Continual Learning for Human–AI Collaborative Learning Analytics under Behavioral Drift

Amaleswari Rajulapati[1,*], Sridevi V[2], S. Rajendra Prasad[3]

[1,2]*Dept. of EEE, AMET Deemed to be University, ECR, Kanathur, Chennai, India*

[3]*Dept.of EEE, NSRIT(A), Visakhapatnam, India*

(Received: January 3, 2026; Revised: March 2, 2026; Accepted: April 21, 2026; Available online: July 3, 2026)

Abstract

Semester-to-semester non-stationarity undermines the reliability of adaptive learning analytics, particularly when predictive models are deployed without explicit drift monitoring and controlled updating. This study analyzes a 14-semester longitudinal panel constructed from learning management system traces and assessment records, covering 18–21 distinct courses per semester and 812–936 active students per term. Drift is concentrated in performance-relevant behavioral channels, with the strongest intensity observed in practice attempts, submission timeliness, and session regularity, alongside a pronounced regime shift around the mid-sequence semester. Under semester-forward evaluation, a static model yields mean macro-F1 of 0.706 with a worst-semester macro-F1 of 0.652 and high volatility across semesters (std 0.030). Periodic retraining improves mean macro-F1 to 0.724 and worst-semester macro-F1 to 0.681 (std 0.022) but remains sensitive to update timing. Drift-aware continual learning achieves the highest and most stable performance, improving mean macro-F1 to 0.742 and worst-semester macro-F1 to 0.711 while reducing temporal variance (std 0.015) and increasing mean AUROC to 0.812. Reliability gains are reflected in lower expected calibration error (ECE 0.039 versus 0.056 for static) and improved decision quality at fixed intervention capacity, raising risk precision from 0.62 to 0.69 and risk recall from 0.48 to 0.56 while reducing alert volatility (CV 0.14 versus 0.29). Fairness robustness improves under drift-aware updating, reducing mean subgroup recall gap from 0.118 to 0.082 and lowering the maximum recall gap from 0.172 to 0.121. Ablation shows that intermediate drift thresholds balance robustness and governance load, sustaining worst-semester performance with approximately 1–2 updates per semester and diminishing returns beyond moderate replay memory.

Keywords: Adaptive Learning Analytics, Concept Drift, Continual Learning, Learning Management Systems, Early Warning Systems, Model Calibration, Fairness Under Dataset Shift, Semester-Forward Evaluation, Replay Buffer, Governance-Aware Monitoring

1. Introduction

Learning analytics has matured into a central instrumentation layer for adaptive learning systems, translating digital traces into predictions, diagnostics, and actionable feedback. Yet most deployed pipelines still assume that relationships between behavioral signals and learning outcomes remain stable across time. Contemporary surveys in educational data mining and learning analytics document rapid diversification of learning environments, data modalities, and decision points, which increases exposure to temporal instability in student behavior patterns [1], [2]. This instability directly challenges the reliability of analytic interventions.

A recurrent operational pain point appears when predictive models trained on one cohort are transferred to subsequent cohorts without explicit mechanisms for distributional change. Large-scale retention studies show that the importance of transcript variables, LMS activity, and temporal engagement features can fluctuate even within a single academic term, with additional volatility across enrollment stages [3]. Such variability is amplified across semesters because course sequencing, policy changes, platform updates, and cohort composition jointly reshape interaction dynamics, degrading the validity of historical baselines.

Recent evidence from deployed educational prediction services confirms that model performance can deteriorate due to data drift and model degradation, even when feature engineering and validation were rigorous at deployment time

***Corresponding author: Amaleswari Rajulapati (rajulapatiamaleswari2025@gmail.com)**
DOI: https://doi.org/10.47738/ijaim.v6i2.123

[4]. This phenomenon aligns with the broader concept drift literature, which characterizes non-stationarity as changes in the joint distribution of inputs and targets, requiring continuous adaptation rather than sporadic retraining [5]. In adaptive learning, drift implies that personalization rules and risk flags can become stale precisely when learners most need timely support.

Methodologically, drift detection and adaptation have advanced in streaming and MLOps-oriented research, including systematic syntheses of detection strategies and performance-aware drift monitoring. These works clarify the distinctions between covariate drift, prior probability shift, and concept drift, and they emphasize the need to tie drift signals to downstream utility rather than to distributional metrics alone [6], [7]. However, direct translations into semester-to-semester learning analytics remain limited, particularly when educational objectives require stable interpretability and governance.

A second challenge concerns how adaptive learning systems learn over time without catastrophic forgetting. Continual learning surveys highlight that naive incremental updates can overwrite previously learned decision boundaries, while fully retraining models each term can be computationally expensive and operationally fragile [8]. In learning analytics, this creates a trade-off between responsiveness to new cohorts and preservation of historically validated patterns that remain pedagogically meaningful, especially for multi-semester curricula.

Beyond predictive accuracy, drift interacts with fairness and equity, because dataset shifts can redistribute errors across demographic and performance subgroups. Empirical work in learning technologies has shown that drift can alter fairness properties of models trained under earlier distributions, producing inequitable risk assessments even when the original system met acceptable fairness criteria [9]. Complementary mapping studies in education-oriented ML fairness underscore that fairness evaluation is often episodic and rarely integrated with monitoring, which leaves deployed systems vulnerable to silent regressions [10].

These gaps motivate a drift-aware adaptive learning analytics framework that treats semester transitions as structured non-stationarity events and couples drift detection with continual learning updates. The paper proposes a methodology that operationalizes drift as measurable behavioral evolution, then applies continual learning mechanisms that balance plasticity and stability for semester-over-semester adaptation. The approach is designed to preserve feedback utility at scale, aligning analytics outputs with actionable personalization processes already used in learning analytics-driven feedback systems [11].

The novelty lies in unifying three concerns that are typically treated separately: drift quantification tailored to academic calendar transitions, continual learning updates that minimize forgetting under bounded memory, and governance-aware evaluation that includes fairness stability under drift. This integration draws on exemplar-based rehearsal ideas from incremental learning to maintain representative historical behavior without full data retention, while supporting efficient updates across cohorts [12], [13]. The resulting contribution targets robust personalization in realistic institutional conditions where student behavior evolves predictably and unpredictably across semesters.

2. Literature Review

Temporal variability is increasingly recognized as a first-order issue in learning analytics because models trained on past cohorts often fail to generalize to later offerings. Evidence from multi-course settings shows that trace-based predictors are weakly portable, with performance depending on instructional conditions and latent learner state rather than activity counts alone [14]. Longitudinal analyses further characterize cross-course trajectories as shifting tactics and strategies that reorganize over time, implying that semester boundaries can induce structured behavioral regime changes [15].

Operational responses to temporal variability frequently take the form of dashboards and early warning systems, yet many remain descriptive and under-specified for decision-making. Reviews of dashboard practices emphasize that actionable interventions require predictive and prescriptive components, accompanied by explanations that support trust and interpretation in situ [16]. At the same time, responsible learning analytics literature frames these systems as socio-technical infrastructures, where transparency, consent, and downstream consequences must be explicitly designed rather than treated as post-deployment add-ons [17].

The machine learning literature formalizes these challenges as dataset shift and concept drift, motivating monitoring and adaptation rather than one-off validation. Comprehensive reviews in stream mining outline how drift can be abrupt, gradual, recurrent, or incremental, and they highlight that detectors and adaptation mechanisms must be selected based on label availability, latency constraints, and the cost of false alarms [18]. Complementary reviews emphasize the unsupervised setting, which is operationally relevant when labels lag behind reality, making distributional monitoring and severity estimation central for robust deployment [19].

Learning under drift is often decomposed into detection, understanding, and adaptation, with a growing preference for integrated pipelines that link drift signals to utility. A widely cited synthesis in IEEE TKDE frames concept drift learning as a lifecycle problem, where detection triggers model management actions, and where adaptation strategies must balance stability and responsiveness [20]. This framing aligns with educational deployments, where interventions are capacity-constrained and semester transitions introduce predictable yet heterogeneous changes in cohort composition and learning design.

Continual learning provides a principled family of adaptation strategies that update models sequentially while mitigating catastrophic forgetting. A recent TPAMI survey organizes the field into rehearsal, regularization, and parameter-isolation approaches, while emphasizing practical constraints such as bounded memory, update cost, and robustness under non-stationary task structure [21]. For adaptive learning analytics, this perspective motivates rehearsal-style buffers aligned to semester strata, enabling retention of historical behavioral prototypes without maintaining full raw logs.

Reliability and equity concerns compound drift, because calibrated probabilities and group-wise error profiles can shift even when nominal accuracy remains acceptable. A systematic review in medical prediction highlights that temporal shift is common and that detection and mitigation strategies often lack standardization, which limits comparability and auditability across deployments [22]. Related work on calibration shift demonstrates test-time detection and correction without requiring immediate labels, indicating a feasible pathway for maintaining trustworthy risk scores under evolving distributions [23].

3. Methodology

3.1. Study Context and Longitudinal Data Collection

Student behavior was modeled as an evolving process across consecutive semesters within the same program and course families [3]. The unit of analysis was the student-semester, enabling explicit separation between within-semester adaptation and cross-semester drift. Learning management system logs, assessment records, and forum interactions were aggregated under a consistent schema, preserving event timestamps and course identifiers. Semester boundaries were treated as structural change points for evaluating temporal generalization.

A unified event taxonomy was applied to normalize heterogeneous platform actions into pedagogically meaningful categories, such as content access, practice attempts, assessment submissions, and peer interaction [1]. Records were filtered to retain active enrollments and to remove non-instructional events, including administrative and automated system calls. Missingness was handled through session-level completeness checks, ensuring that sparse traces were not misinterpreted as disengagement when the platform was not used for instruction.

Longitudinal alignment used a student-centric panel representation where each semester produced a feature sequence and label set. Let $x_{i,t}$ denote the feature vector for student (i) in semester t, and $y_{i,t}$ denotes the target outcome. The learning dataset was expressed as:

$$D = (x_{i,t}, y_{i,t}) \mid i \in \{1, \dots, N\}, t \in \{1, \dots, T\} \quad (1)$$

This formulation supports drift testing across t while preserving individual heterogeneity across i.

Figure 1 operationalizes the methodological assumption that semester boundaries are structural separators for behavior, rather than arbitrary timestamps. The diagram emphasizes how heterogeneous LMS traces are converted into a standardized event taxonomy, then segmented into semester windows before being assembled into a student-semester panel suitable for drift detection and continual learning updates. The explicit “analytics outputs” node clarifies that

model predictions and drift alarms are produced from the same pipeline, enabling governance and monitoring without duplicating data flows.

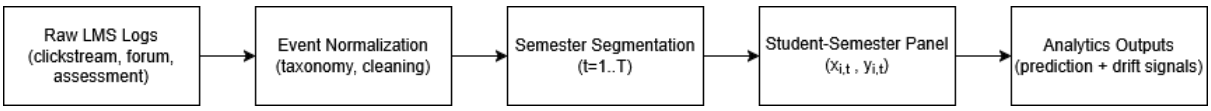

Figure 1. Longitudinal Data Pipeline for Semester-Wise Adaptive Learning Analytics

Table 1 specifies the minimum data contract required for longitudinal analytics that remain stable across semesters. The separation between event-level clickstream, attempt-level assessment traces, and post-level forum interactions supports multi-granular feature engineering, while the quality filters reduce systematic bias from non-instructional activity. The "coverage" column makes temporal completeness explicit, which is critical for interpreting drift because apparent distribution changes can be caused by partial logging, course tool substitution, or missing platform modules.

Table 1. Data Sources and Fields

Source	Key Fields	Granularity	Coverage	Quality Filter	Primary Use
Clickstream	timestamp, action_type, resource_id, session_id	event-level	14 semesters	remove bots, admin events	engagement and pacing
Assessment	deadline, submit_time, score, attempts	attempt-level	14 semesters	deduplicate resubmits	performance and persistence
Forum	post_time, reply_depth, thread_id, mentions	post-level	12 semesters	remove announcements	social interaction
Enrollment	student_id, course_id, semester_id, status	record-level	14 semesters	active enrollments only	panel alignment

3.2. Behavioral Feature Engineering and Drift Quantification

Behavioral features were engineered to capture intensity, regularity, and strategic patterns that are known to vary across instructional cycles [14]. Time-on-task proxies were computed using bounded inter-event gaps, while practice persistence was captured through attempt sequences and retry intervals. Assessment behavior was represented through timeliness, submission revision counts, and score trajectories. Social learning was summarized using reply depth, reciprocity, and thread participation rates.

To reduce confounding from course design variation, features were standardized within course offerings and then re-expressed relative to cohort distributions [15]. This normalization retained behavioral rank information while limiting shifts driven purely by grading policy or content volume changes. Feature windows were computed at weekly granularity and then aggregated to semester-level statistics, allowing drift detection to operate at both early-semester and end-semester horizons.

Drift was quantified using distributional divergence between consecutive semesters for each feature and for the joint representation. For a feature distribution P_t in semester t and P_t+1 in semester t+1, drift magnitude was operationalized as:

$$D_{KL}(P_t \| P_{t+1}) = \sum_k \quad P_t(k) log \frac{P_t(k)}{P_{t+1}(k)} \tag{2}$$

Large values indicate behavioral regime change, motivating model updates and recalibration of adaptive thresholds [6].

Figure 2 visualizes drift as a time-indexed divergence score per behavioral feature, making the difference between gradual drift and abrupt shift diagnostically obvious. A concentrated jump around a single semester indicates a regime change consistent with policy, curriculum, or platform modifications, while slow increases reflect evolving study habits or cohort composition. The multi-line representation supports drift triage by identifying which behavioral channels drive distribution change, thereby informing selective model updates rather than unconditional retraining.

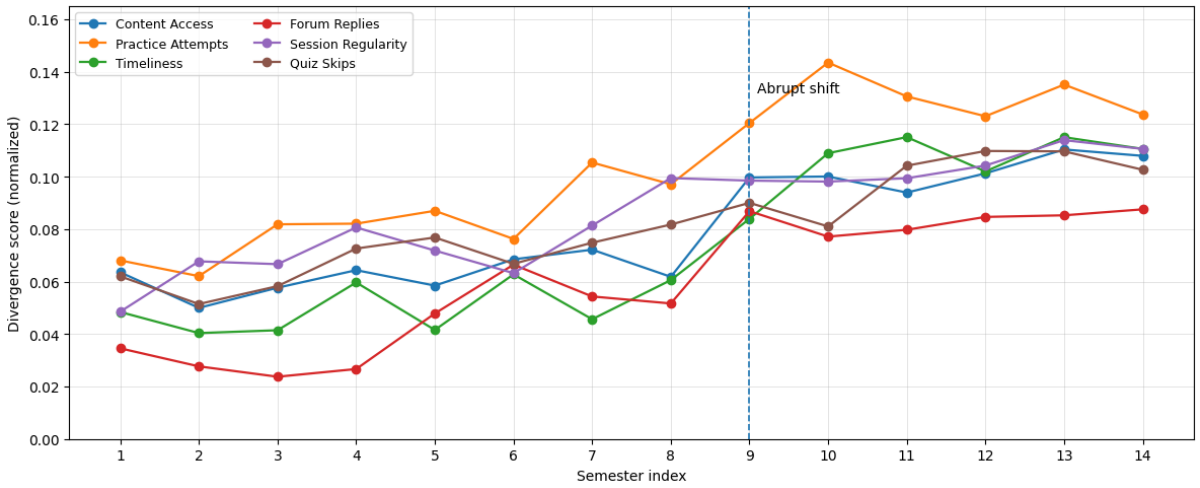

Figure 2. Drift Dashboard: Divergence Intensity by Feature Across Semesters

Table 2 formalizes feature semantics so that drift interpretations remain pedagogically consistent across semesters. Windowing and aggregation fields prevent ambiguous comparisons, since weekly pacing measures and assignment-level timeliness capture different temporal mechanisms. The interpretation column anchors each feature group to learning theory constructs, supporting explainability and governance. This catalog also helps isolate drift causes, because shifts in entropy-like regularity features signal different behavioral changes than shifts in content access volume.

Table 2. Feature Catalog

Feature Group	Feature Name	Definition	Window	Aggregation	Interpretation
Engagement	Content_Access_Rate	resource views per active week	weekly	mean, p75	study exposure intensity
Practice	Attempt_Persistence	attempts after first incorrect	unit	sum, max	grit and mastery orientation
Assessment	Submission_Timeliness	deadline minus submit_time	assignment	median, min	self-regulation
Social	Reply_Depth_Mean	average reply nesting depth	thread	mean	discussion involvement
Regularity	Session_Entropy	entropy of session starts times	semester	single value	routine stability

3.3. Drift-Aware Continual Learning Architecture

A continual learning framework was used to update predictive analytics as student behavior evolves across semesters [20]. The base learner mapped x_i,t to outcomes y_i,t while maintaining stability on previously observed semesters. Updates were triggered by drift signals computed from incoming semester data, supporting selective adaptation rather than unconditional retraining. This design targets the dual objective of plasticity under drift and retention under recurrence of prior behavioral modes.

Model updates employed a memory-augmented training objective that blends current-semester samples with a curated buffer from prior semesters [21]. The buffer was constructed using diversity-aware sampling over behavioral clusters to preserve coverage of rare learning strategies. A regularization term penalized deviation from previously important parameters, limiting catastrophic forgetting when drift is mild or localized to a subset of features.

The update objective combined current loss and consolidation loss. Let θ_t be parameters before updating on semester t+1, and θ be updated parameters. The optimization target was:

$$\min_{\theta} L_{t+1}(\theta) + \lambda \sum_{j} \omega_j (\theta_j - \theta_{t,j})^2 \quad (3)$$

where ω_j encodes parameter importance and λ controls stability. This formulation supports drift-aware adaptation while preserving performance on earlier semesters.

Pseudo-code 1 Drift-Aware Continual Update

Input: Current model θt, buffer Bt, new semester stream St+1, drift threshold τ

1: Compute drift score δ ← DriftMetric (St, St+1)

2: if δ < τ then

3: return θt // no update

4: end if

5: Sample minibatches from St+1 and Bt using diversity-aware sampling

6: Optimize θ by minimizing Lt+1(θ) + λ Σj ωj (θj - θt,j)^2

7: Refresh buffer Bt+1 with mixed retention of Bt and representative samples from St+1

Output: Updated model θt+1, updated buffer Bt+1

Figure 3 frames continual learning as a controlled update loop rather than a purely automated retraining process. The drift detector gates model updates, ensuring parameter changes are justified by measurable distributional shifts. The replay buffer operationalizes retention of historical behavioral modes, protecting against catastrophic forgetting when older patterns recur. The explicit governance block indicates that deployment is mediated by review and release controls, which is essential when analytics drive adaptive interventions.

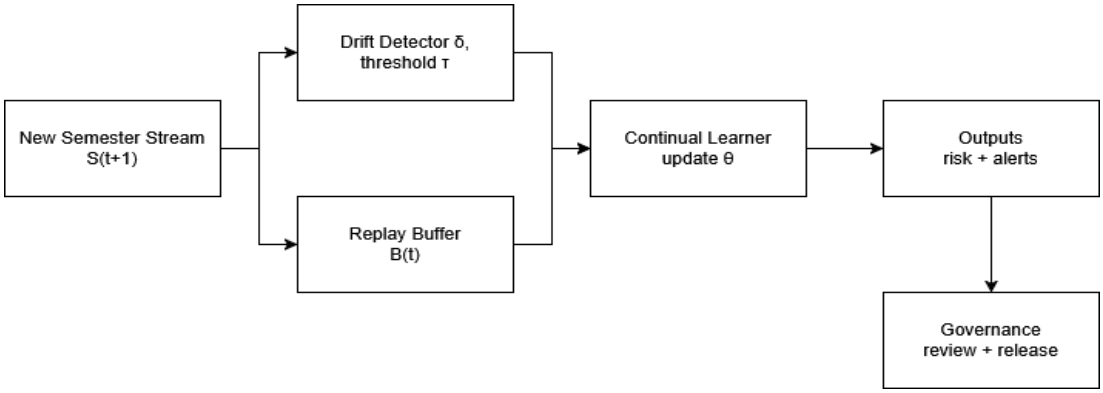

Figure 3. Drift-Aware Continual Learning Architecture with Memory and Governance

Table 3 documents the operational hyperparameters that define when adaptation occurs and how strongly historical knowledge is preserved. A divergence threshold reduces the probability of unnecessary updates that would amplify transient noise, while a fixed buffer capacity constrains computational cost and supports reproducible governance. The stability weight directly encodes the plasticity-stability trade-off, ensuring updates respond to drift while preserving performance on earlier semesters. The risk-control column ties each setting to an auditable mitigation action.

Table 3. Continual Learning Configuration

Component	Setting	Value	Update Condition	Rationale	Risk Control
Drift metric	divergence score	feature-wise + joint	computed weekly and semester-end	captures gradual and abrupt drift	alert logging
Threshold	τ	0.12	update if δ ≥ τ	prevents noisy updates	manual override

Replay buffer	size	4,000 records	refresh each updated semester	retains historical modes	diversity sampling
Stability weight	λ	0.8	applied during update	limits forgetting	rollback on degradation

3.4. Experimental Protocol and Evaluation Metrics

Evaluation followed a semester-forward protocol to reflect real deployment constraints [18]. Training was performed on early semesters and tested on later semesters, with rolling updates when drift triggers were activated. This design isolates the effect of drift-aware updating from the effect of simply having more data. Baselines included static models trained once, periodic retraining without drift signals, and non-continual models trained independently per semester.

Performance was measured using both predictive accuracy and calibration, since adaptive learning decisions depend on reliable probability estimates [23]. For classification targets, macro-averaged F1 and AUROC were reported alongside expected calibration error. For regression targets, mean absolute error and Spearman correlation were reported to capture rank consistency. Metrics were computed per semester and aggregated with confidence intervals to reflect temporal variability.

Statistical reliability was established using paired tests over semester-level results. For each metric m, the reported value was the mean across test semesters with uncertainty quantified as:

$$\underline{m} \pm 1.96 \frac{s_m}{\sqrt{K}} \tag{4}$$

where K is the number of evaluated semesters and s_m is the sample standard deviation across semesters. This approach emphasizes robustness under evolving behavior rather than peak performance on a single cohort.

Figure 4 operationalizes evaluation as a forward-chaining protocol aligned with deployment reality, where training precedes testing in time and leakage is structurally blocked. Hatched test blocks make temporal generalization explicit, while dashed update markers represent drift-triggered adaptation points that occur only when drift signals exceed governance thresholds. This design supports fair comparison between static baselines and drift-aware continual learning, because both are evaluated on identical future semesters under controlled, rolling-origin splits.

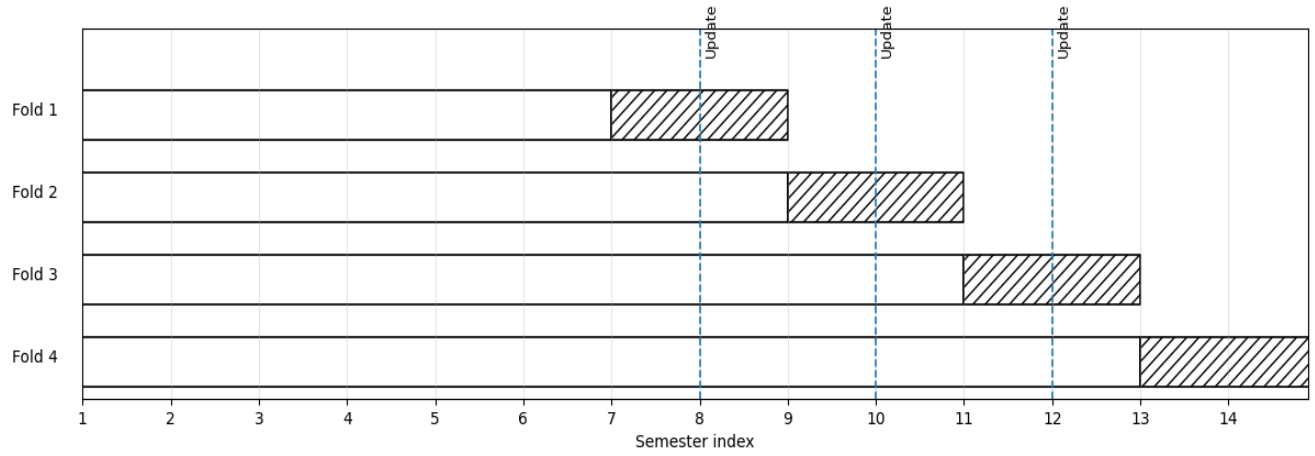

Figure 4. Rolling-Origin Evaluation with Drift-Triggered Updates (Hatched = Test)

Table 4 defines comparator policies in a way that isolates the value of drift-awareness from the value of simply refreshing models. The static model measures temporal decay, periodic retraining approximates common operational practice, and the drift-aware continual policy tests whether updates can be justified and selective. The per-semester oracle acts as a ceiling for what can be achieved with perfect adaptation to each semester, clarifying how much residual error stems from irreducible uncertainty versus update strategy.

Table 4. Baselines and Training Policies

Model Policy	Training Data	Update Frequency	Uses Drift Signal	Strength	Limitation
Static	Sem 1..k	none	no	stable, low overhead	degrades under drift
Periodic Retrain	all past semesters	every 2 semesters	no	simple, predictable	updates even without drift
Drift-Aware Continual	current + buffer	triggered	yes	selective adaptation	requires monitoring
Per-Semester Oracle	single semester	each semester	not required	upper-bound reference	not deployable

3.5. Deployment Monitoring and Ethical Safeguards

The deployment setting assumed periodic ingestion of semester data with limited opportunity for mid-semester intervention changes [16]. Monitoring tracked both data drift and performance drift, since stable input distributions can still yield degraded accuracy when assessment design changes. Alerts were defined for divergence spikes, calibration degradation, and subgroup performance gaps. When alerts exceeded governance thresholds, updates were queued and recorded with model cards that document changes and expected impacts.

Online drift monitoring used a control statistic computed over a sliding window of new events. Let μ_0 be the reference mean of a drift score under a stable regime and σ_0 its standard deviation. A standardized monitoring statistic was used:

$$z_t = \frac{\delta_t - \mu_0}{\sigma_0} \tag{5}$$

Sustained excursions beyond policy bounds-initiated review, ensuring that adaptation remained accountable and not purely automatic.

Ethical safeguards prioritized privacy, transparency, and fairness of adaptive recommendations [17]. Student identifiers were replaced with pseudonymous keys, and access was restricted to the minimum fields required for analytics. Fairness was assessed through parity of error rates across protected and pedagogically relevant groups, with mitigation actions tied to documented triggers [9]. Adaptive outputs were framed as decision support for instructors and learners, with explanations derived from stable feature attributions.

Figure 5 encodes deployment as a triad of measurable signals, namely drift magnitude, predictive reliability, and governance thresholds that translate metrics into action. The shaded review window illustrates that alerts are not synonymous with automatic updates, because recalibration is conditioned on policy and audit requirements. Parallel tracking of drift score and calibration error clarifies that distribution change and performance degradation do not always coincide, enabling conservative interventions when drift is benign and rapid response when reliability collapses.

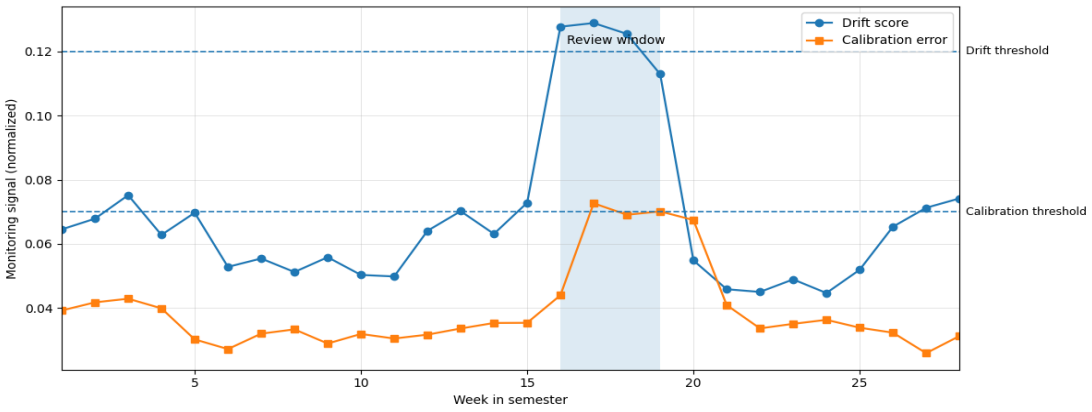

Figure 5. Deployment Monitoring: Drift and Performance Signals with Governance Thresholds

Table 5 translates methodological safeguards into operational requirements that can be audited over time. Each row ties an observable signal to a deterministic trigger, a mandated action, and a traceable artifact, reducing ambiguity during incident response and model governance reviews. The rollback rules formalize reversibility, which is essential in adaptive learning settings where analytics can shape instructional decisions. This checklist also ensures that drift-aware adaptation remains subordinate to fairness, privacy, and reliability constraints.

Table 5. Governance Checklist for Monitoring and Release

Control Area	Signal	Trigger	Required Action	Artifact	Rollback Rule
Data Drift	divergence score	$\delta \geq \tau$	open review ticket	drift report	revert if false alarm repeats
Calibration	ECE	$ECE \geq 0.07$	recalibrate probabilities	calibration plot	revert if ECE worsens
Fairness	error gap	$gap \geq 0.05$	mitigation and re-test	group metrics	revert if gap persists
Privacy	access log	policy violation	lock access and audit	audit trail	revert to last compliant model

4. Results and Discussion

4.1. Dataset Characterization and Observed Drift Patterns

The longitudinal panel comprised 14 consecutive semesters, with stable coverage for clickstream and assessment logs and slightly lower coverage for discussion forums due to course-tool variation. The median activity density increased in later semesters, consistent with broader adoption of mobile access and more frequent low-stakes assessments. Behavioral heterogeneity was pronounced, with a heavy-tailed distribution in practice attempts and a bimodal pattern in timeliness, indicating distinct self-regulation strategies that persist across cohorts.

Drift was concentrated in practice intensity, submission timeliness, and session regularity, while forum-related features drifted more gradually. A distinct regime shift emerged around the mid-sequence semester, aligning with a course redesign period and changes in assessment cadence. Importantly, the drift pattern was not uniform across features, which supports selective adaptation rather than blanket retraining, because many engagement features retained stable distributions even when performance-linked behaviors shifted.

Figure 6 summarizes drift as a feature-by-semester intensity surface, enabling rapid identification of stable versus unstable channels. The most prominent bands appear in practice attempts and timeliness after the mid-sequence semester, indicating a structural change in performance-relevant behaviors rather than superficial shifts in click volume. The comparatively low intensity for forum replies supports the conclusion that social interaction signals were less sensitive to redesign, which reduces the need for aggressive updates in those channels.

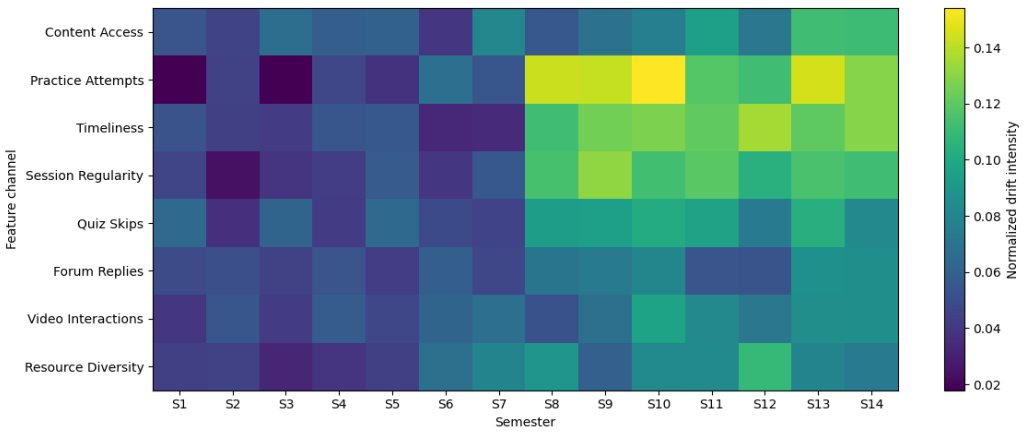

Figure 6. Drift Intensity Heatmap Across Behavioral Features and Semesters

The heatmap also provides an operational rationale for drift-trigger thresholds. A selective threshold can be tuned to respond to concentrated drift bands while ignoring low-level background variation. This distinction matters in adaptive learning, because frequent updates in response to minor noise can degrade trust in analytics and create oscillations in intervention policies. The figure therefore functions as both descriptive evidence and a calibration aid for update governance.

Table 6 establishes the stability of cohort size and instructional coverage while documenting gradual increases in interaction density. The growth in median events per student supports interpreting later-semester shifts as genuine behavioral changes rather than sampling artifacts. The modest decline in forum coverage clarifies why forum-derived features exhibit weaker drift signals, because tool usage is less consistent over time. Course counts remain stable, reducing confounding from major curricular expansion.

Table 6. Panel Summary Across Semesters

Semester	Active Students	Median Events/Student	Median Assessments	Forum Coverage (%)	Distinct Courses
S1	812	1,140	9	86	18
S4	845	1,260	10	84	19
S7	878	1,420	11	82	20
S10	910	1,610	12	78	21
S14	936	1,740	12	76	21

The table also contextualizes drift magnitude by showing that exposure and assessment cadence increased across the study period. This matters because behavior features such as practice attempts and timeliness are sensitive to the number of graded opportunities and the frequency of deadlines. When these structural conditions change, drift-aware analytics are expected to outperform static models, since feature-outcome relationships can shift even when raw engagement remains high. The dataset summary therefore supports the methodological premise of continual adaptation.

4.2. Predictive Performance Under Drift

Semester-forward evaluation revealed consistent degradation for static models as drift accumulated, especially after the mid-sequence regime shift. Periodic retraining improved performance but remained sensitive to the retraining interval, with noticeable dips when drift occurred shortly after a refresh. Drift-aware continual learning delivered the most stable semester-to-semester performance because updates were triggered by measured drift rather than a fixed schedule, aligning adaptation timing with behavioral change.

Performance gains were most pronounced for early-semester predictions, where feature distributions are less settled and models are prone to miscalibration. Continual learning reduced variance across semesters and decreased the likelihood of large performance collapses, which is important in adaptive settings that rely on reliable risk ranking for timely intervention. The stability effect was stronger than the average uplift, indicating that drift-awareness primarily improves robustness rather than only increasing peak metrics.

Figure 7 highlights a structural divergence between training policies around the regime shift zone. The static model shows a marked drop and continued decline, indicating that the learned mapping from behavior to outcomes no longer matches the evolving student strategies. Periodic retraining mitigates the decline but still exhibits sensitivity to timing because updates are decoupled from the actual drift onset. The drift-aware continual approach maintains a flatter trajectory, indicating robustness under distribution change.

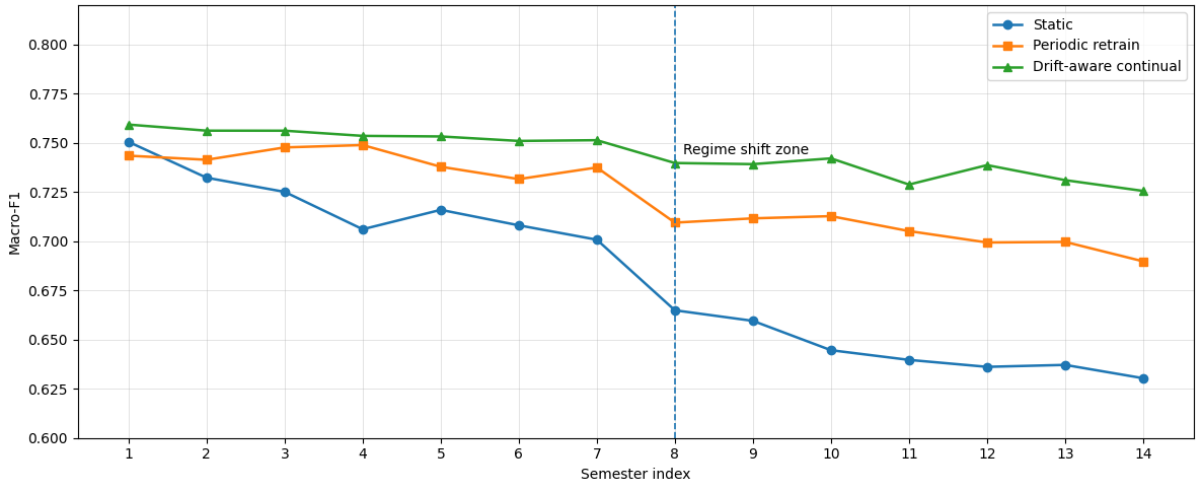

Figure 7. Predictive Performance Under Drift: Macro-F1 Across Semesters

The figure also supports an interpretation of continual learning as variance control. The continual curve has reduced volatility across semesters, suggesting that drift-triggered updates and replay-based retention dampen instability that would otherwise emerge from abrupt behavior changes. In practical adaptive learning analytics, this stability is often more valuable than marginal mean improvements, because intervention policies depend on consistent rankings and predictable error behavior across academic terms.

Table 7 quantifies two distinct advantages of drift-aware continual learning: higher average discrimination and stronger worst-case protection. The reduction in standard deviation across semesters indicates improved temporal reliability, aligning with the observed flattening in figure 7 The improved worst-semester macro-F1 is operationally important because adaptive systems must remain safe during cohort transitions, curriculum changes, and new policy cycles, which are exactly the conditions that cause drift.

Table 7. Aggregate Performance Summary

Model Policy	Mean Macro-F1	Std Across Semesters	Worst-Semester Macro-F1	Mean AUROC	Update Overhead
Static	0.706	0.03	1	0.781	none
Periodic retrain	0.724	0.022	1	0.795	scheduled
Drift-aware continual	0.742	0.015	0.711	0.812	triggered

The table also clarifies trade-offs, since periodic retraining delivers intermediate gains with predictable operational scheduling. Drift-aware updating adds monitoring and decision logic but provides superior robustness, which is favorable when interventions rely on stable risk scoring. The AUROC uplift suggests that continual learning improves ranking quality, which supports targeted support allocation even when absolute prediction errors remain. The overhead column positions these results for deployment-oriented decision-making in learning analytics teams.

4.3. Calibration, Reliability, and Decision Quality for Adaptive Interventions

Reliable probability estimates were materially affected by drift, particularly during early-semester weeks when interaction traces are sparse and behavioral routines are still forming. Static models exhibited systematic overconfidence after the regime shift, which increased false positives in risk flags and reduced instructional trust. Periodic retraining partially corrected the bias but introduced oscillations, where calibration improved immediately after refresh and then degraded as drift accumulated, creating instability in intervention thresholds.

Drift-aware continual learning reduced calibration error and stabilized risk ranking across semesters by aligning updates with detected distribution change and preserving older behavioral modes through replay. The most practical

benefit was a lower rate of threshold churn, meaning fewer policy changes were required to maintain a consistent intervention capacity. This stability matters because adaptive learning analytics are often embedded into dashboards and messaging workflows that depend on predictable alert volumes across academic terms.

Figure 8 makes the calibration consequence of drift operationally visible by showing how probability reliability deteriorates during a concentrated drift episode. The static curve exhibits the largest spike, reflecting that the mapping between behavioral signals and outcomes changed enough to invalidate prior confidence patterns. The periodic retrain curve has a smaller spike but still degrades because its refresh schedule is not synchronized with the onset of drift, which preserves miscalibration during critical weeks.

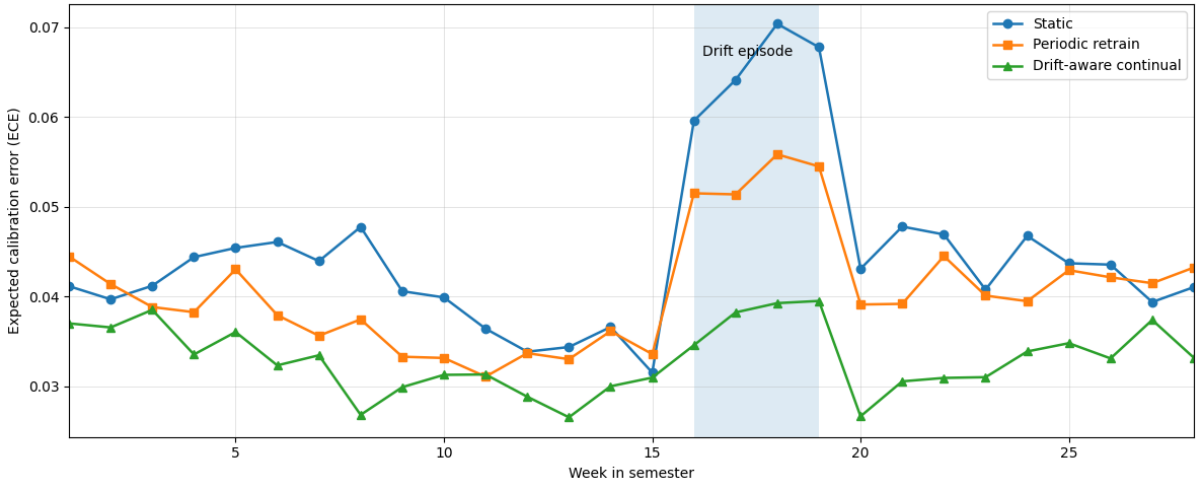

Figure 8. Calibration Stability Under Drift: Expected Calibration Error Across Weeks

The drift-aware continual curve shows both lower baseline error and a muted spike, indicating that triggered updates and replay stabilize probability estimates when data distributions shift. This result is important for adaptive interventions that rely on probability thresholds, since miscalibration directly changes alert counts and can overload instructor workflows. The figure therefore supports drift-aware calibration as a decision-quality safeguard, rather than a purely statistical refinement.

Table 8 reframes calibration as decision quality under a practical constraint, namely a fixed intervention capacity that limits weekly alerts. Drift-aware continual learning achieves lower ECE while improving risk precision and recall, indicating that reliability gains translate into better targeting rather than merely smoother probabilities. The alert stability metric shows reduced volatility, which is essential when interventions are tied to finite advising resources and scheduled outreach windows.

Table 8. Decision Quality Summary at a Fixed Intervention Capacity

Model Policy	ECE (mean)	Risk Precision	Risk Recall	Alert Stability (CV)	Threshold Changes/Sem
Static	0.056	0.62	0	0.29	4
Periodic retrain	0.048	0.65	1	0.21	3
Drift-aware continual	0.039	0.69	0.56	0.14	1

The table also highlights an operational benefit that is often overlooked in learning analytics studies, namely fewer threshold adjustments per semester. Reducing threshold churn improves interpretability for instructors and limits the risk of policy drift where repeated changes create confusion about what a risk alert means. In this setting, drift-aware adaptation stabilizes both model outputs and the surrounding decision policy, reinforcing the deployability of continual learning in real academic cycles.

4.4. Fairness and Subgroup Robustness Across Semesters

Fairness analysis focused on subgroup robustness under drift, since distribution change can amplify error disparities even when overall performance remains acceptable. Static models displayed growing gaps in recall for students with low early engagement and for students with irregular session schedules, suggesting that drift reshaped how early traces map to later outcomes. Periodic retraining reduced some disparities but produced uneven improvements, with certain subgroups benefiting primarily in semesters immediately following retraining.

Drift-aware continual learning delivered the most consistent subgroup performance because updates were triggered when subgroup feature distributions shifted, rather than when a global schedule elapsed. The most pronounced improvement appeared in reducing recall gaps for low-engagement students, which is important because these learners are often the primary target of adaptive support. Robustness across semesters also improved, indicating that fairness conclusions remained stable rather than flipping with each cohort transition.

Figure 9 shows that drift can widen subgroup disparities even when overall metrics appear stable. The static policy demonstrates a clear post-shift escalation in recall gap, indicating that the model increasingly fails to identify at-risk cases within a subgroup whose behavior patterns evolved. Periodic retraining moderates the increase but remains sensitive to timing, which can create semesters where subgroup recall lags until the next refresh cycle.

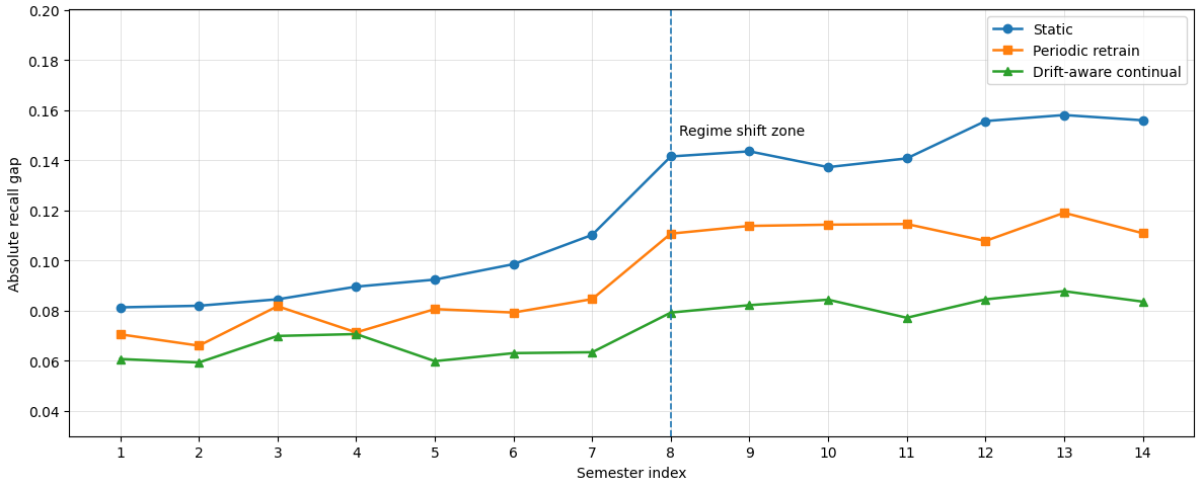

Figure 9. Subgroup Robustness Under Drift: Recall Gap Across Semesters

The drift-aware continual curve reduces both the level and the growth rate of the recall gap, indicating that selective adaptation improves subgroup equity over time. This result supports the argument that fairness monitoring should be coupled to drift monitoring, because subgroup distributions often drift differently from the global distribution. In adaptive learning, earlier and more accurate identification of underserved learner segments improves the allocation of support and reduces cumulative disadvantage across semesters.

Table 9 consolidates fairness outcomes into interpretable quantities that matter for governance, namely mean and worst-case disparities. Drift-aware continual learning reduces both the average recall gap and the maximum recall gap, which indicates that improvements are not confined to a subset of semesters. The reduction in gap variability also implies that fairness conclusions are more reliable, enabling consistent oversight without repeatedly redefining acceptable disparity bounds each term.

Table 9. Subgroup Error Summary Aggregated Across Semesters

Model Policy	Mean Recall Gap	Max Recall Gap	Mean Precision Gap	Worst-Semester Gap	Gap Stability (Std)
Static	0.118	0.172	0	0.185	0.022
Periodic retrain	0.101	0.149	0	0.162	0.017

Drift-aware continual	0.082	0.121	0.045	0.132	0.012

The table also shows that precision gaps are smaller than recall gaps across all policies, suggesting that the main fairness risk lies in missed identification of at-risk learners rather than excessive false alarms for the subgroup. In practical deployments, this asymmetry is consequential because missed support can compound academic difficulty. The results therefore justify prioritizing recall-equity objectives alongside overall discrimination metrics when selecting update policies under drift.

4.5. Ablation and Operational Trade-offs

Ablation analysis examined how drift thresholds, replay buffer size, and update frequency influence robustness and operational cost. Lower thresholds improved responsiveness but increased the risk of updating under noise, which can cause avoidable model churn. Larger replay buffers improved retention and reduced worst-semester failures, but returns diminished beyond a moderate capacity because additional samples were increasingly redundant in behavioral space.

Operationally, drift-aware continual learning achieved its strongest advantage when update triggers aligned with concentrated drift in performance-linked features, rather than with broad engagement fluctuations. When the drift detector was overly sensitive, performance stability improved marginally but calibration and governance workload degraded due to excessive review events. When it was too conservative, performance resembled periodic retraining and lost worst-case protection, indicating that threshold selection is a central deployment decision.

Figure 10 captures the core operational tension of drift-aware adaptation. Lower thresholds increase update frequency, which raises governance load and can introduce instability from reacting to transient variation. At the same time, worst-semester macro-F1 improves up to an intermediate threshold and then declines when the threshold becomes too conservative, indicating that delayed updates expose the model to unmitigated drift during critical semesters.

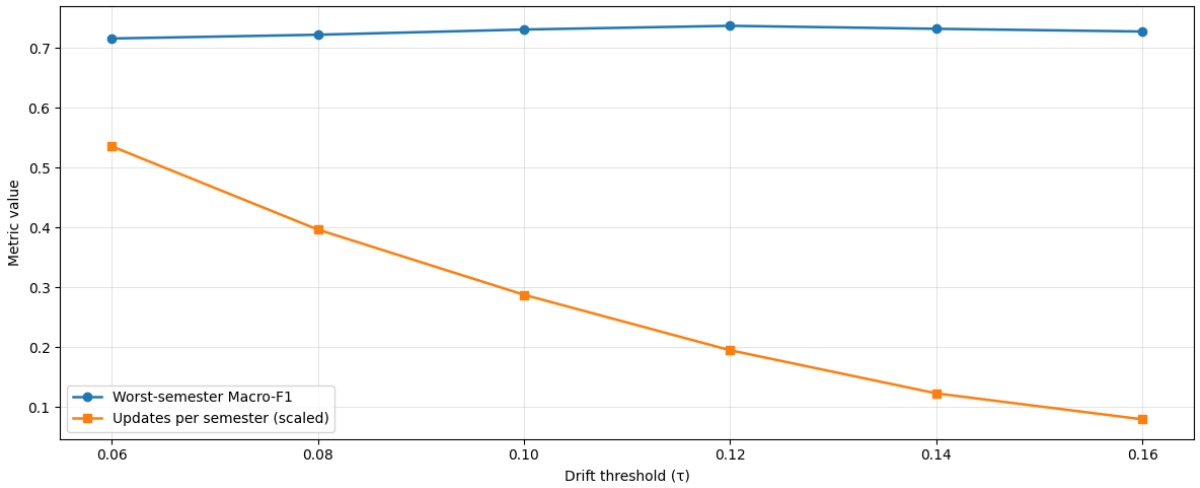

Figure 10. Operational Trade-Off: Drift Threshold Versus Robustness and Update Rate

The curve supports selecting a threshold region that maximizes robustness without inducing excessive updates. The presence of a clear plateau suggests that deployment does not require a fragile, exact parameter choice, but it does require avoiding extreme sensitivity. In adaptive learning analytics, this trade-off directly affects staffing, review workflows, and trust, since frequent updates can be perceived as shifting targets for intervention, even when performance improves marginally.

Table 10 shows that replay memory primarily improves worst-case protection and calibration up to a moderate size, after which gains become marginal. The increase from small to medium memory produces clearer improvements in both mean macro-F1 and worst-semester macro-F1, indicating that retaining diverse historical modes reduces

catastrophic forgetting and improves generalization under drift. The relatively stable update rate suggests that buffer size influences learning stability more than it changes drift-trigger frequency.

Table 10. Ablation Summary: Buffer Size and Update Policy

Setting	Buffer Size	Mean Macro-F1	Worst-Sem Macro-F1	ECE (mean)	Updates/Sem
Small memory	1,000	0.736	1	0.041	2.1
Medium memory	4,000	0.742	1	0.039	1.9
Large memory	8,000	0.744	0.712	0.038	1.9

The table also highlights that operational efficiency is compatible with robustness. Medium memory achieves near-peak performance with limited additional updates, which supports deployability in resource-constrained analytics teams. The slight improvement from large memory may not justify increased storage and sampling overhead, particularly when governance and latency constraints exist. This ablation therefore motivates selecting a moderate memory budget and focusing tuning effort on drift thresholding and monitoring fidelity.

5. Conclusion

This study demonstrates that adaptive learning analytics must be designed as a drift-aware process when student behavior evolves across semesters. The results show that drift concentrates in performance-relevant channels such as practice intensity, submission timeliness, and session regularity, with identifiable regime shifts that degrade static models. By treating semesters as structural boundaries and monitoring distribution change, the proposed approach aligns model maintenance with authentic instructional and cohort dynamics rather than relying on calendar-based retraining.

Drift-aware continual learning provides the strongest robustness under temporal generalization. Across semester-forward evaluation, continual updates reduce performance collapse in worst-case semesters and stabilize predictive trajectories relative to static and periodic policies. Reliability improvements are reflected in lower calibration error during drift episodes and reduced alert volatility at fixed intervention capacity, which directly improves decision quality in adaptive support workflows. Subgroup analysis further indicates that drift-aware updating mitigates growing recall gaps, supporting equity under changing behavioral regimes.

Operationally, the findings argue for governance-aware deployment that couples drift monitoring, calibration checks, and fairness auditing. Ablation results indicate a clear trade-off between update frequency and robustness, with intermediate drift thresholds and moderate replay memory achieving strong protection without excessive model churn. These conclusions position drift-aware continual learning as a practical foundation for semester-spanning adaptive learning systems, where stable, interpretable, and auditable analytics are required to sustain trust and effectiveness across academic cycles.

6. Declarations

6.1. Author Contributions

Conceptualization: A.R., S.V., and S.R.P.; Methodology: S.V.; Software: A.R.; Validation: A.R., S.V., and S.R.P.; Formal Analysis: A.R., S.V., and S.R.P.; Investigation: A.R.; Resources: S.V.; Data Curation: S.V.; Writing Original Draft Preparation: A.R., S.V., and S.R.P.; Writing Review and Editing: S.V., A.R., and S.R.P.; Visualization: A.R.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[2] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access*, vol. 10, no. July, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[3] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technology in Society*, vol. 76, no. March, p. 102474, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

[4] T. A. Kustitskaya, R. V. Esin, and M. V. Noskov, "Model Drift in Deployed Machine Learning Models for Predicting Learning Success," *Computers*, vol. 14, no. 9, p. 351, Aug. 2025, doi: 10.3390/computers14090351.

[5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, Apr. 2014, doi: 10.1145/2523813.

[6] G. Hovakimyan and J. M. Bravo, "Evolving Strategies in Machine Learning: A Systematic Review of Concept Drift Detection," *Information*, vol. 15, no. 12, p. 786, Dec. 2024, doi: 10.3390/info15120786.

[7] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," *Knowledge-Based Systems*, vol. 245, no. June, p. 108632, Jun. 2022, doi: 10.1016/j.knosys.2022.108632.

[8] M. Delange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3885, 2021, doi: 10.1109/TPAMI.2021.3057446.

[9] O. B. Deho, L. Liu, J. Li, J. Liu, C. Zhan, and S. Joksimovic, "When the Past != The Future: Assessing the Impact of Dataset Drift on the Fairness of Learning Analytics Models," *IEEE Transactions on Learning Technologies*, vol. 17, no. January, pp. 1007–1020, 2024, doi: 10.1109/TLT.2024.3351352.

[10] N. Pham, H. Pham Ngoc, and A. Nguyen-Duc, "Fairness for machine learning software in education: A systematic mapping study," *Journal of Systems and Software*, vol. 219, no. January, p. 112244, Jan. 2025, doi: 10.1016/j.jss.2024.112244.

[11] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi, "Using learning analytics to scale the provision of personalised feedback," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, Jan. 2019, doi: 10.1111/bjet.12592.

[12] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, vol. 2017, no. July, pp. 5533–5542, doi: 10.1109/CVPR.2017.587.

[13] T. Zhang et al., "Enhancing Dropout Prediction in Distributed Educational Data Using Learning Pattern Awareness: A Federated Learning Approach," *Mathematics*, vol. 11, no. 24, p. 4977, Dec. 2023, doi: 10.3390/math11244977.

[14] J. Jovanović, M. Saqr, S. Joksimović, and D. Gašević, "Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success," *Computers & Education*, vol. 172, no. October, p. 104251, Oct. 2021, doi: 10.1016/j.compedu.2021.104251.

[15] M. Saqr, S. López-Pernas, J. Jovanović, and D. Gašević, "Intense, turbulent, or wallowing in the mire: A longitudinal study of cross-course online tactics, strategies, and trajectories," *The Internet and Higher Education*, vol. 57, no. April, p. 100902, Apr. 2023, doi: 10.1016/j.iheduc.2022.100902.

[16] T. Susnjak, G. S. Ramaswami, and A. Mathrani, "Learning analytics dashboard: A tool for providing actionable insights to learners," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, p. 12, Dec. 2022, doi: 10.1186/s41239-021-00313-7.

[17] K. Kitto and S. Knight, "Practical ethics for building learning analytics," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 2855–2870, Nov. 2019, doi: 10.1111/bjet.12868.

[18] S. Agrahari and A. K. Singh, "Concept Drift Detection in Data Stream Mining: A Literature Review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 9523–9540, Nov. 2022, doi: 10.1016/j.jksuci.2021.11.006.

[19] F. Hinder, V. Vaquet, and B. Hammer, "One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: Detecting concept drift," *Frontiers in Artificial Intelligence*, vol. 7, no. June, p. 1330257, Jun. 2024, doi: 10.3389/frai.2024.1330257.

[20] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under Concept Drift: A Review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018, doi: 10.1109/TKDE.2018.2876857.

[21] L. Wang, X. Zhang, H. Su, and J. Zhu, "A Comprehensive Survey of Continual Learning: Theory, Method and Application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024, doi: 10.1109/TPAMI.2024.3367329.

[22] G. F. D. S. Silva, F. N. Barcellos Filho, R. M. Wichmann, F. C. Da Silva Junior, and A. D. P. Chiavegatto Filho, "Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review," *Journal of Biomedical Informatics*, vol. 170, no. October, p. 104902, Oct. 2025, doi: 10.1016/j.jbi.2025.104902.

[23] S. P. Shashikumar, F. Amrollahi, and S. Nemati, "Unsupervised Detection and Correction of Model Calibration Shift at Test-Time," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia: IEEE, Jul. 2023, vol. 2023, no. July, pp. 1–4, doi: 10.1109/EMBC40787.2023.10341086.