Rossi, F., & Yger, F. (2023). Meta-survey on outlier and anomaly detection. *Neurocomputing, 555*, 126634. https://doi.org/10.1016/j.neucom.2023.126634

Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys, 54(2)*, Article 38. https://doi.org/10.1145/3439950

Ravishanker, N., & Chen, R. (2021). An introduction to persistent homology for time series. *WIREs Computational Statistics, 13(3)*, e1548. https://doi.org/10.1002/wics.1548

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery, 8(2)*, e1236. https://doi.org/10.1002/widm.1236

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLOS Computational Biology, 15(7)*, e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Schackart III, K. E., Istrate, A.-M., Cook, C. E., & colleagues. (2024). Detailed Implementation of a Reproducible Machine Learning-Enabled Workflow. *Data Science Journal, 23*, 23. https://doi.org/10.5334/dsj-2024-023

Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision, 130*, 1526–1565. https://doi.org/10.1007/s11263-022-01611-x

Tuli, S., Casale, G., & Jennings, N. R. (2022). TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment, 15(6)*, 1201–1214. https://doi.org/10.14778/3514061.3514067

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing, 167*, 107299. https://doi.org/10.1016/j.sigpro.2019.107299

Xu, R., Yu, Y., Cui, H., Kan, X., Zhu, Y., Ho, J. C., Zhang, C., & Yang, C. (2023). Neighborhood-Regularized Self-Training for Learning with Few Labels. *Proceedings of the AAAI Conference on Artificial Intelligence, 37(9)*, 10611–10619. https://doi.org/10.1609/aaai.v37i9.26260

Yao, J., Li, J., Wu, J., Yang, M., & Wang, X. (2025). Change point detection in financial market using topological data analysis. *Systems, 13(10)*, 875. https://doi.org/10.3390/systems13100875

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C. C., & Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys, 57(1)*, Article 15. https://doi.org/10.1145/3691338

Zhang, F., Liu, Z., Zeng, Y., & Chen, X. (2025). Topological Time Series Analysis of Market Crashes. *Proceedings of the ACM on Management of Data*. https://doi.org/10.1145/3745533.3745634

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34(12)*, 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: bioinformatics and beyond. *Briefings in Bioinformatics, 24(6)*, bbad375. https://doi.org/10.1093/bib/bbad375