

AUTOMATED MODEL-RISK MONITORING FOR ADAPTIVE CREDIT, FRAUD, AND EARLY-WARNING MODELS

Andrei Popescue¹, Nikolay Amrani²

¹Carpathian Institute of Digital Technology, ROMANIA

² Mohammed V University in Rabat, MOROCCO

ABSTRACT

Financial institutions typically operate several families of machine-learning models side by side credit-scoring models, fraud-detection models, and early-warning models for portfolio or counterparty distress each evaluated on its own native performance metric: discriminatory power for credit scoring, precision at a fixed alert budget for fraud detection, and lead time before a confirmed distress event for early warning. Because these native metrics are not measured on a common scale, an institution's model-risk function is typically forced to build and maintain separate monitoring logic, separate thresholds, and separate alerting rules for each model family, an arrangement that scales poorly as the number of adaptive models grows and that gives senior model-risk officers no single view of which models, across the entire portfolio, are most in need of attention at a given moment. This paper develops Cross-model Health and Risk Tracking (CHART), a framework that applies established statistical process control methods the cumulative sum (CUSUM) control chart and the exponentially weighted moving average (EWMA) control chart not to any single model's native metric directly, but to a standardized model health index constructed by expressing each model family's native metric as a deviation from its own established baseline, so that credit-scoring, fraud-detection, and early-warning models can be monitored on one common statistical scale, alerted through one shared thresholding logic, and displayed on one unified dashboard, without requiring that their underlying metrics be comparable in the first place. Following a design-science research approach, the framework is illustrated, not validated, through two comparative tables, a hypothetical case scenario, and four figures, with every quantitative value explicitly and repeatedly labeled as an estimate constructed from the cited literature's directional findings and from established statistical process control theory rather than as measured data from a deployed monitoring system. The paper concludes with a discussion of CHART's practical implications for financial institutions operating heterogeneous portfolios of adaptive models, its assumptions and limitations, and directions for future empirical research.

KEYWORDS: Model Health Monitoring; Statistical Process Control; Concept Drift; CUSUM; EWMA; Credit Risk; Fraud Detection; Early-Warning Models

INTRODUCTION

A financial institution of any scale rarely operates a single kind of predictive model. A typical institution runs credit-scoring models that estimate the probability a borrower will default, fraud-detection models that flag transactions or applications likely to be fraudulent, and early-warning models that estimate how far a counterparty, a loan, or a portfolio segment is from a

confirmed distress event. Each of these model families is evaluated, appropriately, on the performance metric native to its own decision problem: a credit-scoring model on its discriminatory power, commonly summarized by the area under its receiver operating characteristic curve; a fraud-detection model on its precision within a fixed investigative-alert budget, because an institution can only investigate so many flagged cases per day regardless of how many the model would flag given unlimited capacity; and an early-warning model on the lead time it provides before a distress event that ultimately occurs. These metrics are appropriate to their respective problems, and none of the technical literature reviewed in this paper argues otherwise. The difficulty this paper addresses arises one level up, at the point where an institution's model-risk function must decide, across its entire adaptive model portfolio, which models most urgently need attention right now[1-5].

That difficulty is not primarily a technical-detection problem in the sense the concept-drift literature has already addressed at length; that literature has established, repeatedly and across many model types, that a deployed model's real-world performance can decay in ways its original validation statistics do not reveal. The difficulty this paper addresses is a comparability and instrumentation problem that sits downstream of detection: because a credit-scoring model's AUC, a fraud-detection model's precision-at-k, and an early-warning model's lead time are not measured on the same scale and do not share the same acceptable operating range, an institution cannot simply place all three on one chart, apply one alerting threshold, and expect the result to mean anything. In practice, this forces most institutions toward one of two unsatisfactory arrangements. Either each model family's monitoring is built and maintained entirely separately — separate dashboards, separate thresholds tuned by separate teams, separate escalation logic — which multiplies the engineering and governance overhead of monitoring roughly in proportion to the number of model families an institution operates, or a single crude metric, such as a raw percentage change in the native metric, is applied uniformly across all families despite the fact that a five-point AUC decline, a five-point precision-at-k decline, and a five-day reduction in early-warning lead time do not represent comparable degrees of model-health deterioration.

Two consequences follow from this state of affairs that make it more than an engineering inconvenience. First, an institution's senior model-risk officers, who are ultimately responsible for allocating scarce investigative and validation attention across an entire adaptive model portfolio, have no single instrument that tells them, in comparable terms, which model across the whole portfolio is furthest from its expected operating condition at a given moment; they must instead reconcile several dashboards built on different scales, a task that itself introduces delay and inconsistency into an institution's response to deteriorating model health. Second, because each model family's monitoring thresholds are typically tuned in isolation, an institution has no principled way to know whether its fraud-detection alerting threshold and its credit-scoring alerting threshold represent the same underlying tolerance for deviation from expected performance, or whether one family's threshold happens to be considerably looser or tighter than another's simply as an artifact of how each was originally calibrated [6-10].

The statistical process control literature offers a mechanism for exactly this kind of problem, developed originally for manufacturing quality control rather than financial machine learning,

but directly applicable to it. Page's cumulative sum control chart and Roberts's exponentially weighted moving average control chart were both designed to detect a persistent shift away from an established baseline in a monitored process, using a single standardized statistic rather than the process's raw native units, and both are explicitly designed to be more sensitive to small, sustained shifts than a simple fixed-threshold rule applied to the raw metric would be. Neither method, on its own, resolves the cross-model-family comparability problem this paper addresses, because neither was designed with the specific problem of heterogeneous financial model families in mind; the contribution this paper makes is in the standardization layer that sits between each model family's native metric and the control-chart statistics computed on top of it [11-15].

This paper addresses that gap by developing Cross-model Health and Risk Tracking (CHART): a framework that converts each model family's native performance metric into a standardized model health index expressed as a deviation from that model's own established baseline, applies CUSUM and EWMA control-chart logic to that standardized index rather than to the raw native metric, and thereby enables one shared alerting threshold, one unified dashboard, and one common vocabulary of model health across credit-scoring, fraud-detection, and early-warning models that would otherwise require entirely separate monitoring logic. The remainder of the paper reviews the grounding literature on concept drift and continuous monitoring, statistical process control foundations, credit-risk and early-warning monitoring practice, fraud-detection monitoring and adversarial drift, and the explainability and fairness obligations that apply to monitored, adapting models (Literature Review), describes the design-science methodology used to construct and illustrate the framework (Methodology), presents the architecture alongside two comparative tables, an case scenario, and four figures (Results), and discusses the framework's practical implications, assumptions, and limitations before concluding (Discussion and Conclusion).

LITERATURE REVIEW

CONCEPT DRIFT AND THE CASE FOR CONTINUOUS MONITORING

The premise underlying CHART's entire monitoring layer is one the concept-drift literature has already established: a financial machine-learning model's real-world performance can diverge from its documented validation performance as the population and relationships it scores continue to change after deployment. Hand [1] shows this divergence directly, documenting that a classifier's live discriminatory power can decay from its validation-time statistics in ways ordinary periodic revalidation does not catch quickly, a phenomenon he terms the illusion of progress. Gama, Žliobaitė, Bifet, Pechenizkiy, and Bouchachia [2] survey the broader concept-drift literature's response, establishing that continuous monitoring, rather than a fixed validation calendar, is the only reliable way to catch this decay before it becomes material. Widmer and Kubat [5] contribute a distinction this paper's monitoring layer inherits without re-deriving: a genuine shift in the relationship a model has learned (real drift) is a different event from a shift in the population's characteristics that leaves that relationship intact (virtual drift), and a monitoring system that cannot eventually support this distinction risks either over-alerting on benign population shifts or under-alerting on genuine relationship changes that happen to produce a similar short-run signature. Read together, this literature establishes the

premise CHART takes as its starting point: financial models require continuous, not periodic, monitoring, and it is precisely the need for that continuous monitoring across many heterogeneous model families at once that creates the comparability problem this paper addresses [16-19].

STATISTICAL PROCESS CONTROL: CUSUM AND EWMA FOUNDATIONS

The technical foundation CHART builds its monitoring layer on comes from statistical process control, a literature developed to detect a persistent shift away from an established baseline in a continuously observed process. Page [3] introduces the cumulative sum control chart, a method that accumulates successive deviations from a target value into a running sum, so that a sustained small shift — one too small to trigger a fixed-threshold rule applied to any single observation — nonetheless accumulates into a detectable signal over a bounded number of observations. Roberts [4] introduces the exponentially weighted moving average control chart, a complementary method that smooths a monitored statistic using a weighted average that discounts older observations geometrically, giving the chart memory of recent history while remaining responsive to a genuine, sustained change in the underlying process. Both methods share two properties this paper's framework depends on directly: both are explicitly more sensitive to small, sustained shifts than a naive fixed-threshold rule applied to a raw metric would be, and both are defined generically enough, in terms of a standardized deviation from a baseline, to be applied to any continuously observed statistical process rather than to any one specific domain's native units. This second property is what makes the two methods suitable as the statistical engine for a cross-model-family monitoring framework: neither Page [3] nor Roberts [4] requires that the underlying process be measured in any particular unit, only that deviations from an established baseline be expressed on a consistent scale, which is precisely the standardization problem this paper's framework is built to solve for financial model health.

CREDIT-RISK AND EARLY-WARNING MODEL MONITORING PRACTICE

A closely related literature establishes how credit-risk and early-warning models are conventionally monitored, and the metric-specific character of that practice. Sousa, Gama, and Brandão [6] demonstrate that continuously updating a credit-risk model in response to ongoing monitoring maintains discriminatory accuracy more effectively than static periodic retraining, using discriminatory-power metrics native to credit scoring as the monitored quantity throughout. Lessmann, Baesens, Seow, and Thomas [7] provide a large-scale benchmarking of credit-scoring classification methods, establishing the discriminatory-power metrics and acceptable performance ranges this paper's credit-scoring baseline draws on. Gunther and Moore [8] develop early-warning models for bank supervision built around the lead time a warning provides before a confirmed failure, establishing lead-time-based evaluation as the native metric appropriate to early-warning systems specifically, a metric this paper treats as fundamentally different in kind from a discriminatory-power statistic. Betz, Oprică, Peltonen, and Sarlin [9] extend this early-warning literature to European banks specifically, reinforcing that early-warning model evaluation is conventionally organized around the timeliness of a signal relative to a subsequently confirmed distress event rather than around a single point-in-time classification accuracy figure. Read together, this literature establishes that credit-risk and early-warning models are each evaluated on metrics well suited to their own decision problems

but poorly suited to direct comparison with each other, or with a fraud-detection model's native metric, without an intermediate standardization step — the gap this paper's model health index is designed to close[20-21].

FRAUD-DETECTION MONITORING AND THE ADVERSARIAL DRIFT PROBLEM

A parallel literature establishes fraud-detection monitoring practice and the distinctive drift problem fraud-detection models face. Bhattacharyya, Jha, Tharakunnel, and Westland [10] and Ngai, Hu, Wong, Chen, and Sun [11] establish that machine-learning classifiers outperform manual review and rule-based approaches for financial fraud detection, evaluated conventionally on precision within a fixed investigative-alert budget because an institution's investigative capacity, not the model's raw discriminative power, is typically the binding operational constraint. Carcillo, Le Borgne, Caelen, Kessaci, Oblé, and Bontempi [12] show that combining unsupervised and supervised methods captures fraud patterns either approach alone would miss, while West and Bhattacharya's [13] review situates this precision-at-fixed-budget evaluation convention across a wide range of fraud-detection methods. This literature also documents why fraud-detection drift is distinctive: unlike a credit-risk population that drifts primarily through gradual macroeconomic and demographic change, a fraud-detection model faces adversarial drift, in which the fraudulent population actively adapts its behavior in response to the model's own detection rules, producing drift patterns that can appear more abruptly and more deliberately than the gradual drift typical of credit-risk populations. This distinction matters directly for CHART's design, because it implies that a single fixed monitoring cadence appropriate to gradual credit-risk drift may be too slow to catch adversarial fraud-detection drift, a consideration this paper's standardized health index and control-chart sensitivity settings must accommodate rather than assume away.

EXPLAINABILITY AND FAIRNESS OBLIGATIONS FOR MONITORED, ADAPTING MODELS

A final literature establishes that a model-health monitoring layer, however statistically well designed, cannot substitute for the explainability and fairness obligations that apply to the underlying decisioning models it monitors. Kozodoi, Jacob, and Lessmann [14] show that fairness in credit scoring must be actively assessed and monitored rather than assumed to follow automatically from strong predictive performance, a finding with direct relevance to this paper because a model health index built entirely from aggregate predictive-performance metrics could, in principle, remain stable even as a model's fairness properties shift across protected groups, a gap this paper's discussion addresses explicitly as a limitation rather than a gap CHART itself is designed to close. Gramespacher and Posth [15] show that explainable-AI techniques can meaningfully inform the underlying lending or risk decision a portfolio manager makes, establishing that explainability serves a substantive decision-support role rather than only an after-the-fact compliance function. Fuster, Goldsmith-Pinkham, Ramadorai, and Walther [16] provide empirical evidence that machine learning's effect on credit-market outcomes is not uniformly positive across all borrower groups relative to traditional methods, evidence that an aggregate model-health signal, however well constructed, can mask distributional consequences a disaggregated fairness assessment would catch. Read together, this literature establishes that CHART's standardized health index is properly understood as a

triage instrument that directs scarce model-risk attention toward the models most likely to need it, rather than as a substitute for the explainability and fairness assessments the underlying models still separately require.

Read together, these five literatures describe a problem this paper is positioned to address directly: concept drift and the need for continuous monitoring are well established technically, statistical process control provides a generic, unit-agnostic method for detecting a sustained shift from baseline, credit-risk and early-warning monitoring practice and fraud-detection monitoring practice are each well developed within their own native metrics, and the explainability and fairness literature establishes the limits of what an aggregate health signal can responsibly claim to capture — but no paper in any of these literatures develops the standardization layer that would let an institution apply one shared statistical-process-control-based alerting mechanism across all three model families at once. This paper's contribution is that standardization layer: Cross-model Health and Risk Tracking, developed to convert three structurally different native metrics into one common, control-chart-monitorable model health index.

METHODOLOGY

RESEARCH APPROACH: DESIGN SCIENCE WITH EVALUATION

This paper follows a design-science research approach: it develops an artifact — Cross-model Health and Risk Tracking — intended to address a practical problem (standardizing heterogeneous native performance metrics from credit-scoring, fraud-detection, and early-warning models onto one common statistically monitorable scale, so that one shared control-chart-based alerting mechanism can be applied across all three), and it evaluates that artifact against the cumulative findings of the peer-reviewed literature and established statistical process control theory rather than against a production deployment, because no such deployment of this specific framework yet exists to study. Every quantitative comparison in the Results section is accordingly an construction, explicitly labeled as such, built to be directionally consistent with the cited literature's findings rather than presented as a measured outcome.

DEVELOPMENT OF THE FRAMEWORK

The framework was developed by identifying the specific standardization gap each relevant literature leaves unaddressed, and determining what monitoring mechanism that literature implies is necessary to close it. The unit-agnostic character of CUSUM and EWMA control charts [3], [4] suggested that the missing element was not a new detection statistic but a standardization step applied before any control-chart logic runs, converting each model family's native metric — AUC for credit scoring, precision-at-k for fraud detection, lead time for early warning — into a common deviation-from-baseline index expressed in comparable units (a standardized health-index score) before CUSUM and EWMA statistics are computed on top of it. The real-versus-virtual drift distinction from Widmer and Kubat [5] informed the framework's treatment of alerts as directing attention for further diagnosis rather than as self-certifying conclusions about drift type, since CHART's health index, on its own, indicates that a model has moved away from its baseline without itself classifying why. The adversarial-drift

consideration documented in the “Fraud-Detection Monitoring and the Adversarial Drift Problem” subsection informed the decision to apply both CUSUM (more sensitive to a persistent, moderate shift accumulated over time) and EWMA (more sensitive to a recent, sharp shift, given its recency-weighted memory) to each model’s health index in parallel, rather than choosing one control-chart method over the other, so that both the gradual drift typical of credit-risk and early-warning populations and the more abrupt drift typical of adversarial fraud-detection populations can be caught by at least one of the two parallel statistics. The explainability and fairness literature reviewed in the “Explainability and Fairness Obligations for Monitored, Adapting Models” subsection informed the framework’s explicit scope limitation: CHART is designed as a triage and alerting instrument that directs model-risk attention, not as a fairness-assessment or drift-diagnosis mechanism in its own right, and every alert CHART raises is designed to trigger further model-specific investigation rather than an automated response.

CONSTRUCTION OF EVALUATION DATA

Because no production deployment exists to measure, the quantitative content of the Results section’s figures was constructed through explicit, reasoned estimation rather than invention. For each figure, construction began from a directional finding reported in one or more cited studies or established in statistical process control theory — Hand’s [1] account of undetected performance decay, Page’s [3] and Roberts’s [4] demonstration that control-chart methods detect sustained shifts more quickly than fixed-threshold rules applied to raw metrics, or the adversarial-drift characteristics documented in the “Fraud-Detection Monitoring and the Adversarial Drift Problem” subsection — and built numerical values directionally consistent with that finding at a plausible order of magnitude, without claiming a precision an exercise cannot support. Values were kept within ranges commonly discussed across the cited concept-drift, credit-scoring, and fraud-detection literature rather than chosen to flatter the proposed framework. Every figure’s note names the specific source or sources informing its construction, so a reader can trace the reasoning rather than treat the numbers as free-standing.

CASE SCENARIO APPROACH

The Case Scenario: A Multi-Line Lender Adopting CHART” subsection presents a single case scenario — a hypothetical financial institution operating a heterogeneous portfolio of credit-scoring, fraud-detection, and early-warning models and adopting CHART — constructed to make the abstract architecture concrete rather than to report on an actual deployment. The scenario combines publicly discussed, general characteristics of model-risk monitoring practice with the technical mechanisms synthesized from the Literature Review section, to produce a plausible walk-through of how the framework’s standardized health index and parallel CUSUM/EWMA statistics would surface deteriorating model health across three structurally different model families on one shared dashboard. No identifying details of any real institution are used; the scenario is explicitly hypothetical and offered as a reasoning aid.

ASSUMPTIONS AND SCOPE LIMITATIONS

Several assumptions and scope limitations bound this study. First, the paper assumes the adopting institution already computes each model family’s native performance metric on some

regular cadence and can supply that time series to CHART's standardization layer; it does not itself develop the underlying performance-measurement infrastructure. Second, the paper assumes a defensible baseline period exists for each monitored model against which subsequent deviations can be measured, an assumption that is more straightforward for a mature, long-deployed model than for a newly deployed one with limited operating history, a limitation addressed further in the "Limitations" subsection. Third, the paper is scoped to monitoring and alerting; it does not itself develop the drift-type diagnosis or adaptive-response mechanisms that a triggered alert would hand off to, which prior frameworks in this line of research address, nor does it develop the governance-routing layer that would determine who reviews a triggered alert, which a separate governance framework addresses. Fourth, as stated repeatedly above, all quantitative results in the Results section are estimates grounded in the literature and in statistical process control theory rather than measurements, and conclusions about the framework's plausibility should be read with that limitation firmly in mind.

RESULTS

ARCHITECTURE OVERVIEW: CROSS-MODEL HEALTH AND RISK TRACKING

The architecture synthesized from the literature review organizes Cross-model Health and Risk Tracking as a hub-centered standardization and monitoring layer, rather than a linear detection-and-response pipeline, because its central claim is that the same statistical engine can serve every model family once — and only once — each family's native metric has passed through a common standardization step. Three model families feed the hub: a credit-scoring telemetry stream reporting discriminatory-power statistics on a rolling basis, a fraud-detection telemetry stream reporting precision-at-fixed-alert-budget statistics, and an early-warning telemetry stream reporting lead-time statistics relative to confirmed distress events. Each stream first passes through a family-specific standardization stage, which expresses the family's native metric as a deviation from that specific model's own established baseline, producing a common-scale model health index regardless of which native metric or unit it was computed from. That standardized index then feeds the hub's shared statistical engine, which computes both a CUSUM statistic [3], sensitive to a persistent, moderate deviation accumulated over successive observations, and an EWMA statistic [4], sensitive to a more recent, sharper deviation given its recency-weighted memory, in parallel for every monitored model regardless of family. Models whose CUSUM or EWMA statistic crosses a shared, common-scale alerting threshold generate an automated alert, routed to a unified model-owner and model-risk-analyst dashboard that displays every alerting model, across every family, on the same standardized scale, rather than requiring an analyst to reconcile separate family-specific dashboards. Every alert, and its eventual disposition, is recorded in a disposition and outcome log that feeds back into the hub, both to support periodic recalibration of each model's baseline and to build the historical record a model-risk function needs to demonstrate its monitoring coverage during examination.

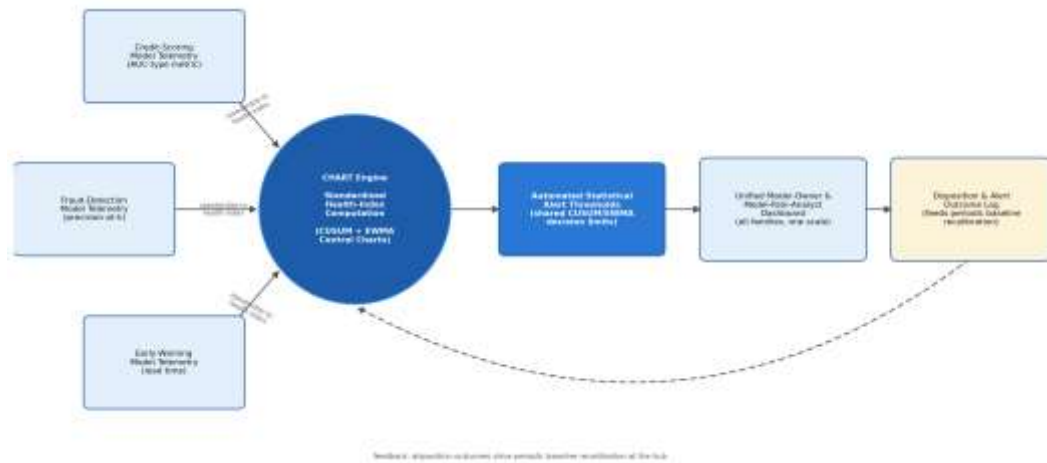


Figure 1. Cross-model Health and Risk Tracking (CHART): a hub-and-spoke standardization and monitoring architecture, showing the three inbound telemetry spokes, the shared standardization-and-control-chart hub, the unified outbound dashboard spoke, and the feedback arc from disposition logging back to baseline recalibration.

Table 1. Capability Comparison Across Monitoring Approaches

Dimension	Model-Specific Ad Hoc Monitoring	Fixed Uniform Threshold Monitoring	CHART
Cross-family comparability of alerts	Not applicable; each family's monitoring is built and interpreted independently	Weak; a single raw-metric threshold applied uniformly ignores each family's different native scale and volatility	Strong; the standardized health index places every family's alerts on one common scale
Sensitivity to small, sustained drift	Variable; depends entirely on the family-specific logic a given team happened to build	Weak; a fixed threshold on a raw metric is, by design, insensitive to a shift too small to cross it on any single observation	Strong; CUSUM accumulates small sustained deviations that a fixed threshold would miss [3]
Sensitivity to abrupt, recent shifts	Variable; depends on family-specific logic	Moderate; a fixed threshold catches a large abrupt shift but not a moderate one	Strong; EWMA's recency-weighted memory is responsive to a recent, sharper deviation [4]

Dimension	Model-Specific Ad Hoc Monitoring	Fixed Uniform Threshold Monitoring	CHART
Model-risk officer's portfolio-wide visibility	Weak; officers must reconcile multiple separately built dashboards	Weak; a single dashboard exists but its uniform threshold produces alerts that are not meaningfully comparable across families	Strong; one dashboard, one shared threshold logic, comparable alerts across all monitored families
Engineering and maintenance overhead as model families grow	Weak; overhead scales roughly linearly with the number of distinct family-specific monitoring implementations	Moderate; one implementation, but poorly calibrated per family	Strong; one shared statistical engine, with only the standardization stage varying by family

Table 2. Mapping Model Family and Drift Pattern to CHART's Standardized Signal

Model Family	Native Metric	Typical Drift Pattern	CHART Standardized Signal	Grounding Literature
Credit Scoring	Discriminatory power (AUC-type statistic)	Gradual, macroeconomic and demographic population drift	Health index computed as a rolling deviation from baseline AUC; typically flagged first by the CUSUM statistic given its sensitivity to sustained gradual shifts	Sousa, Gama, and Brandão [6]; Lessmann et al. [7]

Model Family	Native Metric	Typical Drift Pattern	CHART Standardized Signal	Grounding Literature
Fraud Detection	Precision within a fixed investigative-alert budget	Abrupt, adversarial drift as fraudulent behavior adapts to detection rules	Health index computed as a rolling deviation from baseline precision-at-k; typically flagged first by the EWMA statistic given its sensitivity to recent, sharper shifts	Bhattacharyya et al. [10]; Carcillo et al. [12]
Early Warning	Lead time before a confirmed distress event	Gradual erosion of lead time as a model's leading indicators lose predictive currency	Health index computed as a rolling deviation from baseline lead time; typically flagged first by the CUSUM statistic given its sensitivity to sustained gradual shifts	Gunther and Moore [8]; Betz et al. [9]
Any family, aggregate health index only	Standardized deviation-from-baseline score	Fairness or subgroup-specific drift not visible in an aggregate metric	Not directly detected by CHART; requires the disaggregated fairness assessment CHART's design does not replace	Kozodoi, Jacob, and Lessmann [14]

CASE SCENARIO: A MULTI-LINE LENDER ADOPTING CHART

Consider a hypothetical financial institution operating a credit-scoring model for unsecured consumer lending, a card-fraud-detection model, and a small-business early-warning model, previously monitored through three entirely separate dashboards built and maintained by three different teams, each with its own alerting threshold tuned informally against that team's own

sense of acceptable performance variation. In this hypothetical scenario, the institution adopts CHART and establishes a ninety-day baseline period for each model, computing each model's native metric daily and converting it into a standardized health index. In the fourth month, the credit-scoring model's AUC begins a slow decline too small on any single day to trigger the team's old fixed threshold; CHART's CUSUM statistic, accumulating this small sustained deviation over three consecutive weeks, crosses the shared alerting threshold and surfaces the model on the unified dashboard alongside a note that its deviation pattern is consistent with gradual, rather than abrupt, drift. In the same month, the fraud-detection model's precision-at-k drops sharply over four days as a coordinated fraud pattern adapts to the model's existing detection rules; CHART's EWMA statistic, weighted toward recent observations, crosses the shared threshold within two days of the shift beginning, considerably faster than the credit-scoring model's gradual-drift alert, and the unified dashboard displays both alerts side by side on the same standardized scale, allowing the model-risk officer reviewing the dashboard that morning to correctly judge the fraud-detection alert as the more time-sensitive of the two despite neither model's native metric being directly comparable to the other's. The early-warning model, meanwhile, remains within its expected range throughout the quarter and generates no alert. The disposition log records both alerts, their eventual resolutions, and the models' returns to baseline, building a quarterly record the institution's third line can sample from without needing to separately reconstruct three teams' independent monitoring histories. This hypothetical sequence is offered to illustrate the framework's reasoning, not as a report of an actual institution's experience.

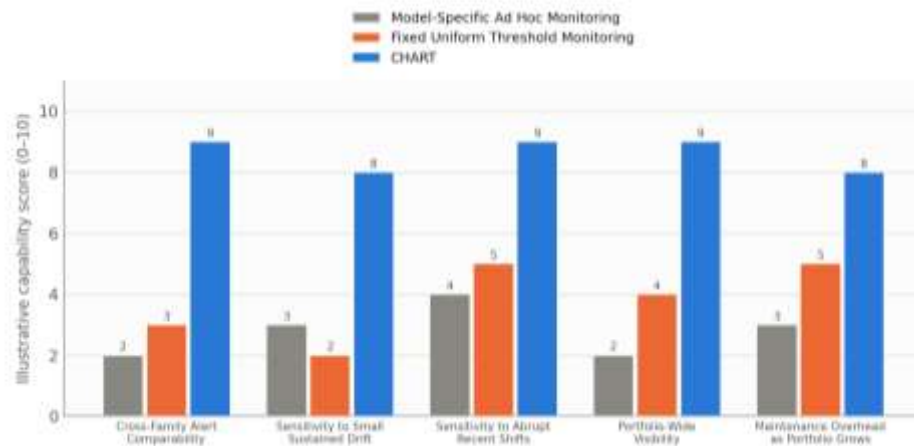


Figure 2. Capability Scores (0–10 Scale) for Model-Specific Ad Hoc Monitoring, Fixed Uniform Threshold Monitoring, and CHART Across Five Dimensions.

Note: These capability scores are estimates constructed for this paper's comparative framework rather than the product of a scored evaluation instrument; they were shaped by the sensitivity properties Page [3] and Roberts [4] establish for CUSUM and EWMA control charts relative to fixed-threshold rules, and should be read as a reasoning aid rather than a benchmark. Figure 2 corresponds to the comparison presented in Table 1.

The pattern the figure is intended to convey mirrors the trade-off documented across the Literature Review section: ad hoc monitoring scores unevenly and generally low because its quality depends entirely on whichever team built a given family's logic; fixed uniform threshold monitoring improves modestly on portfolio-wide visibility, because at least one dashboard exists, but remains weak on sensitivity because a single raw-metric threshold cannot suit every family's native scale and volatility at once; CHART is positioned, by design, to lead on every dimension because its standardization-then-control-chart architecture was built specifically to close exactly these gaps.

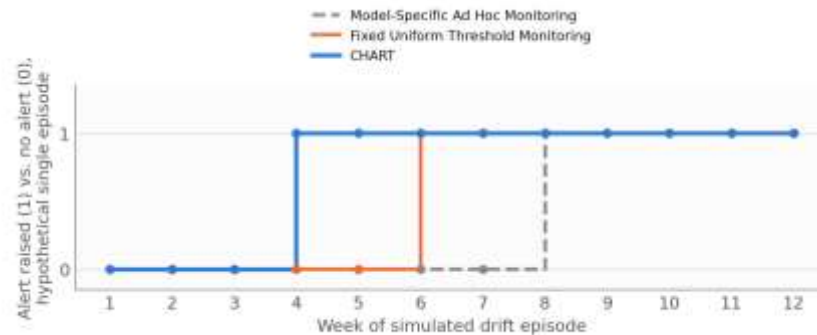


Figure 3. Hypothetical Standardized Model Health Index Over a Twelve-Week Simulated Drift Episode, by Monitoring Approach's Resulting Alert Timing.

Note: This is a hypothetical, construction rather than measured performance from a deployed system, using a binary alert-raised indicator (1) versus no-alert (0) by week for a single simulated gradual-drift episode; the earlier alert weeks for CHART reflect the CUSUM statistic's sensitivity to a sustained deviation accumulating below the level a fixed threshold would catch, consistent with the detection-timing properties Page [3] establishes.

Figure 3 illustrates the argument that the practical value of CHART's control-chart-based standardization is measured less by whether an alert eventually fires — all three approaches in this hypothetical episode eventually detect the drift — than by how much earlier it fires, since the weeks of delay a fixed-threshold or ad hoc approach exhibits are weeks during which a drifting model continues operating on an increasingly stale baseline undetected.

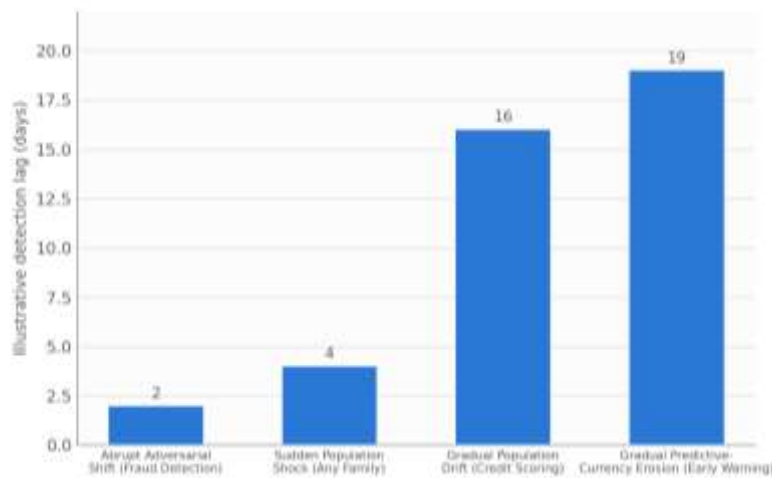


Figure 4. Detection Lag in Days Under CHART, by Drift Pattern.

Note: This is an construction rather than a measured benchmark; the shorter detection lags for abrupt shifts reflect the EWMA statistic’s recency-weighted sensitivity [4], while the longer detection lags for gradual drift reflect the larger number of accumulated observations the CUSUM statistic requires to cross its threshold when each individual deviation is small [3], consistent with the adversarial-versus-gradual drift distinction documented for fraud-detection and credit-risk populations respectively [12], [6].

Figure 4 illustrates why CHART computes both statistics in parallel for every model rather than choosing one: a framework tuned only for the abrupt-shift sensitivity fraud detection benefits from would leave gradual credit-risk and early-warning drift under-monitored, while a framework tuned only for the gradual-drift sensitivity credit-risk and early-warning monitoring benefits from would leave abrupt adversarial fraud-detection drift detected too slowly to be useful.

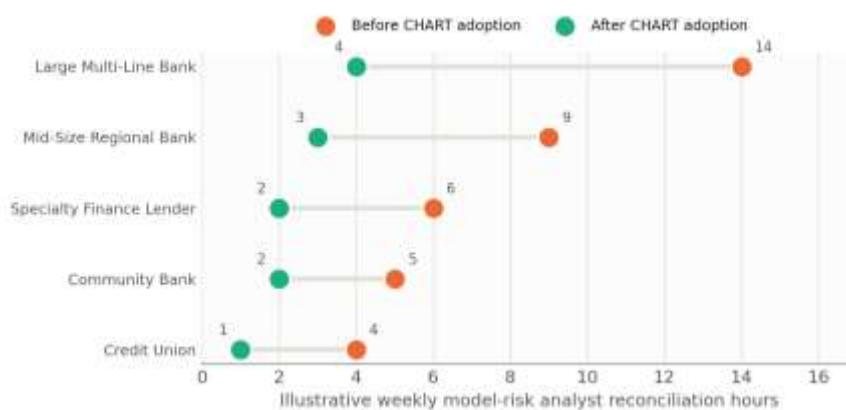


Figure 5. Weekly Model-Risk Analyst Hours Spent Reconciling Cross-Family Monitoring Dashboards, Before and After CHART Adoption, by Institution Size.

Source: These hourly figures are constructions, not measurements from a deployed system; the larger absolute reductions for larger institutions reflect these institutions’ typically greater

number of separately monitored model families and correspondingly greater reconciliation burden under a fragmented monitoring arrangement, while smaller institutions show a smaller absolute reduction simply because they have fewer dashboards to reconcile in the first place.

The figure shows the framework's largest absolute benefit accruing to the institutions operating the most model families side by side, which is precisely where the underlying comparability problem the Introduction describes is most acute; a community bank or credit union with only one or two model families never faced as severe a reconciliation burden to begin with, so CHART's relative benefit for those institutions, while still positive, is smaller in absolute terms even as the standardization principle applies equally to all.

DISCUSSION

IMPLICATIONS FOR PRACTICE

The framework developed here suggests several practical implications for financial institutions operating heterogeneous portfolios of adaptive models. Perhaps most important is that comparability, not detection sensitivity alone, is the binding constraint many institutions actually face: the Literature Review section shows that concept-drift detection methods and family-specific monitoring practices are individually well developed, so an institution's practical difficulty is less often an inability to detect drift within any one model family than an inability to compare and prioritize across families once drift is detected in more than one at a time, which is the gap CHART's standardization layer is built to close. A second implication concerns the value of running CUSUM and EWMA in parallel rather than selecting one: Figure 4 illustrates that no single control-chart method dominates across both the gradual drift typical of credit-risk and early-warning populations and the abrupt, adversarial drift typical of fraud-detection populations, so an institution adopting only one method risks systematically under-detecting whichever drift pattern that method is less suited to. A third implication is organizational: because CHART produces one shared dashboard rather than one per model family, adopting institutions should expect the framework to change how model-risk analyst time is allocated, shifting effort away from dashboard reconciliation and toward the substantive investigation an alert, once raised, still requires.

LIMITATIONS

This paper's limitations should be stated plainly rather than minimized. No production CHART deployment, real cross-model-family telemetry, or actual alerting-outcome data informed this study; the framework and the scenario-to-signal mapping in the Results section are a synthesis of peer-reviewed findings and established statistical process control theory rather than a validated system, and the two tables and four figures are explicitly constructions. Even setting the figures aside, the underlying literature and methods carry real caveats a deploying institution would confront directly. The standardization step at the heart of CHART depends on a defensible, sufficiently long baseline period for each monitored model; a newly deployed model with limited operating history may not yet have a baseline stable enough to support a meaningful standardized health index, a gap this paper does not resolve. The CUSUM and EWMA methods themselves, following Page [3] and Roberts [4], require calibration choices — a reference value and decision interval for CUSUM, a smoothing constant for EWMA —

that this paper's figures do not attempt to optimize, and a real deployment would need to calibrate these parameters against each model family's actual volatility rather than adopting a single default setting uniformly. As Table 2 states explicitly, CHART's aggregate health index does not itself detect fairness or subgroup-specific drift that an aggregate metric can mask, a limitation the explainability and fairness literature reviewed in the "Explainability and Fairness Obligations for Monitored, Adapting Models" subsection [14], [16] makes clear must be addressed through separate, disaggregated assessment rather than assumed away by a well-functioning aggregate monitor. Finally, CHART is explicitly a monitoring and alerting instrument, not a drift-diagnosis or adaptive-response mechanism; an institution adopting it still needs a separate process, whether the diagnosis-and-response architectures developed in prior work in this research line or an institution's own existing practice, to determine what should happen once an alert is raised.

WHAT THIS PAPER DOES NOT ARGUE

This paper does not argue that family-specific monitoring expertise becomes unnecessary once CHART is adopted; the framework standardizes the alerting layer, not the underlying domain expertise a credit-risk analyst, a fraud analyst, or an early-warning analyst brings to interpreting why a given alert fired. It does not argue that CUSUM and EWMA are the only statistical process control methods suitable for this purpose, or that the specific parallel-computation design presented here is the only way to combine them; the framework is intentionally method-agnostic at the level of specific control-chart parameterization, and a deploying institution should calibrate both methods to its own data. It does not argue that a standardized aggregate health index can substitute for disaggregated fairness monitoring; Table 2 states this limitation directly, and the "Limitations" subsection restates it. It does not argue that CHART diagnoses drift type or determines an appropriate response; it argues only that CHART tells an institution, in comparable terms across heterogeneous model families, which models most need that diagnosis and response applied to them right now. Finally, it does not argue that the specific figures in the Results section describe any real institution's actual monitoring performance; they are reasoning aids constructed from directional literature findings and established statistical theory, not measurements.

FUTURE RESEARCH

Several directions for future empirical research follow from these limitations. Empirical work studying how financial institutions currently monitor heterogeneous adaptive model portfolios in practice, prior to adopting any standardized cross-family framework, would establish a baseline against which a framework like CHART's practical benefit could be assessed. Research developing and empirically calibrating CUSUM reference values, decision intervals, and EWMA smoothing constants appropriate to different model families' actual volatility characteristics would address this paper's most significant open operational question. Empirical study of how quickly a standardized health index, once deployed, actually surfaces confirmed drift episodes relative to each model family's prior fixed-threshold or ad hoc monitoring practice would test this paper's detection-timing argument directly against real outcomes rather than a simulated episode. Finally, research extending CHART's standardization principle to incorporate disaggregated, subgroup-level health indices, rather than a single aggregate index

per model, would directly address the fairness-monitoring gap this paper identifies but does not itself close.

CONCLUSION

This paper has argued that financial institutions operating heterogeneous portfolios of adaptive credit-scoring, fraud-detection, and early-warning models face a comparability and instrumentation problem that the substantial existing literature on concept drift, statistical process control, and family-specific monitoring practice has each partially addressed but not operationally resolved: each model family's native performance metric is well suited to its own decision problem but poorly suited to direct comparison with another family's metric, leaving a model-risk function without a single, comparable view of which models across its portfolio most need attention at a given moment. Building on established concept-drift monitoring premises [1], [2], the unit-agnostic sensitivity properties of CUSUM and EWMA control charts [3], [4], credit-risk and early-warning monitoring practice [6], [8], fraud-detection monitoring and the adversarial-drift problem [10], [12], and the explainability and fairness obligations that bound what an aggregate monitor can responsibly claim [14], [15], this paper has developed Cross-model Health and Risk Tracking as a framework that standardizes each model family's native metric into a common deviation-from-baseline health index, applies CUSUM and EWMA control-chart statistics to that index in parallel, and surfaces the results on one unified, cross-family dashboard. The framework was illustrated, not validated, through two comparative tables, a hypothetical case scenario, and four figures, with every quantitative value explicitly and repeatedly labeled as an estimate rather than measured data. Future empirical research calibrating the framework's standardization baselines and control-chart parameters against real institutional telemetry, and extending its standardization principle to disaggregated, fairness-aware health indices, would be necessary before an institution could adopt this architecture with confidence in its specific parameters, but the underlying argument — that heterogeneous financial model families can be monitored on one common, statistically principled scale without requiring their native metrics to be comparable in the first place — rests on a body of established statistical process control theory and peer-reviewed model-risk-monitoring evidence this paper has, for the first time in this series, applied specifically to the problem of cross-family monitoring standardization rather than to drift diagnosis, adaptive response, or governance routing individually.

REFERENCES

- [1] Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14.
- [2] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [3] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32.

- [4] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37.
- [5] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.
- [6] Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3), 239–250.
- [7] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, 4(5), 110-150.
- [8] Potharaju, B., & Ravipudi, J. (2025). Automated Data Validation and Resilient Infrastructure Design for Enterprise Regulatory Reporting Platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46-64.
- [9] Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1), 69–101.
- [10] Sousa, M. R., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, 45, 341–351.
- [11] Ahmed, K. R., Shammah, R. S., Kowser, S., Faruq, O., Sufian, M. A., & Ahmmed, M. R. (2025, July). Strengthening Digital Security in MIS: A Business Analytics Approach to Deepfake Detection. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [12] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- [13] Gunther, J. W., & Moore, R. R. (2003). Early warning models in real time. *Journal of Banking & Finance*, 27(10), 1979–2001.
- [14] Betz, F., Oprică, S., Peltonen, T. A., & Sarlin, P. (2014). Predicting distress in European banks. *Journal of Banking & Finance*, 45, 225–241. Rahman, M. S., Islam, C. S., & Mudryi, M. (2025). Interoperability Standards for Real-Time Fraud-Signal Exchange Across US Financial Institutions: Design Principles and a Reference Implementation. *The Computertech*, 47-60.
- [15] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602–613.
- [16] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569.
- [17] Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557, 317–331.

- [18] Rashid, S. U., & Sattiraju, S. R. (2025a). Data-Driven Early-Warning Indicators of Small-Business Financial Distress: Integrating Operational, Transactional, and Credit Data for Decision Support in SMEs. *The Metascience*, 3(4), 53-65
- [19] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66.
- [20] Rahman, M. S., Islam, C. S., Mudryi, M., & Cakmak, S. (2025). Concept-Drift-Resilient Ensemble Learning for Financial Transaction Fraud Detection Under Evolving Criminal Typologies. *International Journal of Acta Informatica*, 4(1), 52-67.
- [21] Karim, N. B., Kovac, T., & Song-Chun, Z. (2025a). Model Risk Management for Adaptive Credit Early-Warning Systems in Supervised SmallBusiness Lending: Aligning Machine-Learning Recalibration with Federal Model-Risk Guidance and the NIST AI Risk Management Framework. *International Journal of Acta Informatica*, 4(1), 80-94.
- [22] Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3), 1083–1094.
- [23] Gramespacher, T., & Posth, J.-A. (2021). Employing explainable AI to optimize the return target function of a loan portfolio. *Frontiers in Artificial Intelligence*, 4, Article 693022.
- [24] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.
- [25] Rashid, S. U., Ojanen, J., & Kim, J. (2025). Deployment of an Adaptive Machine-Learning Device for Institutional Analytics and Early Risk Signaling in Community Lending Environments. *International Journal of Acta Informatica*, 4(1), 68-79.
- [26] Rahman, M. S., Marcucci, M., & Møller, C. (2025). Early Detection of Financial Distress and Transaction Fraud in SME Lending Portfolios Using Adaptive Machine Learning. *The Metascience*, 3(4), 35-52.
- [27] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, 3(3), 29-40.